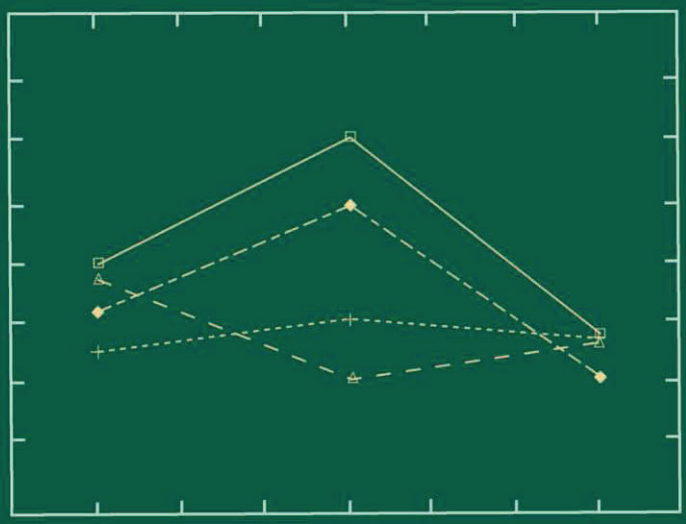
SPRINGER TEXTS IN STATISTICS

Ronald Christensen

Log-Linear Models *and* Logistic Regression

Second Edition



Springer

Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Statistics in Scientific Investigation: Its Basis, Application, and Interpretation
Mueller: Basic Principles of Structural Equation Modeling

(continued after index)

Ronald Christensen

Log-Linear Models and Logistic Regression

Second Edition

Springer

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131
USA

BMDP Statistical Software is distributed by SPSS Inc., 444 N. Michigan Avenue, Chicago, IL, 60611, telephone: (800) 543-2185.

MINITAB is a registered trademark of Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801, telephone: (814) 238-3280, telex: 881612.

MSUSTAT is marketed by the Research and Development Institute Inc., Montana State University, Bozeman, MT 59717-0002, Attn: R.E. Lund.

Library of Congress Cataloging-in-Publication Data
Christensen, Ronald, 1951-
Log-linear models and logistic regression / Ronald Christensen. — 2nd ed.
p. cm. — (Springer texts in statistics)
Earlier ed. published under title: Log-linear models. 1990.
Includes bibliographical references and index.
ISBN 0-387-98247-7
1. Log-linear models. I. Christensen, Ronald, 1951- Log-linear models. II. Title. III. Series.
QA278.C49 1997
519.5'35—dc21 97-12465

ISBN 0-387-98247-7 Springer-Verlag New York Berlin Heidelberg SPIN 10747963

To Sharon and Fletch

Preface to the Second Edition

As the new title indicates, this second edition of *Log-Linear Models* has been modified to place greater emphasis on logistic regression. In addition to new material, the book has been radically rearranged. The fundamental material is contained in Chapters 1-4. Intermediate topics are presented in Chapters 5 through 8. Generalized linear models are presented in Chapter 9. The matrix approach to log-linear models and logistic regression is presented in Chapters 10-12, with Chapters 10 and 11 at the applied Ph.D. level and Chapter 12 doing theory at the Ph.D. level.

The largest single addition to the book is Chapter 13 on Bayesian binomial regression. This chapter includes not only logistic regression but also probit and complementary log-log regression. With the simplicity of the Bayesian approach and the ability to do (almost) exact small sample statistical inference, I personally find it hard to justify doing traditional large sample inferences. (Another possibility is to do exact conditional inference, but that is another story.)

Naturally, I have cleaned up the minor flaws in the text that I have found. All examples, theorems, proofs, lemmas, etc. are numbered consecutively within each section with no distinctions between them, thus Example 2.3.1 will come before Proposition 2.3.2. Exercises that do not appear in a section at the end have a separate numbering scheme. Within the section in which it appears, an equation is numbered with a single value, e.g., equation (1). When reference is made to an equation that appears in a different section, the reference includes the appropriate chapter and section, e.g., equation (2.1.1).

The primary prerequisite for using this book is knowledge of analysis of variance and regression at the masters degree level. It would also be advantageous to have some prior familiarity with the analysis of two-way tables of count data. Christensen (1996a) was written with the idea of preparing people for this book and for Christensen (1996b). In addition, familiarity with masters level probability and mathematical statistics would be helpful, especially for the later chapters. Sections 9.3, 10.2, 11.6, and 12.3 use ideas of the convergence of random variables. Chapter 12 was originally the last chapter in my linear models book, so I would recommend a good course in linear models before attempting that. A good course in linear models would also help for Chapters 10 and 11.

The analysis of logistic regression and log-linear models is not possible without modern computing. While it certainly is not the goal of this book to provide training in the use of various software packages, some examples of software commands have been included. These focus primarily on SAS and BMDP, but include some GLIM (of which I am still very fond).

I would particularly like to thank Ed Bedrick for his help in preparing this edition and Ed and Wes Johnson for our collaboration in developing the material in Chapter 13. I would also like to thank Turner Ostler for providing the trauma data and his prior opinions about it.

Most of the data, and all of the larger data sets, are available from STATLIB as well as by anonymous ftp. The web address for the datasets option in STATLIB is **http://www.stat.cmu.edu/datasets/**. The data are identified as "christensen-llm". To use ftp, type **ftp stat.unm.edu** and login as "anonymous", enter **cd /pub/fletcher** and either get **llm.tar.Z** for Unix machines or **llm.zip** for a DOS version. More information is available from the file "readme.llm" or at **http://stat.unm.edu/~fletcher**, my web homepage.

Ronald Christensen
Albuquerque, New Mexico
February, 1997

BMDP Statistical Software is distributed by SPSS Inc., 444 N. Michigan Avenue, Chicago, IL, 60611, telephone: (800) 543-2185.

MINITAB is a registered trademark of Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801, telephone: (814) 238-3280, telex: 881612.

MSUSTAT is marketed by the Research and Development Institute Inc., Montana State University, Bozeman, MT 59717-0002, Attn: R.E. Lund.

Preface to the First Edition

This book examines log-linear models for contingency tables. Logistic regression and logistic discrimination are treated as special cases and generalized linear models (in the GLIM sense) are also discussed. The book is designed to fill a niche between basic introductory books such as Fienberg (1980) and Everitt (1977) and advanced books such as Bishop, Fienberg, and Holland (1975), Haberman (1974a), and Santner and Duffy (1989). It is primarily directed at advanced Masters degree students in Statistics but it can be used at both higher and lower levels. The primary theme of the book is using previous knowledge of analysis of variance and regression to motivate and explicate the use of log-linear models. Of course, both the analogies and the distinctions between the different methods must be kept in mind.

[From the first edition, Chapters I, II, and III are about the same as the new 1, 2, and 3. Chapter IV is now Chapters 5 and 6. Chapter V is now 7, VI is 10, VII is 4 (and the sections are rearranged), VIII is 11, IX is 8, X is 9, and XV is 12.]

The book is written at several levels. A basic introductory course would take material from Chapters I, II (deemphasizing Section II.4), III, Sections IV.1 through IV.5 (eliminating the material on graphical models), Section IV.10, Chapter VII, and Chapter IX. The advanced modeling material at the end of Sections VII.1, VII.2, and possibly the material in Section IX.2 should be deleted in a basic introductory course. For Masters degree students in Statistics, all the material in Chapters I through V, VII, IX, and X should be accessible. For an applied Ph.D. course or for advanced Masters students, the material in Chapters VI and VIII can be incorporated. Chapter VI recapitulates material from the first five chapters using matrix notation. Chapter VIII recapitulates Chapter VII. This material is necessary (a) to get standard errors of estimates in anything other than the saturated model, (b) to explain the Newton-Raphson (iteratively reweighted least squares) algorithm, and (c) to discuss the weighted least

squares approach of Grizzle, Starmer, and Koch (1969). I also think that the more general approach used in these chapters provides a deeper understanding of the subject. Most of the material in Chapters VI and VIII requires no more sophistication than matrix arithmetic and being able to understand the definition of a column space. All of the material should be accessible to people who have had a course in linear models. Throughout the book, Chapter XV of Christensen (1987) is referenced for technical details. For completeness, and to allow the book to be used in nonapplied Ph.D. courses, Chapter XV has been reprinted in this volume under the same title, Chapter XV.

The prerequisites differ for the various courses described above. At a minimum, readers should have had a traditional course in statistical methods. To understand the vast majority of the book, courses in regression, analysis of variance, and basic statistical theory are recommended. To fully appreciate the book, it would help to already know linear model theory.

It is difficult for me to understand but many of my acquaintance view me as quite opinionated. While I admit that I have not tried to keep my opinions to myself, I have tried to clearly acknowledge them as my opinions.

There are many people I would like to thank in connection with this work. My family, Sharon and Fletch, were supportive throughout. Jackie Damrau did an exceptional job of typing the first draft. The folks at BMDP provided me with copies of 4F, LR, and 9R. MINITAB provided me with Versions 6.1 and 6.2. Dick Lund gave me a copy of MSUSTAT. All of the computations were performed with this software or GLIM. Several people made valuable comments on the manuscript; these include Rahman Azari, Larry Blackwood, Ron Schrader, and Elizabeth Slate. Joe Hill introduced me to statistical applications of graph theory and convinced me of their importance and elegance. He also commented on part of the book. My editors, Steve Fienberg and Ingram Olkin, were, as always, very helpful. Like many people, I originally learned about log-linear models from Steve's book. Two people deserve special mention for how much they contributed to this effort. I would not be the author of this book were it not for the amount of support provided in its development by Ed Bedrick and Wes Johnson. Wes provided much of the data used in the examples. I suppose that I should also thank the legislature of the state of Montana. It was their penury, while I worked at Montana State University, that motivated me to begin the project in the spring of 1987. If you don't like the book, blame them!

Ronald Christensen
Albuquerque, New Mexico
April 5, 1990
(Happy Birthday Dad)

Contents

1
Introduction

This book is concerned with the analysis of cross-classified categorical data using log-linear models and with logistic regression. Log-linear models have two great advantages: they are flexible and they are interpretable. Log-linear models have all the modeling flexibility that is associated with analysis of variance and regression. They also have natural interpretations in terms of odds and frequently have interpretations in terms of independence. This book also examines logistic regression and logistic discrimination, which typically involve the use of continuous predictor variables. Actually, these are just special cases of log-linear models. There is a wide literature on log-linear models and logistic regression and a number of books have been written on the subject. Some additional references on log-linear models that I can recommend are: Agresti (1984, 1990), Andersen (1991), Bishop, Fienberg, and Holland (1975), Everitt (1977), Fienberg (1980), Haberman (1974a), Plackett (1981), Read and Cressie (1988), and Santner and Duffy (1989). Cox and Snell (1989) and Hosmer and Lemeshow (1989) have written books on logistic regression. One reason I can recommend these is that they are all quite different from each other and from this book. There are differences in level, emphasis, and approach. This is by no means an exhaustive list; other good books are available.

In this chapter we review basic information on conditional independence, random variables, expected values, variances, standard deviations, covariances, and correlations. We also review the distributions most commonly used in the analysis of contingency tables: the binomial, the multinomial, product multinomials, and the Poisson. Christensen (1996a, Chapter 1) contains a more extensive review of most of this material.

1.1 Conditional Probability and Independence

This section introduces two subjects that are fundamental to the analysis of count data. Both subjects are quite elementary, but they are used so extensively that a detailed review is in order. One subject is the definition and use of *odds*. We include as part of this subject the definition and use of *odds ratios*. The other is the use of independence and conditional independence in characterizing probabilities. We begin with a discussion of odds.

Odds will be most familiar to many readers from their use in sporting events. They are not infrequently confused with probabilities. (I once attended an American Statistical Association chapter meeting at which a government publication on the Montana state lottery was disbursed that presented probabilities of winning but called them odds of winning.) In log-linear model analysis and logistic regression, both odds and ratios of odds are used extensively.

Suppose that an event, say, the sun rising tomorrow, has a probability p. The odds of that event are

$$Odds = \frac{p}{1-p} = \frac{\Pr(\text{Event Occurs})}{\Pr(\text{Event Does Not Occur})}.$$

Thus, supposing the probability that the sun will rise tomorrow is .8, the odds that the sun will rise tomorrow are $.8/.2 = 4$. Writing 4 as 4/1, it might be said that the odds of the sun rising tomorrow are 4 to 1. The fact that the odds are greater than one indicates that the event has a probability of occurring greater than one-half. Conversely, if the odds are less than one, the event has probability of occurring less than one-half. For example, the probability that the sun *will not* rise tomorrow is $1 - .8 = .2$ and the odds that the sun will not rise tomorrow are $.2/.8 = 1/4$.

The larger the odds, the larger the probability. The closer the odds are to zero, the closer the probability is to zero. In fact, for probabilities and odds that are very close to zero, there is essentially no difference between the numbers. As for all lotteries, the probability of winning big in the Montana state lottery was very small. Thus, the mistake alluded to above is of no practical importance. On the other hand, as probabilities get near one, the corresponding odds approach infinity.

Given the odds that an event occurs, the probability of the event is easily obtained. If the odds are O, then the probability p is easily seen to be

$$p = \frac{O}{O+1}.$$

For example, if the odds of breaking your wrist in a severe bicycle accident are .166, the probability of breaking your wrist is $.166/1.166 = .142$ or about 1/7. Note that even at this level, the numerical values of the odds and the probability are similar.

Examining odds really amounts to a rescaling of the measure of uncertainty. Probabilities between zero and one half correspond to odds between zero and one. Probabilities between one half and one correspond to odds between one and infinity. Another convenient rescaling is the log of the odds. Probabilities between zero and one half correspond to log odds between minus infinity and zero. Probabilities between one half and one correspond to odds between zero and infinity. The log odds scale is symmetric about zero just as probabilities are symmetric about one half. One unit above zero is comparable to one unit below zero. From above, the log odds that the sun will rise tomorrow are $\log(4)$, while the log odds that it will not rise are $\log(1/4) = -\log(4)$. These numbers are equidistant from the center 0. This symmetry of scale fails for the odds. The odds of 4 are three units above the center 1, while the odds of 1/4 are three-fourths of a unit below the center. For most mathematical purposes, the log odds are a more natural transformation than the odds.

EXAMPLE 1.1.1. *N.F.L. Football*
On January 5, 1990, I decided how much of my meager salary to bet on the upcoming Superbowl. There were eight teams still in contention. *The Albuquerque Journal* reported *Harrah's Odds* for each team. The teams and their odds are given below.

Team	Odds
San Francisco Forty-Niners	even
Denver Broncos	5 to 2
New York Giants	3 to 1
Cleveland Browns	9 to 2
Los Angeles Rams	5 to 1
Minnesota Vikings	6 to 1
Buffalo Bills	8 to 1
Pittsburgh Steelers	10 to 1

These odds were designed for the benefit of Harrah's and were not really anyone's idea of the odds that the various teams would win. (This will become all too clear later.) Nonetheless, we examine these odds as though they determine probabilities for winning the Superbowl as of January 5, 1990, and their implications for my early retirement. The discussion of betting is quite general. I have no particular knowledge of how Harrah's works these things.

The odds on the Vikings are 6 to 1. These are actually the odds that the Vikings *will not* win the Superbowl. The odds are a ratio, $6/1 = 6$. The probabilities are

$$\Pr(\text{Vikings do not win}) = \frac{6}{6+1} = \frac{6}{7}$$

and

$$\Pr(\text{Vikings win}) = \frac{\frac{1}{6}}{\frac{1}{6}+1} = \frac{1}{1+6} = \frac{1}{7}.$$

Similarly, the odds on Denver are 5 to 2 or 5/2. The probabilities are

$$\Pr(\text{Broncos do not win}) = \frac{\frac{5}{2}}{\frac{5}{2}+1} = \frac{5}{5+2} = \frac{5}{7}$$

and

$$\Pr(\text{Broncos win}) = \frac{\frac{2}{5}}{\frac{2}{5}+1} = \frac{2}{5+2} = \frac{2}{7}.$$

San Francisco is even money, so their odds are 1 to 1. The probabilities of winning for all eight teams are given below.

Team	Probability of Winning
San Francisco Forty-Niners	.50
Denver Broncos	.29
New York Giants	.25
Cleveland Browns	.18
Los Angeles Rams	.17
Minnesota Vikings	.14
Buffalo Bills	.11
Pittsburgh Steelers	.09

There is a peculiar thing about these probabilities: They should add up to 1 but do not. One of these eight teams had to win the 1990 Superbowl, so the probability of one of them winning must be 1. The eight events are disjoint, e.g., if the Vikings win, the Broncos cannot, so the sum of the probabilities should be the probability that any of the teams wins. This leads to a contradiction. The probability that any of the teams wins is

$$.50 + .29 + .25 + .18 + .17 + .14 + .11 + .09 = 1.73 \neq 1.$$

All of the odds have been deflated. The probability that the Vikings win should not be .14 but .14/1.73 = .0809. The odds against the Vikings should be $(1 - .0809)/.0809 = 11.36$. Rounding this to 11 gives the odds against the Vikings as 11 to 1 instead of the reported 6 to 1. This has severe implications for my early retirement.

The idea behind odds of 6 to 1 is that if I bet \$100 on the Vikings and they win, I should win \$600 and also have my original \$100 returned. Of course, if they lose I am out my \$100. According to the odds calculated above, a fair bet would be for me to win \$1100 on a bet of \$100. (Actually, I should get \$1136 but what is \$36 among friends.) Here, "fair" is used in a

technical sense. In a fair bet, the expected winnings are zero. In this case, my expected winnings for a fair bet are

$$1136(.0809) - 100(1 - .0809) = 0.$$

It is what I win times the probability that I win minus what I lose times the probability that I lose. If the probability of winning is .0809 and I get paid off at a rate of 6 to 1, my expected winnings are

$$600(.0809) - 100(1 - .0809) = -43.4.$$

I don't think I can afford that. In fact, a similar phenomenon occurs for a bet on any of the eight teams. If the probabilities of winning add up to more than one, the true expected winnings on any bet will be negative. Obviously, it pays to make the odds rather than the bets.

Not only odds but ratios of odds arise naturally in the analysis of logistic regression and log-linear models. It is important to develop some familiarity with *odds ratios*. The odds on San Francisco, Los Angeles, and Pittsburgh are 1 to 1, 5 to 1, and 10 to 1, respectively. Equivalently, the odds that each team will not win are 1, 5, and 10. Thus, L.A. has odds of not winning that are 5 times larger than San Francisco's and Pittsburgh's are 10 times larger than San Francisco's. The ratio of the odds of L.A. not winning to the odds of San Francisco not winning is $5/1 = 5$. The ratio of the odds of Pittsburgh not winning to San Francisco not winning is $10/1 = 10$. Also, Pittsburgh has odds of not winning that are twice as large as L.A.'s, i.e., $10/5 = 2$.

An interesting thing about odds ratios is that, say, the ratio of the odds of Pittsburgh not winning to the odds of L.A. not winning is the same as the ratio of the odds of L.A. winning to the odds of Pittsburgh winning. In other words, if Pittsburgh has odds of not winning that are 2 times larger than L.A.'s, L.A. must have odds of winning that are 2 times larger than Pittsburgh's. The odds of L.A. not winning are 5 to 1, so the odds of them winning are 1 to 5 or 1/5. Similarly, the odds of Pittsburgh winning are 1/10. Clearly, L.A. has odds of winning that are 2 times those of Pittsburgh. The odds ratio of L.A. winning to Pittsburgh winning is identical to the odds ratio of Pittsburgh not winning to L.A. not winning. Similarly, San Francisco has odds of winning that are 10 times larger than Pittsburgh's and 5 times as large as L.A.'s.

In logistic regression and log-linear model analysis, one of the most common uses for odds ratios is to observe that they equal one. If the odds ratio is one, the two sets of odds are equal. It is certainly of interest in a comparative study to be able to say that the odds of two things are the same. In this example, none of the odds ratios that can be formed is one because no odds are equal.

Another common use for odds ratios is to observe that two of them are the same. For example, the ratio of the odds of Pittsburgh not winning

relative to the odds of L.A. not winning is the same as the ratio of the odds of L.A. not winning to the odds of the Denver not winning. We have already seen that the first of these values is 2. The odds for L.A. not winning relative to Denver not winning are also 2 because $\frac{5}{1}/\frac{5}{2} = 2$. Even when the corresponding odds are different, odds ratios can be the same.

Marginal and *conditional probabilities* play important roles in logistic regression and log-linear model analysis. If $\Pr(B) > 0$, the conditional probability of A given B is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}\,.$$

It is the proportion of the probability of B in which A also occurs. To deal with conditional probabilities when $\Pr(B) = 0$ requires much more sophistication. It is an important topic in dealing with continuous observations, but it is not something we need to consider.

If knowing that B occurs does not change your information about A, then A is *independent* of B. Specifically, A is independent of B if

$$\Pr(A|B) = \Pr(A)\,.$$

This definition gets tied up in details related to the requirement that $\Pr(B) > 0$. A simpler and essentially equivalent definition is that A and B are independent if

$$\Pr(A \cap B) = \Pr(A)\Pr(B)\,.$$

EXAMPLE 1.1.2. Table 1.1 contains probabilities for nine combinations of hair and eye color. The nine outcomes are all combinations of three hair colors, Blond (BlH), Brown (BrH), and Red (RH), and three eye colors, Blue (BlE), Brown (BrE), and Green (GE).

Table 1.1
Hair-Eye Color Probabilities

		Eye Color		
		Blue	Brown	Green
Hair Color	Blond	.12	.15	.03
	Brown	.22	.34	.04
	Red	.06	.01	.03

The (*marginal*) probabilities for the various hair colors are obtained by summing over the rows:

$$\begin{aligned} \Pr(\text{BlH}) &= .12 + .15 + .03 = .3 \\ \Pr(\text{BrH}) &= .6 \\ \Pr(\text{RH}) &= .1\,. \end{aligned}$$

Probabilities for eye colors come from summing the columns. Blue, Brown, and Green eyes have probabilities .4, .5, and .1, respectively. The conditional probability of Blond Hair given Blue Eyes is

$$\begin{aligned}\Pr(\text{BlH}|\text{BlE}) &= \Pr((\text{BlH},\text{BlE}))/\Pr(\text{BlE})\\ &= .12/.4\\ &= .3\,.\end{aligned}$$

Note that $\Pr(\text{BlH}|\text{BlE}) = \Pr(\text{BlH})$, so the events BlH and BlE are independent. In other words, knowing that someone has blue eyes gives no additional information about whether that person has blond hair.

On the other hand,

$$\begin{aligned}\Pr(\text{BrH}|\text{BlE}) &= .22/.4\\ &= .55\,,\end{aligned}$$

while

$$\Pr(\text{BrH}) = .6,$$

so knowing that someone has blue eyes tells us that they are relatively less likely to have brown hair.

Now condition on blond hair,

$$\Pr(\text{BlE}|\text{BlH}) = .12/.3 = .4 = \Pr(\text{BlE})\,.$$

We again see that BlE and BlH are independent. In fact, it is also true that

$$\Pr(\text{BrE}|\text{BlH}) = \Pr(\text{BrE})$$

and

$$\Pr(\text{GE}|\text{BlH}) = \Pr(\text{GE})\,.$$

Knowing that someone has blond hair gives no additional information about any eye color.

Example 1.1.3. Consider the eight combinations of three factors: economic status (High, Low), residence (Montana, Haiti), and beverage of preference (Beer, Other). Probabilities are given below.

	Beer		Other		Total
	Montana	Haiti	Montana	Haiti	
High	.021	.009	.049	.021	.1
Low	.189	.081	.441	.189	.9
Total	.210	.090	.490	.210	1.0

The factors in this table are completely independent. If we condition on either beverage category, then economic status and residence are independent. If we condition on either residence, then economic status and beverage

are independent. If we condition on either economic status, residence and beverage are independent. No matter what you condition on and no matter what you look at, you get independence. For example,

$$\begin{aligned} \Pr(\text{High}|\text{Montana, Beer}) &= .021/.210 \\ &= .1 \\ &= \Pr(\text{High})\,. \end{aligned}$$

Similarly, knowing that someone has low economic status gives no additional information relative to whether their residence is Montana or Haiti.

The phenomenon of complete independence is characterized by the fact that every probability in the table is the product of the three corresponding marginal probabilities. For example,

$$\begin{aligned} \Pr(\text{Low, Montana, Beer}) &= .189 \\ &= (.9)(.7)(.3) \\ &= \Pr(\text{Low})\Pr(\text{Montana})\Pr(\text{Beer})\,. \end{aligned}$$

EXAMPLE 1.1.4. Consider the eight combinations of socioeconomic status (High, Low), political philosophy (Liberal, Conservative), and political affiliation (Democrat, Republican). Probabilities are given below.

	Democrat		Republican		Total
	Liberal	Conservative	Liberal	Conservative	
High	.12	.12	.04	.12	.4
Low	.18	.18	.06	.18	.6
Total	.30	.30	.10	.30	1.0

For any combination in the table, one of the three factors, socioeconomic status, is independent of the other two, political philosophy and political affiliation. For example,

$$\begin{aligned} \Pr(\text{High, Liberal, Republican}) &= .04 \\ &= (.4)(.1) \\ &= \Pr(\text{High})\Pr(\text{Liberal, Republican})\,. \end{aligned}$$

However, the other divisions of the three factors into two groups do not display this property. Political philosophy is not always independent of socioeconomic status and political affiliation, e.g.,

$$\begin{aligned} \Pr(\text{High, Liberal, Republican}) &= .04 \\ &\neq (.4)(.16) \\ &= \Pr(\text{Liberal})\Pr(\text{High, Republican})\,. \end{aligned}$$

Also, political affiliation is not always independent of socioeconomic status and political philosophy, e.g.,

$$\begin{aligned} \Pr(\text{High}, \text{Liberal}, \text{Republican}) &= .04 \\ &\neq (.4)(.16) \\ &= \Pr(\text{Republican})\Pr(\text{High}, \text{Liberal})\,. \end{aligned}$$

EXAMPLE 1.1.5. Consider the twelve outcomes that are all combinations of three factors, one with three levels and two with two levels. The factors and levels are given below. They are similar to those in a study by Reiss et al. (1975) that was reported in Fienberg (1980).

Factor	Levels
Attitude on Extramarital Coitus	Always Wrong, Not Always Wrong
Virginity	Virgin, Nonvirgin
Use of Contraceptives	Regular, Intermittent, None

The probabilities are

	Use of Contraceptives: Regular	
	Virgin	Nonvirgin
Always Wrong	3/50	12/50
Not Always	3/50	12/50
	Intermittent	
	Virgin	Nonvirgin
Always Wrong	1/80	2/80
Not Always	3/80	2/80
	None	
	Virgin	Nonvirgin
Always Wrong	3/40	1/40
Not Always	6/40	2/40

Consider the relationship between attitude and virginity given regular use of contraceptives. The probability of regular use is

$$\begin{aligned} \Pr(\text{Regular}) &= \frac{3}{50} + \frac{3}{50} + \frac{12}{50} + \frac{12}{50} \\ &= 30/50\,. \end{aligned}$$

The conditional probabilities given regular use are computed by dividing the entries in the 2×2 subtable for regular use by the (marginal) probability of regular use, 30/50, e.g., the probability for Always Wrong, Virgin given Regular is $(3/50)/(30/50) = .1$.

Conditional Probabilities Given
Regular Use of Contraceptives

	Virgin	Nonvirgin	Total
Always Wrong	.1	.4	.5
Not Always	.1	.4	.5
Total	.2	.8	1.0

Note that each entry is the product of the row total and the column total, e.g.,

$$\begin{aligned} &\Pr(\text{Always Wrong and Virgin}|\text{Regular}) \\ &\quad = \quad .1 \\ &\quad = \quad (.2)(.5) \\ &\quad = \quad \Pr(\text{Always Wrong}|\text{Regular})\Pr(\text{Virgin}|\text{Regular})\,. \end{aligned}$$

Because this is true for the entire 2×2 table, attitude and virginity are independent given regular use of contraceptives.

Similarly, the conditional probabilities given no use of contraceptives are

	Virgin	Nonvirgin	Total
Always Wrong	3/12	1/12	1/3
Not Always	6/12	2/12	2/3
Total	3/4	1/4	1

Again, it is easily seen that attitude and virginity are independent given no use of contraceptives.

Although we have independence given either no use or regular use, the probabilities of virginity and attitude change drastically. For regular use, nonvirginity is four times more probable than virginity. For no use, virginity is three times more probable. For regular use, attitudes are evenly split. For no use, the attitude that extramarital coitus is not always wrong is twice as probable as the attitude that it is always wrong.

If the conditional probabilities given intermittent use also display independence, we can describe the entire table as having attitude and virginity independent given use. Unfortunately, this does not occur. Conditional on intermittent use, the probabilities are

	Virgin	Nonvirgin	Total
Always Wrong	1/8	1/4	3/8
Not Always	3/8	1/4	5/8
Total	1/2	1/2	1

Virgins are three times as likely to think extramarital coitus is not always wrong, but nonvirgins are evenly split.

Conditional odds are readily obtained from the unconditional probabilities and other conditional probabilities. The odds that a virgin intermittent

contraceptive user thinks that extramarital coitus is not always wrong are

$$\begin{aligned}
&\frac{\Pr(\text{Not Always}|\text{Virgin, intermittent use})}{\Pr(\text{Always Wrong}|\text{Virgin, intermittent use})} \\
&\quad = \frac{\Pr(\text{Not Always, Virgin}|\text{intermittent use})}{\Pr(\text{Always Wrong, Virgin}|\text{intermittent use})} \\
&\quad = \frac{\Pr(\text{Not Always, Virgin, intermittent use})}{\Pr(\text{Always Wrong, Virgin, intermittent use})} \\
&\quad = 3\,.
\end{aligned}$$

The reader should verify that all of these probability ratios give 3/1. Similarly, the odds that a nonvirgin intermittent contraceptive user thinks that extramarital coitus is not always wrong is

$$\begin{aligned}
\frac{\Pr(\text{Not Always}|\text{Nonvirgin, intermittent use})}{\Pr(\text{Always Wrong}|\text{Nonvirgin, intermittent use})} &= (1/4)\Big/(1/4) \\
&= (2/80)\Big/(2/80) \\
&= 1\,.
\end{aligned}$$

The odds for virgin intermittent users are different than for nonvirgin intermittent users; thus, independence does not hold. For nonusers, the odds for both virgins and nonvirgins are 2, so independence holds. For regular users, the odds for both virgins and nonvirgins are 1, so again independence holds. Rather than checking for equality of the odds for virgins and nonvirgins, we could look at the ratio of the odds. If the odds ratio is one, then the odds are equal and conditional independence given a particular use holds.

1.2 Random Variables and Expectations

A *random variable* is simply a function from a set of outcomes to the real numbers. A *discrete random variable* is one that takes on values in a countable set. The *distribution* of a discrete random variable is a list of the possible values for the random variable along with the probabilities that the values will occur. The *expected value* of a random variable is a number that characterizes the middle of the distribution. For a random variable y with a discrete distribution, the expected value is

$$\mathrm{E}(y) = \sum_{\text{all } r} r\Pr(y = r)\,.$$

Distributions with the same expected value can be very different. For example, the expected value indicates the middle of a distribution but

does not indicate how spread out it is. The *variance* is a measure of how spread out a distribution is from its expected value. Let $\mathrm{E}(y) = \mu$, then the variance of y is

$$\mathrm{Var}(y) = \sum_{\text{all } r} (r - \mu)^2 \mathrm{Pr}(y = r) .$$

One problem with the variance is that it is measured on the wrong scale. If y is measured in meters, $\mathrm{Var}(y)$ involves the terms $(r - \mu)^2$; hence, it is measured in meters squared. To get things back on a comparable scale, we consider the *standard deviation* of y

$$\text{Std. dev.}(y) = \sqrt{\mathrm{Var}(y)} .$$

Standard deviations and variances are useful as measures of the relative dispersions of different random variables. The actual numbers themselves do not mean much. Moreover, there are other equally good measures of dispersion that can give results that are inconsistent with these. One reason standard deviations and variances are so widely used is because they are convenient mathematically. Of particular importance in applied work is the fact that the commonly used normal (Gaussian) distributions are completely characterized by their expected values (means) and variances. With these two numbers, one knows everything about a normal distribution. Normal distributions are widely used in statistics, so variances and their cousins, standard deviations, are also widely used.

The *covariance* is a measure of the linear relationship between two random variables. Suppose y_1 and y_2 are random variables. Let $\mathrm{E}(y_1) = \mu_1$ and $\mathrm{E}(y_2) = \mu_2$. The covariance between y_1 and y_2 is

$$\mathrm{Cov}(y_1, y_2) = \sum_{\text{all } (r,s)} (r - \mu_1)(s - \mu_2) \mathrm{Pr}(y_1 = r, y_2 = s) .$$

It is immediate that

$$\mathrm{Var}(y_1) = \mathrm{Cov}(y_1, y_1) .$$

In an attempt to get a handle on what the numerical value of the covariance means, it is often rescaled into a *correlation coefficient*.

$$\mathrm{Corr}(y_1, y_2) = \mathrm{Cov}(y_1, y_2) \Big/ \sqrt{\mathrm{Var}(y_1)\mathrm{Var}(y_2)} .$$

A perfect increasing linear relationship is indicated by a 1. A perfect decreasing linear relationship gives a -1. The absence of any linear relationship is indicated by a value of 0.

Exercise 1.6.5 contains important results on the expected values, variances, and covariances of linear combination of random variables.

1.3 The Binomial Distribution

There are a few distributions that are used in the vast majority of statistical applications. The reason for this is that they tend to occur naturally. The normal distribution is one. It occurs in practice because the central limit theorem dictates that other distributions will approach the normal. Two other distributions, the binomial and the multinomial, occur in practice because they are so simple. A fourth distribution, the Poisson, also occurs in nature because it is the distribution arrived at in another limit theorem. In this section, we discuss the binomial. Subsequent sections discuss the multinomial and the Poisson.

If you have independent identical trials and are counting how often something (anything) occurs, the appropriate distribution is the binomial. What could be simpler? Typically, the outcome of interest is referred to as a success. If the probability of a success is p in each of N independent identical trials, then the number of successes n has a binomial distribution with parameters N and p. Write

$$n \sim \text{Bin}(N, p)\,.$$

The distribution of n is

$$\Pr(n = r) = \binom{N}{r} p^r (1-p)^{N-r}$$

for $r = 0, 1, \ldots, N$. Here,

$$\binom{N}{r} = \frac{N!}{r!(N-r)!}$$

and for any positive integer m, $m! = m(m-1)(m-2)\cdots(2)(1)$.

Given the distribution, we can find the mean (expected value) and variance. By definition, the mean is

$$\text{E}(n) = \sum_{r=0}^{N} r \binom{N}{r} p^r (1-p)^{N-r}\,.$$

By writing n as the sum of N independent $\text{Bin}(1, p)$ random variables and using Exercise 1.6.5a, it is easily seen that

$$\text{E}(n) = Np\,.$$

The variance of n is

$$\text{Var}(n) = \sum_{r=0}^{N} (r - Np)^2 \binom{N}{r} p^r (1-p)^{N-r}\,.$$

Again, by writing n as the sum of N independent $\text{Bin}(1,p)$ random variables and now using Exercise 1.6.5b, it is easily seen that

$$\text{Var}(n) = Np(1-p)\,.$$

In this book, we will often need to look at both the number of successes and the number of failures at the same time. If the number of successes is n_1 and the number of failures is n_2, then

$$\begin{aligned} n_2 &= N - n_1 \\ n_1 &\sim \text{Bin}(N,p) \end{aligned}$$

and

$$n_2 \sim \text{Bin}(N, 1-p)\,.$$

The last result holds because, with independent identical trials, the number of outcomes that we call failures must also have a binomial distribution. If p is the probability of success, the probability of failure is $1-p$. Of course,

$$\begin{aligned} \text{E}(n_2) &= N(1-p) \\ \text{Var}(n_2) &= N(1-p)p\,. \end{aligned}$$

Note that $\text{Var}(n_1) = \text{Var}(n_2)$, regardless of the value of p. Finally,

$$\text{Cov}(n_1, n_2) = -Np(1-p)$$

and

$$\text{Corr}(n_1, n_2) = -1\,.$$

There is a perfect linear relationship between n_1 and n_2. If n_1 goes up one unit, n_2 goes down one unit. When we look at both successes and failures, write

$$(n_1, n_2) \sim \text{Bin}\big(N, p, (1-p)\big)\,.$$

1.4 The Multinomial Distribution

The multinomial distribution is a generalization of the binomial to more than two categories. Suppose we have N independent identical trials. On each trial, we check to see which of q events occurs. In such a situation, we assume that on each trial, one of the q events must occur. Let n_i, $i = 1, \ldots, q$, be the number of times that the ith event occurs. Let p_i be the probability that the ith event occurs on any trial. Note that the p_i's must satisfy $p_1 + p_2 + \cdots + p_q = 1$. In this situation, we say that $(n_1, \ldots, n_q)$ has a multinomial distribution with parameters $N, p_1, \ldots, p_q$. Write

$$(n_1, \ldots, n_q) \sim \text{Mult}(N, p_1, \ldots, p_q)\,.$$

The distribution is

$$\begin{aligned}\Pr(n_1 = r_1, \ldots, n_q = r_q) &= \frac{N!}{r_1! \cdots r_q!} p_1^{r_1} \cdots p_q^{r_q} \\ &= \frac{N!}{\prod_{i=1}^{q} r_i!} \prod_{i=1}^{q} p_i^{r_i}\end{aligned}$$

for $r_i \geq 0$ and $r_1 + \cdots + r_q = N$. Note that if $q = 2$, this is just a binomial distribution. In general, each individual component is

$$n_i \sim \text{Bin}(N, p_i)$$

so

$$\text{E}(n_i) = Np_i$$

and

$$\text{Var}(n_i) = Np_i(1 - p_i)\,.$$

Also, it can be shown that

$$\text{Cov}(n_i, n_j) = -Np_ip_j \qquad \text{for} \qquad i \neq j\,.$$

EXAMPLE 1.4.1. In Example 1.1.4, probabilities were given for the eight categories determined by combining high and low socioeconomic status, liberal and conservative political philosophy, and Democratic and Republican political affiliation. Suppose a sample of 50 individuals was taken from a population that had the probabilities associated with Example 1.1.4,

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	.12	.12	.04	.12	.4
Low	.18	.18	.06	.18	.6

The number of individuals falling into each of the eight categories has a multinomial distribution with $N = 50$ and these p_i's. The expected numbers of observations for each category are given by Np_i. It is easily seen that the expected counts for the cells are

	Democrat		Republican	
	Liberal	Conservative	Liberal	Conservative
High	6	6	2	6
Low	9	9	3	9

Note that the expected counts need not be integers.

The variance for, say, the number of high liberal Republicans is $50(.04)(1 - .04) = 1.92$. The variance of the number of high liberal

Democrats is $50(.12)(1-.12) = 5.28$. The covariance between the number of high liberal Republicans and the number of high liberal Democrats is $-50(.04)(.12) = -.24$. The correlation between the numbers of high liberal Democrats and Republicans is $-.24/\sqrt{(1.92)(5.28)} = -0.075$.

Now, suppose that the 50 individuals fall into the categories as listed in the table below.

	Democrat		Republican	
	Liberal	Conservative	Liberal	Conservative
High	5	7	4	6
Low	8	7	3	10

The probability of getting this particular table is

$$\frac{50!}{5!7!4!6!8!7!3!10!}(.12)^5(.12)^7(.04)^4(.12)^6(.18)^8(.18)^7(.06)^3(.18)^{10} = .000007.$$

The fact that this is a very small number is not surprising. There are a lot of possible tables, so the probability of getting any particular one is small. In fact, the table that has the highest probability can be shown to have a probability of only .000142. Although this probability is also very small, it is more than 20 times larger than the probability of the table given above.

Product-Multinomial Distributions

For $i = 1, \ldots, t$, take independent multinomials where the ith has s_i possible outcomes, i.e.,

$$(n_{i1}, \ldots, n_{is_i}) \sim \text{Mult}(N_i, p_{i1}, \ldots, p_{is_i})\,;$$

then we say that the n_{ij}'s have a product-multinomial distribution. By independence, the probability of any set of outcomes, say $\Pr(n_{ij} = r_{ij} \text{ all } i, j)$, is the product of the multinomial probabilities for each i. In other notation,

$$\Pr(n_{ij} = r_{ij} \text{ all } i, j) = \prod_{i=1}^{t} \Pr(n_{ij} = r_{ij} \text{ all } j)$$

and for $r_{ij} \geq 0$, $j = 1, \ldots, s_i$, with $\sum_{j=1}^{s_i} r_{ij} = N_i$, we have

$$\Pr(n_{ij} = r_{ij} \text{ all } j) = \left(N_i! \Big/ \prod_{j=1}^{s_i} r_{ij}! \right) \prod_{j=1}^{s_i} (p_{ij})^{r_{ij}}\,.$$

Thus,

$$\Pr(n_{ij} = r_{ij} \text{ all } i, j) = \prod_{i=1}^{t} \left(N_i! \Big/ \prod_{j=1}^{s_i} r_{ij}! \right) \prod_{j=1}^{s_i} (p_{ij})^{r_{ij}}\,,$$

where $r_{ij} \geq 0$ all i, j and $r_{i1} + \cdots + r_{is_i} = N_i$ for all i. Expected values, variances, and covariances within a particular multinomial are obtained by ignoring the other multinomials. Covariances between counts in different multinomials are zero because such observations are independent.

EXAMPLE 1.4.2. In Example 1.4.1 we considered taking a sample of 50 people from a population with the probabilities given in Example 1.1.4. Suppose we can identify and sample two subpopulations, the high socioeconomic group and the low socioeconomic group. If we take independent random samples of 30 from the high group and 20 from the low group, the numbers of individuals in the eight categories has a product-multinomial distribution with $t = 2$, $N_1 = 30$, $s_1 = 4$, $N_2 = 20$, and $s_2 = 4$. The probabilities of the four categories associated with high socioeconomic status are the conditional probabilities given high status. For example, the probability of a liberal Republican in the high group is $.04/.4 = .1$; the probability of a liberal Democrat is $.12/.4 = .3$. Similarly, the probability of a liberal Republican in the low socioeconomic group is $.06/.6 = .1$. The table of probabilities appropriate for the product-multinomial sampling described is the table of conditional probabilities given socioeconomic status:

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	.3	.3	.1	.3	1.0
Low	.3	.3	.1	.3	1.0

Although the probabilities for each category are the same for both high and low status, this is just an oddity of the particular example under consideration. Typically, the probabilities will be different in the different groups. In fact, the different groups do not even need to be divided into the same categories, although in most of our applications, the categories will be identical for all groups.

The expected counts for cells are computed separately for each multinomial. The expected count for high liberal Republicans is $30(.1) = 3$. With samples of 30 from the high group and 20 from the low group, the expected counts are

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	9	9	3	9	30
Low	6	6	2	6	20

Similarly, variances and covariances are found for each multinomial separately. The variance of the number of high liberal Republicans is $30(.1)(1 - .1) = 2.7$. The covariance between the numbers of low liberal Democrats and low liberal Republicans is $-20(.3)(.1) = -0.6$. The covariance between

counts in different multinomials is zero because counts in different multinomials are independent, e.g., the covariance between the numbers of high liberal Democrats and low liberal Republicans is zero because all high status counts are independent of all low status counts.

To find the probability of any particular table, find the probability associated with the high group and multiply it by the probability of the low group. The probability of getting the table

	Democrat		Republican		
	Liberal	Conservative	Liberal	Conservative	Total
High	10	10	2	8	30
Low	5	8	1	6	20

is

$$\left[\frac{30!}{10!10!2!8!}(.3)^{10}(.3)^{10}(.1)^{2}(.3)^{8}\right]\left[\frac{20!}{5!8!1!6!}(.3)^{5}(.3)^{8}(.1)^{1}(.3)^{6}\right]$$
$$= (.045716)(.008117) = .000371\,.$$

Exercise 1.1. Find the expected counts for a sample of size 20 from the population with probabilities given in Example 1.1.3. Now, conditioning on residence, find the expected counts for a sample of size 8 from Montana and a sample of size 12 from Haiti.

1.5 The Poisson Distribution

The binomial and multinomial distributions are appropriate and useful when the number of trials are not too large (whatever that means) and the probabilities of occurrences are not too small. For phenomena that have a very small probability of occurring on any particular trial, but for which an extremely large number of trials are available, the Poisson distribution is appropriate. For example, the number of suicides in a year might have a Poisson distribution. The probability of anyone committing suicide is very small, but in a large population, a substantial number of people will do it.

One of the most famous examples of a Poisson distribution is due to Bortkiewicz (1898). He examines the yearly total of men in the Prussian army corps who were kicked by horses and died of their injuries. Again, the probability that any individual will be mortally hoofed on a given day is very small, but for an entire army corps over an entire year, the number is fairly substantial. In particular, Fisher (1925) cites the 200 observations from 10 corps over a 20-year period as:

Deaths	0	1	2	3	4	5+
Frequencies	109	65	22	3	1	0

The idea is to view these as the results of a random sample of size 200 from a Poisson distribution. (Incidentally, the identity of the individual who introduced this example is one of the compelling mysteries in the history of statistics. It has been ascribed to at least four different people: Bortkiewicz, Bortkewicz, Bortkewitsch, and Bortkewitch.)

A third example of Poisson sampling is the number of microorganisms in a solution. One can imagine dividing the solution into a very large number of hypothetical units with very small volume (i.e., just big enough for the microorganism to be contained in the unit). If microorganisms are rare in the solution, then the probability of getting an organism in any particular unit is very small. Now, if we extract say one cubic centimeter of the solution, we have a very large number of trials. The number of organisms in the extracted solution should follow a Poisson distribution. Note that the Poisson distribution depends on having relatively few organisms in the solution. If that assumption is not true, one can dilute the solution until it is true.

Finally, and perhaps most importantly, the number of people who arrive during a 5-minute period to buy tickets for a Bruce Springsteen concert can be modeled with a Poisson distribution. Time can be divided into arbitrarily small intervals. The probability that anyone in the population will show up during any particular interval is very small. However, in 5 minutes there are a very large number of intervals.

The Poisson distribution can be arrived at as the limit of a $\text{Bin}(N, p)$ distribution where $N \to \infty$ and $p \to 0$. However, the two convergences must occur in such a way that $Np \to \lambda$. (To do this rigorously, we would let p be a function of N, say p_N.) The value λ is the parameter of the Poisson distribution. If n is a random variable with a Poisson distribution and parameter λ, write

$$n \sim \text{Pois}(\lambda)\,.$$

The distribution is defined by giving the probabilities and outcomes, i.e.,

$$\Pr(n = r) = \lambda^r e^{-\lambda}/r! \tag{1}$$

for $r = 0, 1, \ldots$.

It is not difficult to arrive at (1) by looking at binomial probabilities. The corresponding binomial probability for $n = r$ is

$$\binom{N}{r} p^r (1-p)^{N-r} = [(Np)^r (1-p)^N / r!](1-p)^{-r} \frac{N!}{(N-r)!N^r}\,. \tag{2}$$

With $N \to \infty$, $p \to 0$, and $Np \to \lambda$,

$$\begin{aligned} (Np)^r &\to \lambda^r \\ (1-p)^N &\to e^{-\lambda} \\ (1-p)^{-r} &\to 1 \\ N!/(N-r)!N^r &\to 1 \end{aligned}$$

Substituting these limits into the right-hand side of (2) gives the probability displayed in (1).

Using (1), we can compute the expected value and the variance of n. It is not difficult to show that

$$\mathrm{E}(n) = \lambda$$

and

$$\mathrm{Var}(n) = \lambda\,.$$

Lindgren (1993) gives a more detailed discussion of the assumptions behind the Poisson model. Fisher (1925) gives a nice discussion of the uses of the Poisson model and the analysis of Poisson data.

We close with two facts about independent Poisson random variables. If $n_1, \ldots, n_q$ are independent with $n_i \sim \mathrm{Pois}(\lambda_i)$, then the total of all the counts is

$$n_1 + n_2 + \cdots + n_q \;\sim\; \mathrm{Pois}(\lambda_1 + \cdots + \lambda_q)$$

and the counts given the total are

$$(n_1, \ldots, n_q)|N \;\sim\; \mathrm{Mult}(N, p_1, \ldots, p_q)$$

where

$$N = n_1 + \cdots + n_q$$

and

$$p_i = \lambda_i/(\lambda_1 + \cdots + \lambda_q), \quad i = 1, \ldots, q\,.$$

The conditional distribution is important for the analysis of log-linear models. If we have a table of counts that is comprised of independent Poisson random variables, we can always compute the grand total for the table. Looking at the conditional distribution given the total leads us to an analysis based on a multinomial distribution. The multinomial is the most commonly assumed distribution for tables of counts. Our discussion will focus almost entirely on multinomial and product-multinomial sampling.

1.6 Exercises

Exercise 1.6.1. In a *Newsweek* article on "The Wisdom of Animals" (May 23, 1988), one of the key issues considered was whether animals (other than humans) understand relationships between symbols. Some animals can associate symbols with objects; the question is whether they can tell the difference between commands such as "take the red ball to the blue ball" and "take the blue ball to the red ball." In discussing sea lions, it was indicated that out of a large pool of objects, they correctly identify symbols 95% of the time but are only correct 45% of the time on relationships. Presumably, this referred to a simple relationship between two objects; for

example, a sea lion could be shown symbols for "blue ball," "take to," "red ball." It was then concluded that, "considering the number of objects present in the pool, the odds are exceedingly long of getting even that proportion [45%] right by sheer chance." Assume a simple model in which sea lions know the nature of the relationship (it is repeated in a long series of trials), e.g., take one object to another, but make independent choices for identifying each object and the order in the relationship. Assume also that they have no idea what the correct order should be in the relationship, i.e., the two possible orders are equally probable. Compute the probability a sea lion will perform the task correctly. Why is the conclusion given in the article wrong? What does the number of objects present in the pool have to do with all this?

EXERCISE 1.6.2. Consider a 2×2 table of multinomial probabilities that models how subjects respond on two separate occasions.

	First Trial	
Second Trial	A	B
A	p_{11}	p_{12}
B	p_{21}	p_{22}

Show that

$$\Pr(A \text{ Second Trial } | B \text{ First Trial }) = \Pr(B \text{ Second Trial } | A \text{ First Trial })$$

if and only if the event that a change occurs between the first and second trials is independent of the outcome on the first trial.

EXERCISE 1.6.3. Weisberg (1975) reports the following data on the number of boys among the first seven children born to a collection of 1,334 Swedish ministers.

Number of Boys	0	1	2	3	4	5	6	7
Frequency	6	57	206	362	365	256	69	13

Assume that the number of boys has a Bin$(7, .5)$ distribution. Compute the probabilities for each of the eight categories $0, 1, \ldots, 7$. From the sample of 1,334 families, what is the expected frequency for each category? What is the distribution of the number of families that fall into each category? Summarize the fit of the assumed binomial model by computing

$$X^2 = \sum_{i=0}^{7} \frac{(Observation_i - Expected_i)^2}{Expected_i}.$$

The statistic X^2 is known as Pearson's chi-square statistic. For large samples such as this, if the *Expected* values are correct, X^2 should be one observation from a $\chi^2(7)$ distribution. (The 7 is one less than the number

of categories.) Compute X^2 and compare it to tabled values of the $\chi^2(7)$ distribution. Does X^2 seem like it could reasonably come from a $\chi^2(7)$? What does this imply about how well the binomial model fits the data? Can you distinguish which assumptions made in the binomial model may be violated?

EXERCISE 1.6.4. The data given in the previous problem may be 1,334 independent observations from a $\text{Bin}(7, p)$ distribution. If so, use the defining assumptions of the binomial distribution to show that this is the same as one observation from a $\text{Bin}(1334 \times 7, p)$ distribution. Estimate p with

$$\hat{p} = \frac{\text{Total number of boys}}{\text{Total number of trials}}.$$

Repeat the previous problem, replacing .5 with $\hat{p}$. Compare X^2 to a $\chi^2(6)$ distribution, reducing the degrees of freedom by one because the probability p is being estimated from the data.

EXERCISE 1.6.5. Let y_1, y_2, y_3, and y_4 be random variables and let a_1, a_2, a_3, and a_4 be real numbers. Show that the following relationships hold for finite discrete distributions.

(a) $\text{E}(a_1y_1 + a_2y_2 + a_3) = a_1\text{E}(y_1) + a_2\text{E}(y_2) + a_3$.

(b) $\text{Var}(a_1y_1 + a_2y_2 + a_3) = a_1^2\text{Var}(y_1) + a_2^2\text{Var}(y_2)$ for y_1 and y_2 independent.

(c) $\text{Cov}(a_1y_1 + a_2y_2, a_3y_3 + a_4y_4) = \sum_{i=1}^{2}\sum_{j=3}^{4} a_ia_j\text{Cov}(y_i, y_j)$.

EXERCISE 1.6.6. Assume that

$$(n_1, \ldots, n_q) \sim \text{Mult}(N, p_1, \ldots, p_q)$$

and let t be an integer less than q. Define $y = n_1 + \cdots + n_t$ and $\tilde{p} = p_1 + \cdots + p_t$. Show that

$$(y, n_{t+1}, \ldots, n_q) \sim \text{Mult}(N, \tilde{p}, p_{t+1}, \ldots, p_q)$$

so that

$$\text{E}(y) = N\tilde{p}$$

and

$$\text{Var}(y) = N\tilde{p}(1 - \tilde{p}).$$

EXERCISE 1.6.7. Suppose $y \sim \text{Bin}(N, p)$. Let $\hat{p} = y/N$. Show that $\text{E}(\hat{p}) = p$ and that $\text{Var}(\hat{p}) = p(1 - p)/N$.

2

Two-Dimensional Tables and Simple Logistic Regression

At this point, it is not our primary intention to provide a rigorous account of logistic regression and log-linear model theory. Such a treatment demands extensive use of advanced calculus and asymptotic theory. On the other hand, some knowledge of the basic issues is necessary for a correct understanding of *applications of logistic regression and log-linear models*. In this chapter, we address these basic issues for the simple case of two-dimensional tables and simple logistic regression. For a more elementary discussion of two-dimensional tables and simple logistic regression including substantial data analysis, see Christensen (1996a, Chapter 8). In fact, we assume that the reader is familiar with such analyses and use the topics in this chapter primarily to introduce key theoretical ideas.

2.1 Two Independent Binomials

Consider two binomials arranged in a 2×2 table. Our interest is in examining possible differences between the two binomials.

EXAMPLE 2.1.1. A survey was conducted to examine the relative attitudes of males and females about abortion. Of 500 females, 309 supported legalized abortion. Of 600 males, 319 supported legalized abortion. The data can be summarized in tabular form:

OBSERVED VALUES

	Support	Do Not Support	Total
Female	309	191	500
Male	319	281	600
Total	628	472	1,100

Note that the totals on the right-hand side of the table (500 and 600) are fixed by the design of the study. The totals along the bottom of the table are observed random variables. It is assumed that for each sex, the numbers of supporters and nonsupporters form a binomial random vector (ordered pair). We are interested in whether these numbers indicate that a person's sex affects their attitude toward legalized abortion. Note that the categories are Support and Do Not Support legalized abortion. Not supporting legalized abortion is distinct from opposing it. Anyone who is indifferent neither supports nor opposes legalized abortion.

We now introduce the notation that will be used for tables of counts in this book. For a 2×2 table, the observed values are denoted by n_{ij}, $i = 1, 2$ and $j = 1, 2$. Marginal totals are written $n_{i\cdot} \equiv n_{i1} + n_{i2}$ and $n_{\cdot j} \equiv n_{1j} + n_{2j}$. The total of all observations is $n_{\cdot\cdot} \equiv n_{11} + n_{12} + n_{21} + n_{22}$. The probability of having an observation fall in the ith row and jth column of the table is denoted p_{ij}. The number of observations that one would expect to see in the ith row and jth column (based on some statistical model) is denoted m_{ij}. For independent binomial rows, $m_{ij} = n_{i\cdot} p_{ij}$. Marginal totals $p_{i\cdot}$, $p_{\cdot j}$, $m_{i\cdot}$, and $m_{\cdot j}$ are defined like $n_{i\cdot}$ and $n_{\cdot j}$.

All of this notation can be summarized in tabular form.

OBSERVED VALUES

		Columns 1	2	Totals
Rows	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
	Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

PROBABILITIES

		Columns 1	2	Totals
Rows	1	p_{11}	p_{12}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$p_{2\cdot}$
	Totals	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot\cdot}$

EXPECTED VALUES

		Columns		
		1	2	Totals
Rows	1	m_{11}	m_{12}	$m_{1\cdot}$
	2	m_{21}	m_{22}	$m_{2\cdot}$
	Totals	$m_{\cdot 1}$	$m_{\cdot 2}$	$m_{\cdot\cdot}$

Our interest is in finding estimates of the p_{ij}'s, developing models for the p_{ij}'s, and performing tests on the p_{ij}'s. Equivalently, we can concern ourselves with estimates, models, and tests for the m_{ij}'s.

In Example 2.1.1, our interest is in whether sex is related to support for legalized abortion. Note that $p_{11}+p_{12}=1$ and $p_{21}+p_{22}=1$. (Equivalently $m_{11}+m_{12}=500$ and $m_{21}+m_{22}=600$.) If sex has no effect on opinion, the distribution of support versus nonsupport should be the same for both sexes. In particular, it is of interest to test the null hypothesis (model)

$$H_0 : p_{11} = p_{21} \text{ and } p_{12} = p_{22}\,.$$

With $p_{i1}+p_{i2}=1$, the equality $p_{11}=p_{21}$ holds if and only if $p_{12}=p_{22}$ holds. In other words, females and males have the same probability of "support" if and only if they have the same probability for "do not support." It suffices to test that the probability of support is the same for both sexes, i.e.,

$$H_0 : p_{11} = p_{21}$$

or, equivalently,

$$H_0 : p_{11} - p_{21} = 0\,.$$

To test this hypothesis, we need an estimate of $p_{11}-p_{21}$ and the standard error (SE) of the estimate. Each row is binomial with sample size $n_{i\cdot}$, so a natural estimate of p_{ij} is the proportion of observations falling in cell ij relative to the total number of observations in the ith row, i.e.,

$$\hat{p}_{ij} = n_{ij}/n_{i\cdot}\,.$$

For the abortion example, $\hat{p}_{11}=309/500$ and $\hat{p}_{21}=319/600$. The estimate of $p_{11}-p_{21}$ is

$$\hat{p}_{11}-\hat{p}_{21} = (n_{11}/n_{1\cdot}) - (n_{21}/n_{2\cdot})\,.$$

The two rows of the table were sampled independently so the variance of $\hat{p}_{11}-\hat{p}_{21}$ is

$$\begin{aligned}\operatorname{Var}(\hat{p}_{11}-\hat{p}_{21}) &= \operatorname{Var}(\hat{p}_{11}) + \operatorname{Var}(\hat{p}_{21}) \\ &= p_{11}p_{12}/n_{1\cdot} + p_{21}p_{22}/n_{2\cdot}\,,\end{aligned}$$

cf. Exercise 1.6.7. Finally,

$$\operatorname{SE}(\hat{p}_{11}-\hat{p}_{21}) = \sqrt{\hat{p}_{11}\hat{p}_{12}/n_{1\cdot} + \hat{p}_{21}\hat{p}_{22}/n_{2\cdot}}\,.$$

For the abortion example,

$$\mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(309/500)(191/500)}{500} + \frac{(319/600)(281/600)}{600}} = .0298\,.$$

One other thing is required before we can perform a test. We need to know the distribution of $[(\hat{p}_{11} - \hat{p}_{21}) - (p_{11} - p_{21})]/\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})$. By appealing to the Central Limit Theorem and the Law of Large Numbers (cf. Lindgren, 1993), if $n_{1\cdot}$ and $n_{2\cdot}$ are large, we can use the approximate distribution

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - (p_{11} - p_{21})}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} \sim N(0,1)\,.$$

To perform a test of

$$H_0 : p_{11} - p_{21} = 0$$

versus

$$H_A : p_{11} - p_{21} \neq 0,$$

assume that H_0 is true and look for evidence against H_0. If H_0 is true, the approximate distribution is

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} \sim N(0,1)\,.$$

If the alternative hypothesis is true, $\hat{p}_{11} - \hat{p}_{21}$ still estimates $p_{11} - p_{21}$ so the test statistic

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})}$$

tends to be either a large positive value if $p_{11} - p_{21} > 0$ or a large negative value if $p_{11} - p_{21} < 0$. An $\alpha = .05$ level test rejects H_0 if

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} > 1.96$$

or if

$$\frac{(\hat{p}_{11} - \hat{p}_{21}) - 0}{\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})} < -1.96\,.$$

The values -1.96 and 1.96 cut off the probability .025 from the bottom and top of a $N(0,1)$ distribution, respectively. Thus, the total probability of rejecting H_0 when H_0 is true is $.025 + .025 = .05$. Recall that this test is based on a large sample approximation to the distribution of the test statistic.

For the abortion example, the test statistic is

$$\frac{(309/500) - (319/600)}{.0298} = 2.90\,.$$

Because $2.90 > 1.96$, H_0 is rejected at the $\alpha = .05$ level. There is evidence of a relationship between sex and attitudes about legalized abortion. These data indicate that females are more likely to support legalized abortion.

Before leaving this test procedure, we mention an alternative method for computing $\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21})$. Recall that

$$\mathrm{Var}(\hat{p}_{11} - \hat{p}_{21}) = p_{11}p_{12}/n_{1\cdot} + p_{21}p_{22}/n_{2\cdot}\,.$$

If H_0 is true, $p_{11} = p_{21}$ and $p_{12} = p_{22}$. These facts can be used in estimating the variance of $\hat{p}_{11} - \hat{p}_{21}$. A pooled estimate of $p \equiv p_{11} = p_{21}$ is $(n_{11} + n_{21})/(n_{1\cdot} + n_{2\cdot}) = n_{\cdot 1}/n_{\cdot\cdot} = 628/1100$. A pooled estimate of $(1-p) \equiv p_{12} = p_{22}$ is $n_{\cdot 2}/n_{\cdot\cdot} = 472/1100$. This yields

$$\mathrm{Var}(\hat{p}_{11} - \hat{p}_{21}) = p(1-p)(1/n_{1\cdot} + 1/n_{2\cdot})$$

and

$$\begin{aligned}\mathrm{SE}(\hat{p}_{11} - \hat{p}_{21}) &= \sqrt{(628/1100)(472/1100)[(1/500) + (1/600)]} \\ &= .0300\,.\end{aligned}$$

The test statistic computed with the new standard error is

$$\frac{(309/500) - (319/600)}{.0300} = 2.87777\,.$$

For these data, the results are essentially the same.

The test procedures discussed above work nicely for two independent binomials, but, unfortunately, they do not generalize to more than two binomials or to situations in which there are more than two possible outcomes (e.g., support, oppose, no opinion). An alternative test procedure is based on what is known as the *Pearson chi-square test statistic*. This test is equivalent to the test given above using the pooled estimate of the standard error. Moreover, Pearson's chi-square is applicable in more general problems. The Pearson test statistic is based on comparing the observed table values in the 2×2 table with estimates of the expected values that are obtained assuming that H_0 is true.

In the abortion example, if H_0 is true, then $p = p_{11} = p_{21}$ and $\hat{p} = 628/1100$. Similarly, $(1-p) = p_{12} = p_{22}$ and $(1-\hat{p}) = 472/1100$. As before, the expected values are $m_{ij} = n_{i\cdot}p_{ij}$. The estimated expected values under H_0 are $\hat{m}_{ij}^{(0)} = n_{i\cdot}\hat{p}_{ij}$, where $\hat{p}_{ij}$ is $\hat{p}$ if $j = 1$ and $(1-\hat{p})$ if $j = 2$. More generally,

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}(n_{\cdot j}/n_{\cdot\cdot})\,. \tag{1}$$

The Pearson chi-square statistic is defined as

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}}\,.$$

If H_0 is true, then n_{ij} and $\hat{m}_{ij}^{(0)}$ should be near each other, and the terms $(n_{ij} - \hat{m}_{ij}^{(0)})^2$ should be reasonably small. If H_0 is not true, then the $\hat{m}_{ij}^{(0)}$'s, which are estimates based on the assumption that H_0 is true, should do a poor job of predicting the n_{ij}'s. The terms $(n_{ij} - \hat{m}_{ij}^{(0)})^2$ should be larger when H_0 is not true.

Note that a prediction $\hat{m}_{ij}^{(0)}$ that is, say, three away from the observed value n_{ij}, can be either a good prediction or a bad prediction depending on how large the value in the cell should be. If $n_{ij} = 4$ and $\hat{m}_{ij}^{(0)} = 1$, the prediction is poor. If $n_{ij} = 104$ and $\hat{m}_{ij}^{(0)} = 101$, the prediction is good. The $\hat{m}_{ij}^{(0)}$ in the denominator of each term of X^2 is a scale factor that corrects for this problem.

The hypothesis $H_0 : p_{11} = p_{21}$ and $p_{12} = p_{22}$ is rejected at the $\alpha = .05$ level if

$$X^2 \geq \chi^2(.95, 1)\,.$$

The test is based on the fact that if H_0 is true, then as $n_{1\cdot}$ and $n_{2\cdot}$ get large, X^2 has approximately a $\chi^2(1)$ distribution. This is a consequence of the Central Limit Theorem and the Law of Large Numbers, cf. Exercise 2.1.

For the abortion example

$\hat{m}_{ij}^{(0)}$	Support	Do Not Support	Totals
Female	285.5	214.5	500
Male	342.5	257.5	600
Totals	628	472	1100

$$X^2 = 8.2816,$$

$$\chi^2(.95, 1) = 3.84\,.$$

Because $8.2816 > 3.84$, the $\alpha = .05$ test rejects H_0.

Note that $8.2816 = (2.8777)^2$ and that $3.84 = (1.96)^2$. For 2×2 tables, the results of Pearson chi-square tests are exactly equivalent to the results of normal theory tests using the pooled estimate in the standard error. By definition, $\chi^2(1 - \alpha, 1) = [z(1 - \frac{\alpha}{2})]^2$ for $\alpha \in (0, .5]$. Also,

$$X^2 = \frac{(\hat{p}_{11} - \hat{p}_{21})^2}{\hat{p}(1 - \hat{p})\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)}\,. \tag{2}$$

Exercise 2.1. Prove equation (2).

By comparing the n_{ij}'s to the $\hat{m}_{ij}^{(0)}$'s, we can examine the nature of the differences in the two binomials. One simple way to do this comparison is to examine a table of *residuals*, i.e., the $(n_{ij} - \hat{m}_{ij}^{(0)})$'s. In order to make

accurate evaluations of how well $\hat{m}_{ij}^{(0)}$ is predicting n_{ij}, the residuals need to be rescaled or standardized. Define the *Pearson residuals* as

$$\tilde{r}_{ij} = \frac{n_{ij} - \hat{m}_{ij}^{(0)}}{\sqrt{\hat{m}_{ij}^{(0)}}},$$

$i = 1, 2$, $j = 1, 2$. Note that $X^2 = \sum_{ij} \tilde{r}_{ij}^2$. The Pearson residuals for the abortion data are

$\tilde{r}_{ij}$	Support	Do Not Support
Female	1.39	−1.60
Male	−1.27	1.46

The positive residual 1.39 indicates that more females support legalized abortion than would be expected under H_0. The negative residual −1.27 indicates that fewer males support abortion than would be expected under H_0. Together, the values 1.39 and −1.27 indicate that proportionately more females support legalized abortion than males. Equivalently, proportionately more males do not support legalized abortion than females.

2.1.1 The Odds Ratio

A commonly used technique in the analysis of count data is the examination of odds ratios. In the abortion example, the odds of females supporting legalized abortion is p_{11}/p_{12}. The odds of males supporting legalized abortion is p_{21}/p_{22}. The odds ratio is

$$\frac{(p_{11}/p_{12})}{(p_{21}/p_{22})} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Note that if the two binomials are identical, then $p_{11} = p_{21}$ and $p_{12} = p_{22}$, so

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = 1.$$

An alternative to using Pearson's chi-square for examining whether two binomials are the same is to examine the estimated odds ratio. Using $\hat{p}_{ij} = n_{ij}/n_{i\cdot}$ gives

$$\frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

For the abortion example, the estimate is

$$\frac{(309)(281)}{(191)(319)} = 1.425.$$

This is only an estimate of the population odds ratio, but it is fairly far from the target value of 1. In particular, we have estimated that the odds

of a female supporting legalized abortion are about one and a half times as large as the odds of a male supporting legalized abortion.

We may wish to test the hypothesis that the odds ratio equals 1. Equivalently, we can test whether the log of the odds ratio equals 0. The log odds ratio is $\log(1.425) = .354$. The asymptotic (large sample) standard error of the log odds ratio is

$$\begin{aligned}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} &= \sqrt{\frac{1}{309}+\frac{1}{191}+\frac{1}{319}+\frac{1}{281}} \\ &= .123\,.\end{aligned}$$

The estimate minus the hypothesized value over the standard error is

$$\frac{.354-0}{.123} = 2.88\,.$$

Comparing this to a $N(0,1)$ distribution indicates that the log-odds ratio is greater than zero, thus the odds ratio is greater than 1. Note that this test is not equivalent to the other tests considered, even though the numerical value of the test statistic is similar to the other normal theory tests.

A 95% confidence interval for the log odds ratio has the end points $.354 \pm 1.96(.123)$. This gives an interval $(.113, .595)$. The log odds ratio is in the interval $(.113, .595)$ if and only if the odds ratio is in the interval $(e^{.113}, e^{.595})$. Thus, a 95% confidence interval for the odds ratio is $(e^{.113}, e^{.595})$, which simplifies to $(1.12, 1.81)$. We are 95% confident that the true odds of women supporting legalized abortion is between 1.12 and 1.81 times greater than the odds of men supporting legalized abortion.

2.2 Testing Independence in a 2×2 Table

In Section 1, we obtained a 2×2 table by looking at two populations, each divided into two categories. In this section, we consider only one population but divide it into two categories in each of two different ways. The two different ways of dividing the population will be referred to as factors.

In Section 1, we examined differences between the two populations. In this section, we examine the nature of the one population being sampled. In particular, we examine whether the two factors affect the population independently or whether they interact to determine the nature of the population.

EXAMPLE 2.2.1. As part of a longitudinal study, a sample of 3182 people without cardiovascular disease were cross-classified by two factors: personality type and exercise. Personality type was categorized as type A or type B. Type A persons show signs of stress, uneasiness, and hyperactivity. Type

B persons are relaxed, easygoing, and normally active. Exercise is categorized as persons who exercise regularly and those who do not. The data are given in the following table:

		Personality		
	n_{ij}	A	B	Totals
Exercise	Regular	483	477	960
	Other	1101	1121	2222
	Totals	1584	1598	3182

Although notations for observations (n_{ij}'s), probabilities (p_{ij}'s), and expected values (m_{ij}'s) are identical to those in Section 1, the meaning of these quantities has changed. In Section 1, the rows were two independent binomials, so $p_{11}+p_{12} = 1 = p_{21}+p_{22}$. In this section, there is only one population, so the constraint on the probabilities is that $p_{11}+p_{12}+p_{21}+p_{22} = 1$.

In this section, our primary interest is in determining whether the row factor is independent of the column factor and if not, how the factors deviate from independence. The probability of an observation falling in the ith row and jth column of the table is p_{ij}. The probability of an observation falling in the ith row is $p_{i.}$. The probability of the jth column is $p_{.j}$. Rows and columns are independent if and only if for all i and j

$$p_{ij} = p_{i.}p_{.j}\,. \tag{1}$$

The sample size is $n_{..}$, so the expected counts in the table are

$$m_{ij} = n_{..}p_{ij}\,.$$

If rows and columns are independent, this becomes

$$m_{ij} = n_{..}p_{i.}p_{.j}\,. \tag{2}$$

It is easily seen that condition (1) for independence is equivalent to

$$m_{ij} = m_{i.}m_{.j}/n_{..}\,. \tag{3}$$

Pearson's chi-square can be used to test independence. Pearson's statistic is again

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}}$$

where $\hat{m}_{ij}^{(0)}$ is an estimate of m_{ij} based on the assumption that rows and columns are independent. If we take $\hat{m}_{i.} = n_{i.}$ and $\hat{m}_{.j} = n_{.j}$, then equation (3) leads to

$$\hat{m}_{ij}^{(0)} = n_{i.}n_{.j}/n_{..}\,. \tag{4}$$

Equation (4) can also be arrived at via equation (2). An obvious estimate of $p_{i\cdot}$ is

$$\hat{p}_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot} .$$

Similarly,

$$\hat{p}_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot} .$$

Substitution into equation (2) leads to equation (4). It is interesting to note that equation (4) is numerically identical to formula (2.1.1), which gives $\hat{m}_{ij}$ for two independent binomials. Just as in Section 1, for the purposes of testing, X^2 is compared to a $\chi^2(1)$ distribution. Pearson residuals are again defined as

$$\tilde{r}_{ij} = \frac{n_{ij} - \hat{m}_{ij}^{(0)}}{\sqrt{\hat{m}_{ij}^{(0)}}} .$$

For the personality-exercise data, we get

	$\hat{m}_{ij}^{(0)}$	Personality A	Personality B	Totals
Exercise	Regular	477.9	482.1	960
	Other	1106.1	1115.9	2222
	Totals	1584	1598	3182

$$X^2 = .156$$

The test is not significant for any reasonable α level. (The *P value* is greater than .5.) There is no significant evidence against independence of exercise and personality type. In other words, the data are consistent with the interpretation that knowledge of personality type gives no new information about exercise habits or, equivalently, knowledge of exercise habits gives no new information about personality type.

2.2.1 The Odds Ratio

Just as in examining the equality of two binomials, the odds ratio can be used to examine the independence of two factors in a multinomial sample. In the personality-exercise data, the odds that a person exercises regularly are $p_{1\cdot}/p_{2\cdot}$. In addition, the odds of exercising regularly can be examined separately for each personality type. For type A personalities, the odds are p_{11}/p_{21}, and for type B personalities, the odds are p_{12}/p_{22}. Intuitively, if exercise and personality types are independent, then the odds of regular exercise should be the same for both personality types. In particular, the ratio of the two sets of odds should be one.

Proposition 2.2.2. If rows and columns are independent, then the

odds ratio

$$\frac{(p_{11}/p_{21})}{(p_{12}/p_{22})} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

equals one.

Proof. By equation (1)

$$\frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{1\cdot}p_{\cdot 1}p_{2\cdot}p_{\cdot 2}}{p_{1\cdot}p_{\cdot 2}p_{2\cdot}p_{\cdot 1}} = 1.$$

□

If the odds ratio is estimated under the assumption of independence, $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = n_{i\cdot}n_{\cdot j}/(n_{\cdot\cdot})^2$; so the estimated odds ratio is always one. A more interesting approach is to estimate the odds ratio without assuming independence and then see how close the estimated odds ratio is to one. With this approach, $\hat{p}_{ij} = n_{ij}/n_{\cdot\cdot}$ and

$$\frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

In the personality-exercise example, the estimated odds ratio is

$$\frac{(483)(1121)}{(477)(1101)} = 1.03$$

which is very close to one. The log odds are .0305, the asymptotic standard error is $[1/483 + 1/477 + 1/1101 + 1/1121]^{1/2} = .0772$, and the test statistic for H_0 that the log odds equal 0 is $.0305/.0772 = .395$. Again, there is no evidence against independence.

EXERCISE 2.2. Give a 95% confidence interval for the odds ratio. Explain what the confidence interval means.

2.3 $I \times J$ Tables

The situation examined in Section 1 can be generalized to consider samples from I different populations, each of which is divided into J categories. We assume that the samples from different populations are independent and that each sample follows a multinomial distribution. This is *product-multinomial sampling*.

Similarly, a sample from one population that is categorized by two factors can be generalized beyond the case considered in Section 2. We allow one factor to have I categories and the other factor to have J categories. Between the two factors, the population is divided into a total of IJ categories. The distribution of counts within the IJ categories is assumed

to have a multinomial distribution. Consequently, this sampling scheme is multinomial sampling.

An $I \times J$ table of the observations n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, can be written

		Factor 2 (Categories)				
	n_{ij}	1	2	...	J	Totals
	1	n_{11}	n_{12}	...	n_{1J}	$n_{1\cdot}$
Factor 1	2	n_{21}	n_{22}	...	n_{2J}	$n_{2\cdot}$
(Populations)	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\cdot}$
	Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot J}$	$n_{\cdot\cdot}$

with similar tables for the probabilities p_{ij} and the expected values m_{ij}.

The analysis of product-multinomial sampling begins by testing whether all of the I multinomial populations are identical. In other words, we wish to test the model

$$H_0 : \; p_{1j} = p_{2j} = \cdots = p_{Ij} \text{ for all } j = 1, \ldots, J \,. \tag{1}$$

against the alternative

$$H_A : \text{ model (1) is not true} \,.$$

This is described as testing for *homogeneity of proportions*.

We continue to use Pearson's chi-square test statistic to evaluate the appropriateness of the null hypothesis model. Pearson's chi-square requires estimates of the expected values m_{ij}. Each sample i has a multinomial distribution with $n_{i\cdot}$ trials, so

$$m_{ij} = n_{i\cdot} p_{ij} \,.$$

If H_0 is true, p_{ij} is the same for all values of i. A pooled estimate of the common value of the p_{ij}'s is

$$\hat{p}_{ij}^{(0)} = n_{\cdot j}/n_{\cdot\cdot} \,.$$

In other words, if all the populations have the same probability for category j, an estimate of this common probability is the total number of observations in category j divided by the overall total number of observations. From this probability estimate we obtain

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}(n_{\cdot j}/n_{\cdot\cdot}) \,.$$

In both $\hat{p}_{ij}^{(0)}$ and $\hat{m}_{ij}^{(0)}$, the superscript (0) is used to indicate that the estimate was obtained under the assumption that H_0 was true. Pearson's

chi-square test statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}} .$$

For large samples, if H_0 is true, the approximation

$$X^2 \sim \chi^2((I-1)(J-1))$$

is valid. H_0 is rejected in an α level test if

$$X^2 > \chi^2(1-\alpha, (I-1)(J-1)) .$$

Note that if $I = J = 2$, these are precisely the results discussed in Section 1.

The analysis of a multinomial sample begins by testing for independence of the two factors. In particular, we wish to test the model

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J . \tag{2}$$

We again use Pearson's chi-square. The marginal probabilities are estimated as

$$\hat{p}_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}$$

and

$$\hat{p}_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot} .$$

Because $m_{ij} = n_{\cdot\cdot} p_{ij}$, if the model in (2) is true, we can estimate m_{ij} with

$$\begin{aligned} \hat{m}_{ij}^{(0)} &= n_{\cdot\cdot} \hat{p}_{i\cdot} \hat{p}_{\cdot j} \\ &= n_{\cdot\cdot} (n_{i\cdot}/n_{\cdot\cdot})(n_{\cdot j}/n_{\cdot\cdot}) \\ &= n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot} \end{aligned}$$

where the (0) in $\hat{m}_{ij}^{(0)}$ indicates that the estimate is obtained assuming that (2) holds. The Pearson chi-square test statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{m}_{ij}^{(0)}\right)^2}{\hat{m}_{ij}^{(0)}}$$

which, if (2) is true and the sample size is large, is approximately distributed as a $\chi^2\left((I-1)(J-1)\right)$. H_0 is rejected at the α level if

$$X^2 > \chi^2\left(1-\alpha, (I-1)(J-1)\right) .$$

Once again, if $I = J = 2$, we obtain the previous results given for 2×2 tables. Moreover, the test procedures for product-multinomial sampling and

for multinomial sampling are numerically identical. Only the interpretations of the tests differ.

EXAMPLE 2.3.1. Fifty-two males between the ages of 11 and 30 were operated on for knee injuries using arthroscopic surgery. The patients were classified by type of injury: twisted knee, direct blow, or both. The results of the surgery were classified as excellent (E), good (G), and fair or poor (F-P). These data can reasonably be viewed as either multinomial or product-multinomial. As a multinomial, we have 52 people cross-classified by type of injury and result of surgery. However, we can also think of the three types of injuries as defining different populations. Each person sampled from a population is given arthroscopic surgery and then the result is classified. Because our primary interest is in the result of surgery, we choose to *think* of the sampling as product-multinomial. The form of the analysis is identical for both sampling schemes. The data are

		Result			
	n_{ij}	E	G	F-P	Totals
Injury	Twist	21	11	4	36
	Direct	3	2	2	7
	Both	7	1	1	9
	Totals	31	14	7	52

The estimated expected counts under H_0 are

		Result			
	$\hat{m}_{ij}^{(0)}$	E	G	F-P	Totals
Injury	Twist	21.5	9.7	4.8	36
	Direct	4.2	1.9	.9	7
	Both	5.4	2.4	1.2	9
	Totals	31	14	7	52

with

$$X^2 = 3.229$$

and

$$df = (3-1)(3-1) = 4\,.$$

If the sample size is large, X^2 can be compared to a χ^2 distribution with four degrees of freedom. If we do this, the *P value* for the test is .52. Unfortunately, it is quite obvious that the sample size is not large. The number of observations in many of the cells of the table is small. This is a serious problem and aspects of the problem are discussed in Section 4, the subsection of Section 3.5 on Other Sampling Models, and Chapter 8. However, to the extent that this book focuses on distribution theory, it focuses on asymptotic distributions. For now, we merely state that in this

example, the n_{ij}'s and $\hat{m}_{ij}^{(0)}$'s are such that, when taken together with the very large P *value*, we feel safe in concluding that these data provide no evidence of different surgical results for the three types of injuries. (This conclusion is borne out by the fact that an exact small sample test yields a similar P *value*, cf. Section 3.5.)

2.3.1 Response Factors

In Example 2.3.1, the result of surgery can be thought of as a response, whereas the type of injury is used to explain the response. Similarly, in Example 2.1.1, opinions on abortions can be considered as a response and sex can be considered as an explanatory factor.

The existence of response factors is often closely tied to the sampling scheme. Product-multinomial sampling is commonly used with an independent multinomial sample taken for every combination of the explanatory factors and the categories of the multinomials being the categories of the response factors. This is illustrated in Example 2.1.1 where there are two independent multinomials (binomials), one for males and one for females. The categories for each multinomial are Support and Do Not Support legalized abortion. Example 3.5.2 in the next chapter involves two explanatory factors, Sex and Socioeconomic Status, and one response factor, Opinion on Legalized Abortion. Each of the four combinations obtained from the two sexes and the two statuses define an independent multinomial. In other words, there is a separate multinomial sample for each combination of sex and socioeconomic status. The categories of the response factor, Support and Do Not Support legalized abortion, are the categories of the multinomials.

More generally, the categories of a response factor can be cross-classified with other response factors or explanatory factors to yield the categories in a series of independent multinomials. This situation is of most interest when there are several factors involved. Some factors can be cross-classified to define the multinomial populations while other factors can be cross-classified with the response factors to define the categories of the multinomials. Example 2.3.1 illustrates the simplest case in which there is one explanatory factor, Injury, crossed with one response factor, Result, to define the categories of the multinomial. Both Injury and Result have three levels so the multinomial has a total of nine categories. With only two factors in the table, there can be only one multinomial sample because there are no other factors available to define various multinomial populations. Example 3.5.1 is more general in that it has two independent multinomials, one for each sex. Each multinomial has six categories. The categories are obtained by cross-classifying the explanatory factor, Political Party, having three levels, with the response factor, Abortion Opinion, having two levels.

In this more general sampling scheme, one often conditions on all factors other than the response so that the analysis is reduced to that of the original sampling scheme in which every combination of explanatory factors defines an independent multinomial. Again, this is illustrated in Example 2.3.1. While the sampling was multinomial, we treated the sampling as product-multinomial with an independent multinomial for each level of the Injury. The justification for treating the data as product-multinomial is that we conditioned on the Injury.

While the sampling techniques described above are probably the most commonly used, there are alternatives that are also commonly used. For example, in medicine a response factor is often some disease with levels that are various states of the disease. If the disease is at all rare, it may be impractical to sample different populations and see how many people fall into the various levels of the disease. In this case, one may need to take the disease levels as populations, sample from these populations, and investigate various characteristics of the populations. Thus the "explanatory" factors discussed above would be considered descriptive factors here. This sampling scheme is often called *retrospective* for obvious reasons. The other schemes discussed above are called *prospective*. These issues are discussed in more detail in the introduction to Chapter 4 and in Sections 4.7 and 11.7.

2.3.2 Odds Ratios

The null hypotheses (1) and (2) can be rewritten in terms of odds ratios.

Proposition 2.3.2. Under product-multinomial sampling $p_{1j} = \cdots = p_{Ij} > 0$ for all $j = 1, \ldots, J$ if and only if

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = 1$$

for all i, $i' = 1, \ldots, I$ and j, $j' = 1, \ldots, J$.

Proof. a) *Equality of probabilities across rows implies that the odds ratios equal one.* By substitution,

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = \frac{p_{ij}p_{ij'}}{p_{ij'}p_{ij}} = 1.$$

b) *All odds ratios equal to one implies equality of probabilities across rows.* Recall that $p_{i\cdot} = 1$ for all $i = 1, \ldots, I$, so that $p_{\cdot\cdot} = I$. In addition, $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1$ implies $p_{ij}p_{i'j'} = p_{ij'}p_{i'j}$. Note that

$$p_{ij} = p_{ij}p_{\cdot\cdot}/I = \frac{1}{I}\sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij}p_{i'j'}$$

$$
\begin{aligned}
&= \frac{1}{I}\sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij'}p_{i'j} \\
&= \frac{1}{I}\sum_{j'=1}^{J} p_{ij'}\sum_{i'=1}^{I} p_{i'j} \\
&= \frac{1}{I}\sum_{j'=1}^{J} p_{ij'}p_{\cdot j} \\
&= \frac{1}{I}p_{\cdot j}\sum_{j'=1}^{J} p_{ij'} \\
&= \frac{1}{I}p_{\cdot j}p_{i\cdot} \\
&= p_{\cdot j}/I\,.
\end{aligned}
$$

Because this holds for any i and j, $p_{\cdot j}/I = p_{1j} = p_{2j} = \cdots = p_{Ij}$ for $j = 1, \ldots, J$. □

Proposition 2.3.3. Under multinomial sampling, $0 < p_{ij} = p_{i\cdot}p_{\cdot j}$ for all $i = 1, \ldots, I$, $j = 1, \ldots, J$ if and only if

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = 1$$

for all $i, i' = 1, \ldots, I$ and $j, j' = 1, \ldots, J$.

Proof. a) *Independence implies that the odds ratios equal one.*

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = \frac{p_{i\cdot}p_{\cdot j}p_{i'\cdot}p_{\cdot j'}}{p_{i\cdot}p_{\cdot j'}p_{i'\cdot}p_{\cdot j}} = 1.$$

b) *All odds ratios equal to one implies independence.* If $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1$ for all i, i', j, and j', then $p_{ij}p_{i'j'} = p_{ij'}p_{i'j}$. Moreover, because $p_{\cdot\cdot} = 1$,

$$
\begin{aligned}
p_{ij} = p_{ij}p_{\cdot\cdot} = \sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij}p_{i'j'} &= \sum_{i'=1}^{I}\sum_{j'=1}^{J} p_{ij'}p_{i'j} \\
&= \sum_{i'=1}^{I} p_{i'j}\sum_{j'=1}^{J} p_{ij'} \\
&= \sum_{i'=1}^{I} p_{i'j}p_{i\cdot} \\
&= p_{i\cdot}\sum_{i'=1}^{I} p_{i'j}
\end{aligned}
$$

$$= \quad p_{i\cdot}p_{\cdot j} \qquad \square$$

There is a great deal of redundancy in specifying that

$$\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} = 1$$

for all i, i', j, and j'. For example, if $i = i'$, then $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = p_{ij}p_{ij'}/p_{ij'}p_{ij} = 1$ and no real restriction has been placed on the p_{ij}'s. A similar result occurs if $j = j'$. More significantly, if

$$p_{12}p_{23}/p_{13}p_{22} = 1$$

and

$$p_{12}p_{24}/p_{14}p_{22} = 1,$$

then

$$\begin{aligned} 1 &= (p_{12}p_{23}/p_{13}p_{22})(p_{14}p_{22}/p_{12}p_{24}) \\ &= p_{14}p_{23}/p_{13}p_{24}\,. \end{aligned}$$

In other words, the fact that two of the odds ratios equal one implies that a third odds ratio equals one. It turns out that the condition

$$\frac{p_{11}p_{ij}}{p_{1j}p_{i1}} = 1$$

for $i = 2, \ldots, I$ and $j = 2, \ldots, J$ is equivalent to the condition that all odds ratios equal one.

Proposition 2.3.4. $p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1$ for all i, i', j and j' if and only if $p_{11}p_{ij}/p_{1j}p_{i1} = 1$ for all $i \neq 1$, $j \neq 1$.

Proof. Clearly, if the odds ratios are one for all i, i', j, and j', then $p_{11}p_{ij}/p_{1j}p_{i1} = 1$ for all i and j. Conversely,

$$\begin{aligned} 1 &= \left(\frac{p_{11}p_{ij}}{p_{1j}p_{i1}}\right)\left(\frac{p_{11}p_{i'j'}}{p_{1j'}p_{i'1}}\right) \Big/ \left(\frac{p_{11}p_{ij'}}{p_{1j'}p_{i1}}\right)\left(\frac{p_{11}p_{i'j}}{p_{1j}p_{i'1}}\right) \\ &= \frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}} \end{aligned}$$

$\square$

Of course, in practice the p_{ij}'s are never known. We can investigate independence by examining the estimated odds ratios

$$\hat{p}_{ij}\hat{p}_{i'j'}/\hat{p}_{ij'}\hat{p}_{i'j} = n_{ij}n_{i'j'}/n_{ij'}n_{i'j}$$

or, equivalently, we can look at the log of this. For large samples, the log of the estimated odds ratio is normally distributed with standard error

$$\text{SE} = \sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{ij'}} + \frac{1}{n_{i'j}} + \frac{1}{n_{i'j'}}}\,.$$

This allows the construction of asymptotic tests and confidence intervals for the log odds ratio. Of particular interest is the hypothesis

$$H_0 : p_{ij}p_{i'j'}/p_{ij'}p_{i'j} = 1\,.$$

After taking logs, this becomes

$$H_0 : \log(p_{ij}p_{i'j'}/p_{ij'}p_{i'j}) = 0\,.$$

EXAMPLE 2.3.5. We continue with the knee injury data of Example 2.3.1. From Proposition 2.3.4, it is sufficient to examine

$$\begin{aligned}
\frac{n_{11}n_{22}}{n_{12}n_{21}} = 21(2)/11(3) &= 1.27\,,\\
\frac{n_{11}n_{23}}{n_{13}n_{21}} = 21(2)/4(3) &= 3.5\,,\\
\frac{n_{11}n_{32}}{n_{12}n_{31}} = 21(1)/11(7) &= .27\,,\\
\frac{n_{11}n_{33}}{n_{13}n_{31}} = 21(1)/4(7) &= .75\,.
\end{aligned}$$

Although the X^2 test indicated no difference in the populations (populations = injury types), *at least* two of these estimated odds ratios *seem* substantially different from 1. In particular, relative to having an F-P result, the odds of an excellent result are about 3.5 times larger with twist injuries than with direct blows. Also, relative to having a good result, the odds of an excellent result from a twisted knee are only .27 of the odds of an excellent result with both types of injury. These numbers seem quite substantial, but they are difficult to evaluate without some idea of the error to which the estimates are subject. To this end, we use the large sample standard errors for the log odds ratios. Testing whether the log odds ratios are different from zero, we get

odds ratio	log (odds ratio)	SE	z
1.27	0.2412	0.9858	0.24
3.5	1.2528	1.0635	1.18
.27	−1.2993	1.1320	−1.15
.75	−0.2877	1.2002	−0.24

The large standard errors and small z values are consistent with the result of the X^2 test. None of the odds ratios appear to be substantially different from 1. Of course, it should not be overlooked that the standard errors are really only valid for large samples and we do not have large samples. Thus, all of our conclusions about the individual odds ratios must remain tentative.

2.4 Maximum Likelihood Theory for Two-Dimensional Tables

In this section, we introduce the *likelihood function*, *maximum likelihood estimates*, and *(generalized) likelihood ratio tests*. A valuable result for maximum likelihood estimation is given below without proof.

Lemma 2.4.1. Let $f(p_1, \ldots, p_r) = \sum_{i=1}^{r} n_i \log p_i$. If $n_i > 0$ for $i = 1, \ldots, r$, then, subject to the conditions $0 < p_i < 1$ and $p_\cdot = 1$, the maximum of $f(p_1, \ldots, p_r)$ is achieved at the point $(p_1, \ldots, p_r) = (\hat{p}_1, \ldots, \hat{p}_r)$ where $\hat{p}_i = n_i/n_\cdot$.

In this section, we consider product-multinomial sampling of I populations, with each population divided into the same J categories. The I populations will form the rows of an $I \times J$ table. No results will be presented for multinomial sampling in an $I \times J$ table. The derivations of such results are similar to those presented here and are left as an exercise.

The probability of obtaining the data $n_{i1}, \ldots, n_{iJ}$ from the ith multinomial sample is

$$\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{ij}^{n_{ij}} .$$

Because the I multinomials are independent, the probability of obtaining all of the values n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, is

$$\prod_{i=1}^{I} \left[\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{ij}^{n_{ij}} \right] . \tag{1}$$

Thus, if we know the p_{ij}'s, we can find the probability of obtaining any set of n_{ij}'s. In point of fact, we are in precisely the opposite position. We do not know the p_{ij}'s, but we do know the n_{ij}'s. The n_{ij}'s have been observed. If we think of (1) as a function of the p_{ij}'s, we can write

$$L(p) = \prod_{i=1}^{I} \left[\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{J=1}^{J} p_{ij}^{n_{ij}} \right] \tag{2}$$

where $p = (p_{11}, p_{12}, \ldots, p_{IJ})$. $L(p)$ is called the likelihood function for p. Some values of p give a very small probability of observing the n_{ij}'s that

were actually observed. Such values of p are unlikely to be the true value of p. The true value of p is likely to be some value that gives a relatively large probability of observing what was actually observed. If we wish to estimate p, it makes sense to use the value of p that gives the largest probability of seeing what was actually observed. In other words, it makes sense to estimate p with a value $\hat{p}$ that maximizes the likelihood function $L(p)$. Such a value is called a *maximum likelihood estimate (MLE)* of p.

Rather than maximizing the likelihood function (which involves many products), it is often easier to maximize the log of the likelihood function (in which products change to sums). Because the logarithm is a strictly increasing function, the maximum of the likelihood and the maximum of the log of the likelihood occur as the same point.

For product-multinomial sampling, the log-likelihood function is

$$\log L(p) = \sum_{i=1}^{I} \left[\log(n_{i\cdot}!) - \sum_{j=1}^{J} \log(n_{ij}!) + \sum_{j=1}^{J} n_{ij} \log p_{ij} \right].$$

To maximize this as a function of p, we can ignore any terms that do not depend on p. It suffices to maximize

$$\ell(p) = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log p_{ij}.$$

The maximum is achieved when we maximize each of the terms $\sum_{j=1}^{J} n_{ij} \log p_{ij}$. By Lemma 2.4.1, the maximum is achieved at $p = \hat{p}$, where

$$\hat{p}_{ij} \equiv n_{ij}/n_{i\cdot}\,.$$

We can also obtain maximum likelihood estimates for the expected counts m_{ij}. Because $m_{ij} = n_{i\cdot} p_{ij}$, the MLE of m_{ij} is

$$\hat{m}_{ij} = n_{i\cdot} \hat{p}_{ij} = n_{ij}\,.$$

This follows from the *invariance of maximum likelihood estimates*: For any parameter θ and MLE $\hat{\theta}$, the MLE of a function of θ, say $f(\theta)$, is the corresponding function of the MLE, $f(\hat{\theta})$, cf. Cox and Hinkley (1974, p. 287).

If we change the model so that the null hypothesis

$$H_0 : p_{1j} = \ldots = p_{Ij}, \qquad j = 1, \ldots, J,$$

is true, we get different maximum likelihood estimates. Let $\pi_j = p_{1j} = \cdots = p_{Ij}$. The log-likelihood function becomes

$$\log L(p) = \sum_{i=1}^{I} \left[\log(n_{i\cdot}!) - \sum_{j=1}^{J} n_{ij}! + \sum_{j=1}^{J} n_{ij} \log \pi_j \right].$$

Ignoring terms that do not involve the p_{ij}'s, we are led to maximize

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\pi_j$$

or, equivalently,

$$\sum_{j=1}^{J} n_{\cdot j}\log\pi_j\,.$$

By Lemma 2.4.1, the maximum likelihood estimates become

$$\hat{p}_{ij}^{(0)} = \hat{\pi}_j = n_{\cdot j}/n_{\cdot\cdot}$$

where the (0) in $\hat{p}_{ij}^{(0)}$ is used to indicate that the estimate was obtained assuming that H_0 was true.

Maximum likelihood estimates of the m_{ij}'s are easily obtained under the null model H_0. Because $m_{ij} = n_{i\cdot}p_{ij}$,

$$\hat{m}_{ij}^{(0)} = n_{i\cdot}\hat{p}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}\,.$$

Note that $\hat{p}_{ij}^{(0)}$ and $\hat{m}_{ij}^{(0)}$ are precisely the estimates used in Section 3 to test H_0.

The likelihood function can also be used as the basis for a test of whether H_0 is true. The data have a certain likelihood of being observed, which can be summarized as the maximum value that the likelihood function achieves. If we place any restriction on the possible values of the p_{ij}'s, we will reduce the likelihood of observing the data. If placing a restriction on the p_{ij}'s reduces the likelihood too much, we can infer that the restriction on the p_{ij}'s is not likely to be valid. The relative reduction in the likelihood can be measured by looking at the maximum of $L(p)$ subject to the restriction divided by the overall maximum of $L(p)$. If this ratio gets too small, we will reject the assumption that the restriction on the p_{ij}'s is valid. In particular, if the restriction on the p_{ij}'s is that H_0 is true, we reject H_0 when the likelihood ratio is too small.

Again, we can simplify the mathematics by examining the log of the likelihood ratio and rejecting H_0 when the log gets too small. Of course, the log of the likelihood ratio is just the difference in the log-likelihoods. The maximum value of the log-likelihood when the reduced model H_0 is true is

$$\log L(\hat{p}^{(0)}) = \sum_{i=1}^{I}\left[\log(n_{i\cdot}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log(n_{\cdot j}/n_{\cdot\cdot})\right].$$

The overall maximum of the log-likelihood is

$$\log L(\hat{p}) = \sum_{i=1}^{I}\left[\log(n_{i\cdot}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log(n_{ij}/n_{i\cdot})\right].$$

The difference is

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{\cdot j}/n_{\cdot\cdot}) - \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{ij}/n_{i\cdot})\,.$$

If we multiply by -2 and simplify, we get a likelihood ratio test statistic

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right)$$

where $\hat{m}_{ij} = n_{ij}$ is the MLE of m_{ij} in the unrestricted model and $\hat{m}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ is the MLE of m_{ij} under the restriction that H_0 is true.

The reason for multiplying by -2 is that with this multiplication, the approximation

$$G^2 \sim \chi^2\left((I-1)(J-1)\right)$$

is valid when H_0 is true and the samples are large. Note that because H_0 was to be rejected for small values of the likelihood ratio, after taking logs and multiplying by -2, H_0 should be rejected for large values of G^2. In particular, for large samples, an α level test of H_0 is rejected if

$$G^2 > \chi^2\left(1-\alpha,(I-1)(J-1)\right).$$

Example 2.4.2. Computing the likelihood ratio test statistic using the data and estimated expected cell counts on knee operations in Example 2.3.1 gives

$$G^2 = 3.173\,.$$

This is similar to, but distinct from, the Pearson test statistic $X^2 = 3.229$. Both are based on 4 degrees of freedom. In this example, formal tests using either statistic suffer from the fact that the sample is not large.

Larntz (1978) indicates that, for small samples, the actual size of the test that rejects H_0 if

$$X^2 > \chi^2(1-\alpha,(I-1)(J-1))$$

tends to be nearer the nominal size α than the corresponding likelihood ratio test. This is related to the fact that G^2 becomes too large when the observations are small but the estimated expected cell counts are not. Kreiner (1987) comes to similar conclusions. From the results of Larntz and others, Fienberg (1979) concludes that (a) if the minimum expected cell count is about 1, χ^2 tests often work well and (b) if the sample size is 4 to 5 times the number of cells in the table, χ^2 tests give *P values* with the correct order of magnitude. In practice, the first of these conclusions must

compare the *estimated* expected cell counts to 1. In Example 2.4.2, the test statistics are similar, so the choice of test is not important. The data also pass both of the criteria mentioned by Fienberg. The rules of thumb given in this paragraph can be applied to higher-dimensional tables. Although X^2 has the advantage alluded to above, G^2 is more convenient to use in analyzing higher-dimensional tables. The likelihood ratio test statistic will be used almost exclusively after Chapter 3.

Discussion

There are philosophical grounds for preferring the use of G^2. The likelihood principle indicates that one's conclusions should depend on the relative values of the likelihood function. The likelihood function depends only on the data that actually occurred. Because G^2 is computed from the likelihood, its use *can* be consistent with the likelihood principle. Unfortunately, the standard use of G^2 is to compute a *P value* or to perform an α level test. Both of these procedures depend on data that could have happened but did not, so these uses for G^2 violate the likelihood principle. An excellent discussion of the likelihood principle is given by Berger and Wolpert (1984).

Although formal tests are conducted throughout this book, the real emphasis is on informal evaluation of models using G^2 and other tools. The emphasis is on modeling and data analysis, not formal inferential procedures. Nevertheless, a relatively complete account is given of the standard results in formal log-linear model methodology. Bayesian methods are the primary inferential methods that satisfy the likelihood principle. Chapter 13 discusses Bayesian logistic regression — but not log-linear models. Santner and Duffy (1989) include discussion of Bayesian methods.

EXERCISE 2.3. For multinomial sampling, H_0 is the restriction that $p_{ij} = p_{i\cdot}p_{\cdot j}$ for all i and j. Show that

(a) the unrestricted MLE of p_{ij} is $\hat{p}_{ij} = n_{ij}/n_{\cdot\cdot}$

(b) the unrestricted MLE of m_{ij} is $\hat{m}_{ij} = n_{ij}$

(c) the MLE of p_{ij} under H_0 is $\hat{p}_{ij}^{(0)} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = (n_{i\cdot}/n_{\cdot\cdot})(n_{\cdot j}/n_{\cdot\cdot})$

(d) the MLE of m_{ij} under H_0 is $\hat{m}_{ij}^{(0)} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$

(e) the likelihood ratio test rejects H_0 when

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right)$$

gets too large.

2.5 Log-Linear Models for Two-Dimensional Tables

It is our intention to exploit the similarities between analysis of variance (ANOVA) and regression on the one hand and log-linear and logistic regression models on the other. We begin by discussing two-factor analysis of variance.

Consider a balanced ANOVA model $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$. We can change the symbols used to denote the parameters and rewrite the model as

$$y_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + e_{ijk}, \tag{1}$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$. The e_{ijk}'s are assumed to be independent $N(0, \sigma^2)$. We can estimate σ^2 and test for interaction. If interaction exists, we can look at contrasts in the interaction; if no interaction exists, we can test for main effects and look at contrasts in the main effects. If some factor levels correspond to quantitative values, then regression ideas can be incorporated into the ANOVA. The estimate of σ^2 is the mean squared error

$$\text{MSE} = \frac{1}{IJ(K-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{ijk} - \bar{y}_{ij\cdot})^2 \ .$$

Everything in the analysis other than the estimate of σ^2 is a function of the $\bar{y}_{ij\cdot}$'s. In particular, we can form an $I \times J$ table of the $\bar{y}_{ij\cdot}$'s. The goal of the analysis is to explore the structure of this table. The ANOVA model (1) and the corresponding contrasts in interactions and main effects have proved to be very useful tools in exploring this $I \times J$ table.

Let us reconsider what the ANOVA model is really saying. Basically, the ANOVA model is saying that the y_{ijk}'s are independent and that

$$y_{ijk} \sim N(m_{ij}, \sigma^2)$$

where

$$m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \, . \tag{2}$$

Our goal is to examine the structure of the m_{ij}'s. To do that, we use the MLEs of the m_{ij}'s, which are

$$\hat{m}_{ij} = \bar{y}_{ij\cdot} \, .$$

Our statistical inferences are based on the fact that the $\hat{m}_{ij}$'s are independent with

$$\hat{m}_{ij} \sim N(m_{ij}, \sigma^2/K)$$

and that the MSE is an estimate of σ^2, which is independent of the $\hat{m}_{ij}$'s. It is of interest to note that although the MSE is not the MLE of σ^2, exactly

the same tests and confidence intervals for the m_{ij}'s would be obtained if the MLE for σ^2 was used in place of the MSE (and suitable adjustments in distributions were made).

If we impose a restriction on the m_{ij}'s – for example, the restriction of no interaction

$$m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{3}$$

– the MLEs of the m_{ij}'s change. In particular,

$$\hat{m}_{ij} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) \tag{4}$$

and the MLE of σ^2 also changes. It can be shown that the usual F test for no interaction is just the likelihood ratio test for no interaction.

To examine an $I \times J$ table of counts, we use similar techniques. The table entries have the property that

$$\mathrm{E}(n_{ij}) = m_{ij}\,.$$

Again, we are interested in the structure of the m_{ij}'s; however, instead of considering linear models like (2) and (3), we consider log-linear models such as

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

and

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}\,.$$

Our analysis will again rely on the MLEs of the m_{ij}'s and on likelihood ratio tests; however, there are some differences. The n_{ij}'s are typically multinomial or product-multinomial. Small sample results similar to those from standard analysis of variance are not available. Traditionally, tests and confidence intervals have been based on large sample approximate distributions. On the other hand, multinomial distributions depend only on the p_{ij}'s or, equivalently, the m_{ij}'s, so there is no need to deal with a term analogous to σ^2 in normal theory. Finally, the ANOVA model (1) is balanced; it has K observations in each cell of the table. This balance leads to simplifications in the analysis. If there are different numbers of observations in the cells, the simplifications are lost. For example, the simple formula (4) for MLEs under the no-interaction model does not apply. Log-linear models are analogous to ANOVA models with unequal numbers of observations. They almost never display all the simplifications associated with balanced observations in ANOVA and they only occasionally have simple formulas for MLEs. Although most work on log-linear models has used large sample (*asymptotic*) distributions, recently there has been considerable work on exact conditional inference and Bayesian inference for small samples. See the subsection on Other Sampling Methods in Section 3.5 for further discussion of exact conditional inference and Chapter 13 for a discussion of Bayesian methods.

There are several reasons for writing ANOVA type models for the $\log(m_{ij})$'s rather than the m_{ij}'s. One is that the large sample theory can be worked out. In other words, one reason to do it is because it can be done. Another reason is that log-linear models arise in a natural fashion from the mathematics of Poisson sampling, cf. Chapter 9. Multinomial expected cell counts are bounded between 0 and the sample size N, these bounds place awkward limits on the parameters of ANOVA type models for the m_{ij}'s. Such problems do not arise in log-linear modeling. One of the best reasons for considering log-linear models is that they often have very nice interpretations. We now examine the interpretations of log-linear models for two factors.

Consider a multinomial sample. We know that $m_{ij} = n_{\cdot\cdot}p_{ij}$. We can write a log-linear model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \,. \tag{5}$$

This has accomplished absolutely nothing! The terms $u_{12(ij)}$ are sufficient to explain the m_{ij}'s. The u, $u_{1(i)}$, and $u_{2(j)}$ terms are totally redundant; they can have any values at all, and yet, by choosing the $u_{12(ij)}$'s appropriately, equation (5) can be made to hold. Because model (5) has enough u terms to completely explain any set of m_{ij}'s, model (5) is referred to as a *saturated* model.

A more interesting example of a log-linear model occurs when the rows and columns of the table are independent. If

$$m_{ij} = n_{\cdot\cdot}p_{i\cdot}p_{\cdot j} \,,$$

then

$$\log m_{ij} = \log n_{\cdot\cdot} + \log p_{i\cdot} + \log p_{\cdot j} \,.$$

In other words, if rows and columns are independent, a log-linear model of the form

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{6}$$

holds. However, if we are to base our analysis on log-linear models, it is even more important to know that if model (6) holds, then rows and columns are independent.

Theorem 2.5.1. For multinomial sampling in an $I \times J$ table, $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, if and only if $p_{ij} = p_{i\cdot}p_{\cdot j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$.

Proof. We have already shown that independence implies the log-linear model.

If the log-linear model holds, then

$$m_{ij} = e^{u + u_{1(i)} + u_{2(j)}} \,.$$

Let $a = e^u$, $a_{1(i)} = e^{u_{1(i)}}$, and $a_{2(j)} = e^{u_{2(j)}}$. Let $a_{1(\cdot)} = \sum_{i=1}^{I} a_{1(i)}$ and similarly for $a_{2(\cdot)}$. Note that

$$\begin{aligned} p_{ij} &= m_{ij}/n_{\cdot\cdot} = a\, a_{1(i)} a_{2(j)}/n_{\cdot\cdot}\,, \\ p_{i\cdot} &= a a_{1(i)} a_{2(\cdot)}/n_{\cdot\cdot}\,, \\ p_{\cdot j} &= a a_{1(\cdot)} a_{2(j)}/n_{\cdot\cdot}\,, \end{aligned}$$

and

$$1 = p_{\cdot\cdot} = a a_{1(\cdot)} a_{2(\cdot)}/n_{\cdot\cdot}\,.$$

Substitution gives

$$\begin{aligned} p_{i\cdot} p_{\cdot j} &= a a_{1(i)} a_{2(\cdot)} a a_{1(\cdot)} a_{2(j)}/n_{\cdot\cdot}^2 \\ &= (a a_{1(i)} a_{2(j)}/n_{\cdot\cdot})(a a_{1(\cdot)} a_{2(\cdot)}/n_{\cdot\cdot}) \\ &= a a_{1(i)} a_{2(j)}/n_{\cdot\cdot} \\ &= p_{ij}\,. \end{aligned}$$

Thus, the log-linear model implies independence. □

For product-multinomial sampling,

$$m_{ij} = n_{i\cdot} p_{ij} \tag{7}$$

and the log-linear model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

holds trivially. Now, consider the model under H_0. If $\pi_j = p_{1j} = \cdots = p_{Ij}$ for all $j = 1, \ldots, J$, then

$$m_{ij} = n_{i\cdot} \pi_j\,.$$

Theorem 2.5.2. For product-multinomial sampling in an $I \times J$ table where rows are independent samples, $\log m_{ij} = u + u_{1(i)} + u_{2(j)}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, if and only if $p_{1j} = \cdots = p_{Ij}$, $j = 1, \ldots, J$.

Proof. If for each j the probabilities p_{ij} are equal, we have $m_{ij} = n_{i\cdot}\pi_j$ and $\log m_{ij} = \log n_{i\cdot} + \log \pi_j$. Taking $u = 0$, $u_{1(i)} = \log(n_{i\cdot})$, and $u_{2(j)} = \log(\pi_j)$ shows that the log-linear model holds.

Conversely, if $\log m_{ij} = u + u_{1(i)} + u_{2(j)}$, then $m_{ij} = a a_{1(i)} a_{2(j)}$, where $a = e^u$, $a_{1(i)} = e^{u_{1(i)}}$, and $a_{2(j)} = e^{u_{2(j)}}$. Note that $p_{i\cdot} = 1$, so from (7), $m_{i\cdot} = n_{i\cdot}$ and

$$n_{i\cdot} = a\, a_{1(i)} a_{2(\cdot)}\,.$$

Because $p_{ij} = m_{ij}/n_{i\cdot}$,

$$\begin{aligned} p_{ij} &= aa_{1(i)}a_{2(j)}/n_{i\cdot} \\ &= aa_{1(i)}a_{2(j)}/aa_{1(i)}a_{2(\cdot)} \\ &= a_{2(j)}/a_{2(\cdot)} \,. \end{aligned}$$

This is true for any i, so $a_{2(j)}/a_{2(\cdot)} = p_{1j} = p_{2j} = \cdots = p_{Ij}$, $j = 1, \ldots, J$. □

2.5.1 Odds Ratios

In applications with high-dimensional tables, it is rare that there are no important interactions. In order to explore the nature of the interactions, we need to look at contrasts in the interactions. To do this, we need a method of defining contrasts in the interactions. We begin by reviewing methods for examining interactions in analysis of variance.

Let q_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be any set of numbers with the property that $q_{i\cdot} = q_{\cdot j} = 0$. In balanced analysis of variance,

$$\sum_{i=1}^{I}\sum_{j=1}^{J} q_{ij}m_{ij} \tag{8}$$

is a contrast in the interactions. Using model (2) and the fact that $q_{i\cdot} = q_{\cdot j} = 0$, the contrast (8) can also be written as

$$\sum_{i=1}^{I}\sum_{j=1}^{J} q_{ij}u_{12(ij)}$$

which involves only the interactions. The most interpretable way of obtaining a contrast in the interactions is to define the interaction contrast in terms of contrasts in the main effects. Let a_i, $i = 1, \ldots, I$, determine a contrast in the rows (thus, $a_{\cdot} = 0$) and let b_j, $j = 1, \ldots, J$, determine a contrast in the columns (so $b_{\cdot} = 0$). Then, if we take $q_{ij} = a_i b_j$, we get a contrast in the interactions. Recall that if there is no interaction, all interaction contrasts equal zero. Conversely, the interaction has $(I-1)(J-1)$ degrees of freedom, so specifying that any $(I-1)(J-1)$ linearly independent contrasts in the interaction are all zero is equivalent to specifying that there is no interaction.

A valuable data analytic technique for examining interactions in two-way analysis of variance is the *interaction plot*. It consists of plotting the I curves determined by connecting the points $(j, \hat{m}_{ij})$, $j = 1, \ldots, J$, with line segments. In this plot, $\hat{m}_{ij} = \bar{y}_{ij\cdot}$, the estimate of m_{ij} in model (2). If there is no interaction, $m_{ij} = u + u_{1(i)} + u_{2(j)}$ and the I theoretical curves (j, m_{ij}) are parallel. If interaction exists, the theoretical curves are not

parallel. The curves $(j, \hat{m}_{ij})$ estimate the theoretical curves. If the curves $(j, \hat{m}_{ij})$ are approximately parallel, it suggests that there is no interaction. If interaction exists, the estimated curves can suggest the nature of the interaction. Whether the plots are approximately parallel depends on the variability of the $\hat{m}_{ij}$'s.

Rather than plotting the I curves based on $(j, \hat{m}_{ij})$, one can plot the J curves based on $(i, \hat{m}_{ij})$, $i = 1, \ldots, I$. Again, in the absence of interactions, the curves should be approximately parallel. If the column treatments correspond to quantitative levels, say, x_j, $j = 1, \ldots, J$, then plots of $(x_j, \hat{m}_{ij})$ are appropriate. Again, one looks for parallelism. Similar plots can be constructed for row treatments with quantitative levels.

In log-linear models, the same procedures can be applied to the $\log(m_{ij})$'s. In particular, specifying that an odds ratio equals one is equivalent to specifying that an interaction contrast is zero. First, note that odds ratios can be written in terms of expected values. For product-multinomial sampling,

$$m_{ij} = n_{i\cdot} p_{ij}$$

and for multinomial sampling,

$$m_{ij} = n_{\cdot\cdot} p_{ij} \,.$$

In either case,

$$\frac{p_{ij} p_{i'j'}}{p_{ij'} p_{i'j}} = \frac{m_{ij} m_{i'j'}}{m_{ij'} m_{i'j}} \,.$$

If

$$\frac{m_{ij} m_{i'j'}}{m_{ij'} m_{i'j}} = 1 \,,$$

then taking logs gives

$$\log m_{ij} - \log m_{ij'} - \log m_{i'j} + \log m_{i'j'} = 0 \,.$$

This is precisely the statement that the interaction contrast

$$\sum_{r=1}^{I} \sum_{s=1}^{J} q_{rs} \log(m_{rs}) \tag{9}$$

equals zero, where $q_{ij} = q_{i'j'} = 1$, $q_{ij'} = q_{i'j} = -1$, and $q_{rs} = 0$ for all other pairs (r, s). In particular, the coefficients q_{rs} can be obtained by combining the contrast in the rows $a_i = 1$, $a_{i'} = -1$, and $a_r = 0$ for all other r with the contrast in the columns $b_j = 1$, $b_{j'} = -1$, and $b_s = 0$ for all other s. Observe that the contrast (9) can also be written

$$\sum_{r=1}^{I} \sum_{s=1}^{J} q_{rs} u_{12(rs)}$$

where we have used model (5) and the fact that $q_{r\cdot} = q_{\cdot s} = 0$.

If we specify that

$$\frac{m_{11}m_{ij}}{m_{1j}m_{i1}} = 1$$

for all $i = 2, \ldots, I$ and $j = 2, \ldots, J$, then we have specified that $(I-1)(J-1)$ linearly independent interaction contrasts in the $\log(m_{ij})$'s are all equal to zero; hence, there is no interaction.

As with analysis of variance, an interaction plot can be a valuable tool in the analysis of log-linear models. The I curves that connect the sets of points $(j, \log(\hat{m}_{ij}))$, $j = 1, \ldots, J$, are the basis of the interaction plot. The estimated expected counts $\hat{m}_{ij}$ are estimated using model (5), which contains interaction. Under model (5), $\hat{m}_{ij} = n_{ij}$. The I curves estimate the theoretical curves based on $(j, \log(m_{ij}))$. If there is no interaction, the theoretical curves are parallel and estimated curves should indicate this. If interaction exists, the nature of the interaction should be suggested by the estimated curves.

EXAMPLE 2.5.3. Consider the data given below on the relationship between college of enrollment and political affiliation for university students.

		Political Affiliation			
		Rep.	Dem.	Ind.	Total
College	Letters	34	61	16	111
	Engineering	31	19	17	67
	Agriculture	19	23	16	58
	Education	23	39	12	74
	Totals	107	142	61	310

The Pearson and likelihood ratio test statistics for independence (no interaction) are

$$X^2 = 16.16$$

and

$$G^2 = 16.39\,.$$

The test has $(4-1)(3-1) = 6$ degrees of freedom. The 99th percentile of a $\chi^2(6)$ is 16.81, so the *P value* for either statistic is a little above .01. An interaction plot uses the values $\log(n_{ij})$ given below.

	Political Affiliation		
	Rep.	Dem.	Ind.
Letters	3.5	4.1	2.8
Engineering	3.4	2.9	2.8
Agriculture	2.9	3.1	2.8
Education	3.1	3.7	2.5

The interaction plot is given in Figure 2.1. The curves for Letters and Education are almost parallel. The curve for Agriculture is similar but not

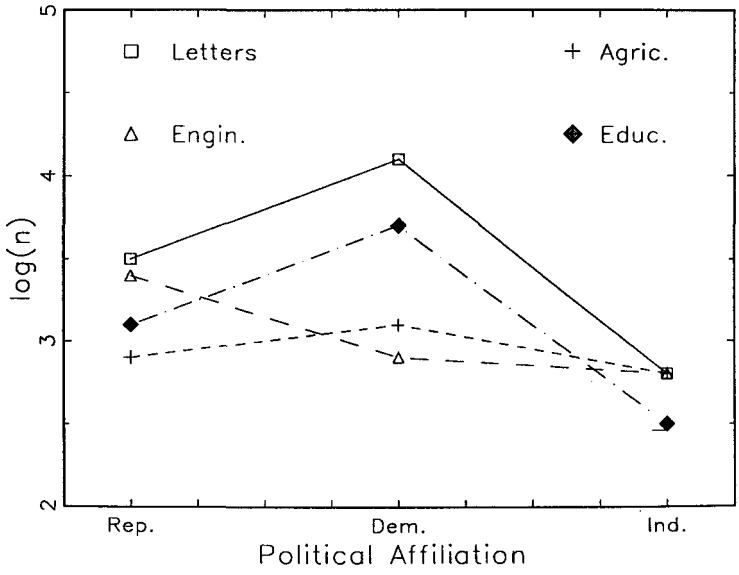


FIGURE 2.1. Interaction Plot

nearly as concave. In fact, the Agriculture curve is nearly horizontal. The Engineering curve is clearly the main source of interaction. It does not behave at all like the other three. It is clearly not parallel to the others. If Engineering is dropped from the table and the resulting 3×3 table is fit, one gets $X^2 = 5.770$ and $G^2 = 5.536$ on 4 degrees of freedom. The *P value* is a bit larger than .2. Without Engineering, there is no evidence for lack of independence. This confirms that the main source of interaction is in Engineering.

2.6 Simple Logistic Regression

In this section, we deal with simple logistic regression in which we use a predictor variable to estimate probabilities. Simple logistic regression, in fact, all of logistic regression, can be viewed as an extension of standard regression analysis. It can also be viewed as modeling the interactions in two-dimensional tables.

EXAMPLE 2.6.1. *O-Ring Data.*
Table 2.1 presents data from Dalal, Fowlkes, and Hoadley (1989) on field O-ring failures in the 23 pre-*Challenger* space shuttle launches. See also Lavine (1991) and Martz and Zimmer (1992). *Challenger* was the shuttle that blew up on take off. Temperature is the predictor variable. The *Challenger* explosion occurred during a takeoff at 31 degrees Fahrenheit. Each flight

is viewed as an independent trial. The result of a trial is 1 if any field O-rings failed on the flight and 0 if all the O-rings functioned properly. A simple logistic regression uses temperature to model the probability that any O-ring failed. Such a model allows us to predict O-ring failure from temperature.

TABLE 2.1. O-Ring Failure Data

Case	Flight	Failure	Success	Temperature
1	14	1	0	53
2	9	1	0	57
3	23	1	0	58
4	10	1	0	63
5	1	0	1	66
6	5	0	1	67
7	13	0	1	67
8	15	0	1	67
9	4	0	1	68
10	3	0	1	69
11	8	0	1	70
12	17	0	1	70
13	2	1	0	70
14	11	1	0	70
15	6	0	1	72
16	7	0	1	73
17	16	0	1	75
18	21	1	0	75
19	19	0	1	76
20	22	0	1	76
21	12	0	1	78
22	20	0	1	79
23	18	0	1	81

Let p_i be the probability that any O-ring fails in case i. A simple linear logistic regression model for these data is

$$\text{logit}(p_i) \equiv \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \tau_i,$$

where τ_i is the known temperature and β_0 and β_1 are unknown intercept and slope parameters (coefficients). The logistic regression model presents the log odds of O-ring failure as a linear function of temperature.

We again use maximum likelihood estimates. The likelihood function for logistic regression is discussed later in this section. The procedure for finding maximum likelihood estimates is discussed later in the book. For now, we merely present results and use analogies to standard regression.

The coefficient estimates, standard errors, and z values are

Variable	Estimate	Std. Error	z
Intercept	15.04	7.316	2.06
Temperature	−0.2321	0.1073	−2.16

The z values are simply the estimate divided by the standard error. They are test statistics for testing whether a coefficient equals zero. In particular, $z = -2.16$ yields a *P value* for $H_0 : \beta_1 = 0$ that is approximately .03. An alternative and preferred test is presented later.

To predict the probability of any O-ring failures for a flight at a temperature of τ,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\tau$$

which can be rearranged into

$$p = \frac{\exp(\beta_0 + \beta_1\tau)}{1 + \exp(\beta_0 + \beta_1\tau)}.$$

The estimated probability is

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\tau)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\tau)}.$$

Figure 2.2 gives a plot of the estimated probabilities as a function of temperature. The *Challenger* was launched at $\tau = 31$ degrees Fahrenheit, so the predicted log odds are $15.04 - (.2321)31 = 7.8449$ and the predicted probability of an O-ring failure is $e^{7.8449}/(1 + e^{7.8449}) = .9996$. Actually, there are problems with this prediction because we are predicting very far from the observed data. The lowest temperature at which a shuttle had previously been launched was 53 degrees, very far from 31 degrees. According to the fitted model, a launch at 53 degrees has probability .939 of O- ring failure, so even with the caveat about predicting beyond the range of the data, the model indicates an overwhelming probability of failure.

Before discussing logistic regression in general, we review standard one-way ANOVA and simple linear regression with normal errors. Suppose we have independent observations y_{ij} on I populations. The one-way ANOVA model is

$$y_{ij} = m_i + \varepsilon_{ij} \tag{1}$$

ε_{ij}'s independent $N(0, \sigma^2)$, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$. Here, $\mathrm{E}(y_{ij}) \equiv m_i$. Alternatively, when a predictor variable x_i is available for each population, a simple linear regression model for the y_{ij}'s is

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}\ . \tag{2}$$

Model (2) is specifying a linear structure for the m_i's defined in model (1), i.e., for $i = 1, \ldots, I$

$$m_i = \beta_0 + \beta_1 x_i\ .$$

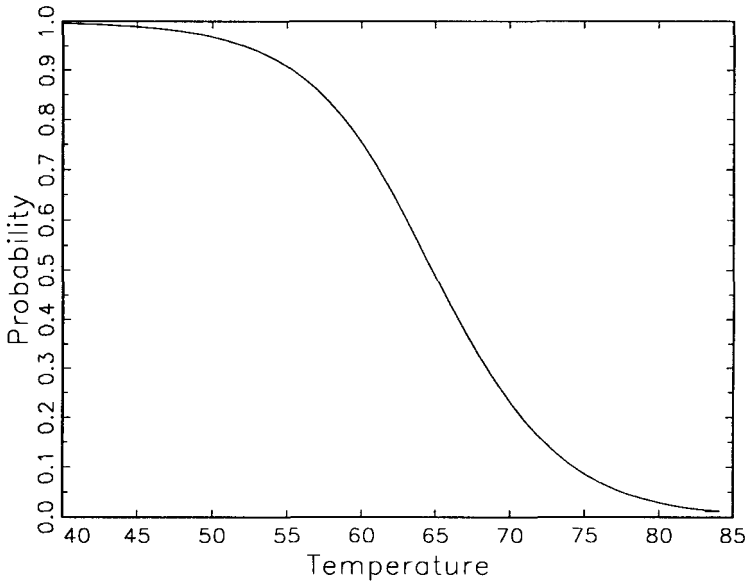


FIGURE 2.2. O-Ring Failure Probabilities

In general, we construct similar models for binomial data, except that the models are for the log odds rather than for the expected values. In a simple logistic regression, we have independent observations from I populations; each is $y_i \sim \text{Bin}(N_i, p_i)$. The N_i trials in the binomial play the same role as the N_i replicate observations in ANOVA. Recall that $\text{E}(y_i) = m_i = N_i p_i$ and that the odds are $p_i/(1 - p_i)$. Logistic regression specifies a linear structure for the log odds

$$\text{logit}(p_i) \equiv \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i, \tag{3}$$

$i = 1, \ldots, I$. Note that by varying x_i, the term on the right of equation (3) can take on any real value. Although the probabilities p_i are restricted to be between 0 and 1, the log odds can also take on any real value.

Alternatively, the logistic regression model for the log odds can be viewed as a log-linear model. For example, we can think of the O-ring data in Table 2.1 as providing a two-way table in which there are independent samples from 23 populations and a binomial response of failure or success. Associated with the 23 populations in this 23×2 table are temperature values that we can use to model the interaction.

In general, we rewrite a logistic regression with I independent binomials as an $I \times 2$ two-way table. This involves substantially changing the notation we have used. The sampling is product-multinomial (actually, product-binomial). $y_i \sim \text{Bin}(N_i, p_i)$, so $(N_i - y_i) \sim \text{Bin}(N_i, 1 - p_i)$. In terms of a two-way table, write $y_i \equiv n_{i1}$ and $N_i - y_i \equiv n_{i2}$. Note that $N_i = n_{i\cdot}$ for all

i. Also, $p_i \equiv p_{i1}$ and $1 - p_i \equiv p_{i2}$ with similar definitions for the expected values. In particular, $m_i = N_i p_i = n_{i\cdot} p_{i1} = m_{i1}$ and $m_{i2} = N_i(1 - p_i)$, so $p_i/(1 - p_i) = m_{i1}/m_{i2}$.

The log-linear version of model (3) is

$$\log(m_{ij}) = u_{1(i)} + u_{2(j)} + \eta_j x_i \tag{4}$$

where the usual interaction term $u_{12(ij)}$ from (2.5.5) is being replaced in the model by a more specific interaction term, $\eta_j x_i$. Of course, x_i is the known predictor variable, but η_j is an unknown parameter. This is an interaction term because it involves both the i and j subscripts, just like $u_{12(ij)}$. The relationship between the logistic model (3) and the log-linear model (4) is that

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \log\left(\frac{m_{i1}}{m_{i2}}\right) \\
&= \log(m_{i1}) - \log(m_{i2}) \\
&= \left[u_{1(i)} + u_{2(1)} + \eta_1 x_i\right] - \left[u_{1(i)} + u_{2(2)} + \eta_2 x_i\right] \\
&= \left[u_{2(1)} - u_{2(2)}\right] + \left[\eta_1 x_i - \eta_2 x_i\right] \\
&\equiv \beta_0 + \beta_1 x_i
\end{aligned}$$

where $\beta_0 \equiv \left[u_{2(1)} - u_{2(2)}\right]$ and $\beta_1 \equiv [\eta_1 - \eta_2]$.

As in Section 4, we can use maximum likelihood to estimate the parameters and to generate tests. The likelihood function $L(p)$ for a two-dimensional table was given in (2.4.2). Equation (3) can be rearranged to give

$$\begin{aligned}
p_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \\
1 - p_i &= \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}.
\end{aligned}$$

Recalling that $p_i \equiv p_{i1}$, $(1 - p_i) \equiv p_{i2}$, $y_i \equiv n_{i1}$, and $N_i - y_i \equiv n_{i2}$, substitution into (2.4.2) gives the likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1}^{I} \left[\frac{n_{i\cdot}!}{\prod_{j=1}^{2} n_{ij}!} \left\{ \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\}^{n_{i1}} \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\}^{n_{i2}} \right].$$

It is by no means obvious what values of β_0 and β_1 will maximize this function. In Chapters 10 and 11, we discuss the *Newton–Raphson* method for obtaining such maxima. For now, we rely on a computer program to give us the maximizing values. (See Subsections 2.6.1 and 4.4.2 for SAS, BMDP, and GLIM computer commands.)

As in the example, if p is the probability for a predictor x,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \quad \text{and} \quad p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Given the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$, we get the estimated probability associated with x:

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

In particular, this formula provides the $\hat{p}_i$'s when doing predictions at the x_i's. It also provides $\hat{m}_{ij}$'s through $\hat{m}_{ij} = n_{i\cdot}\hat{p}_{ij}$. We can try to test model (4) against the more general saturated model (2.5.5). Recall that the MLEs for the expected cell counts under model (2.5.5) are just the n_{ij}'s, so

$$\begin{aligned} G^2 &= 2\sum_{i=1}^{I}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \\ &= 2\sum_{i=1}^{I}[n_{i1}\log(n_{i1}/\hat{m}_{i1}) + n_{i2}\log(n_{i2}/\hat{m}_{i2})] \\ &= 2\sum_{i=1}^{I}[y_i\log(y_i/N_i\hat{p}_i) + (N_i - y_i)\log((N_i - y_i)/N_i(1-\hat{p}_i))]\,. \end{aligned}$$

In this formula, if $y_i = 0$, then $y_i\log(y_i)$ is taken as zero.

The Pearson test statistic is

$$\begin{aligned} X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{2}\frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \\ &= \sum_{i=1}^{I}\left[\frac{(y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i} + \frac{[(N_i - y_i) - N_i(1-\hat{p}_i)]^2}{N_i(1-\hat{p}_i)}\right] \\ &= \sum_{i=1}^{I}\left[\frac{(y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i} + \frac{(y_i - N_i\hat{p}_i)^2}{N_i(1-\hat{p}_i)}\right] \\ &= \sum_{i=1}^{I}\frac{(y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i(1-\hat{p}_i)}\,. \end{aligned}$$

The degrees of freedom for the tests are $23 - 2 = 21$, i.e., the number of cases minus one for the intercept and one for temperature. This computation is based on model (3). Alternatively, based on model (4), the degrees of freedom are the number of cells in the two-way table, 23×2, minus 23 for fitting row effects and the grand mean, minus 1 for column effects, and minus 1 for fitting the interaction term based on temperature, i.e., $46 - 23 - 1 - 1 = 21$.

G^2 and X^2 are appropriate test statistics, but, unfortunately, for them to have large sample χ^2 distributions, we need the n_{ij}'s to get large. In this example, the n_{ij}'s are 0 or 1, so a χ^2 test is inappropriate for this example. In general, *a χ^2 test of a logistic regression model against the saturated model (2.5.5) is appropriate* **only** *when the sample sizes N_i for the I populations are all large.*

We can also use model (4) as a full model and test it against a reduced model. Since models (4) and (3) are equivalent, we specify a reduced model for model (3), say

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0. \tag{5}$$

Testing model (5) against model (3) is equivalent to testing $H_0 : \beta_1 = 0$. Given an estimate $\hat{\beta}_0$ for model (5), we get $\hat{p}_i = e^{\hat{\beta}_0}/(1+e^{\hat{\beta}_0})$, estimates $\hat{p}_{ij}$, and estimates, say, $\hat{m}_{ij}^{(0)} = n_{i\cdot}\hat{p}_{ij}$, where the (0) indicates that the expected cell count is estimated under $H_0 : \beta_1 = 0$. The test statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{2}\hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}_{ij}^{(0)}}\right).$$

Unlike standard regression analysis where the t test for $H_0 : \beta_1 = 0$ is equivalent to the F test, in logistic regression the z test described earlier can give different results than the G^2 test described here. Both tests of H_0 will generally be valid whenever I is large.

EXERCISE 2.4 Show that the independence model (2.5.6) implies model (5). Hint: Use the same method as was used to show that model (4) implies model (3).

Crude standardized residuals can be defined as

$$\tilde{r}_i = \frac{y_i - N_i\hat{p}_i}{\sqrt{N_i\hat{p}_i(1-\hat{p}_i)}}, \tag{6}$$

so that Pearson's chi-squared is $X^2 = \sum_{i=1}^{s}\tilde{r}_i^2$. Note that $\mathrm{Var}(y_i) = N_i p_i(1-p_i)$, making this definition of crude standardized residuals an estimate of $y_i - \mathrm{E}(y_i)/\sqrt{\mathrm{Var}(y_i)}$. (These are "crude" in that they ignore the variability of $\hat{p}_i$.) When $N_i = 1$, the residuals will not have an asymptotic normal distribution, which is a major reason why these residuals do not behave like residuals in normal theory models.

EXAMPLE 2.6.1 CONTINUED. For the simple linear logistic regression model, $G^2 = 20.315$ with 21 degrees of freedom. For the intercept-only model, $G^2 = 28.267$ with 22 degrees of freedom. Since $N_i = 1$ for all i, neither of these G^2's is compared directly to a chi-squared distribution.

The model based test for $H_0 : \beta_1 = 0$ has $G^2 = 28.267 - 20.315 = 7.952$ on $df = 22-21 = 1$. Comparing this to a $\chi^2(1)$ distribution, the *P value* for the test is approximately .005. It is considerably smaller than the *P value* for the z test of H_0. This test is generally preferred to the z test.

Since $N_i = 1$ for all i, we delay consideration of residuals until Chapter 4.

All of the methods presented in this section carry over to multiple logistic regression in which there is more than one predictor variable. Such models are discussed in Chapter 4.

2.6.1 Computer Commands

The data are in a file 'oring.dat' that looks like Table 2.1 except it has an extra column at the right (which contains the actual number of O-rings that failed on each flight). Perhaps the simplest way to fit the logistic regression model in SAS is to use PROC GENMOD.

```
options ps=60 ls=72 nodate;
data oring;
   infile 'oring.dat';
   input ID flt f s temp junk;
   n = 1;
proc genmod data = oring;
   model f/n = temp  / link=logit
                       dist=binomial;
run;
```

The first line controls printing of the output. The next four lines define the data. The variable "n" is used to specify that there is only one trial in each of the 23 binomials. PROC GENMOD needs the data specified: "data = oring". GENMOD also needs information on the model. "link = logit" and "dist = binomial" are both needed to specify that a logistic regression is being fitted. "model f/n = temp" indicates that we are modeling the number of failures in "f" out of "n" trials using the predictor "temp" (and implicitly an intercept).

A more powerful SAS program for logistic regression is PROC LOGISTIC. Commands for this, BMDP-LR, and GLIM are given in Subsection 4.4.2.

2.7 Exercises

EXERCISE 2.7.1. The data in Table 2.2 are on graduate admissions by sex at the University of California, Berkeley, and are given by Bickel et al. (1975) and Freedman et al. (1978). Test for independence, examine the

Pearson residuals, and evaluate the odds ratio. What conclusions do you reach? (Do not put too much credence in your analysis; the data will be reanalyzed in Exercise 3.6.4.)

TABLE 2.2. Graduate Admissions at Berkeley

	Male	Female
Admitted	1198	557
Rejected	1493	1278

EXERCISE 2.7.2. Cramér (1946) presents data on the distribution of birth dates for males and females born in Sweden in 1935. The data given in Table 2.3 presume a natural ordering for the months of the year that Cramér does not specify. Analyze the data. Is it better to think of this as one multinomial sample or as two independent multinomial samples?

TABLE 2.3. Swedish Birth Dates

Month	Female	Male
January	3537	3743
February	3407	3550
March	3866	4017
April	3711	4173
May	3775	4117
June	3665	3944
July	3621	3964
August	3596	3797
September	3491	3712
October	3391	3512
November	3160	3392
December	3371	3761

EXERCISE 2.7.3. Gilby (1911) presents data on the relationships among instructor's evaluation of general intelligence, quality of clothing, and school standard. General intelligence was classified using a system of Karl Pearson's that was reported in Waite (1911). Briefly, the Intelligence classifications are A – Mentally Defective, B – Dull, C – Slow, E – Fairly Intelligent, F – Capable, and G – Very Able. Clothing was classified as I – Very Well Clad, II – Well Clad, III – Poor but Passable, IV – Insufficient, V – Worse than Insufficient. Throughout, intelligence category A was combined with B and clothing category V was combined with IV. This was done because of small numbers of observations. The third variable, Standard, seems to

be similar to the American idea of a school grade. For example, roughly half of 10-year-olds were in Standard III with most of the others in II or IV. For $10\frac{1}{2}$-year-olds, about two-thirds were in standards III or IV with most of the rest in II or V. Data were collected from 36 instructors spread over eight different primary schools. Tables 2.4, 2.5, and 2.6 summarize some of the data; use the methods of Chapter 2 to analyze these data.

TABLE 2.4. Intelligence versus Clothing

	Intelligence					
Clothing	B	C	D	E	F	G
I	33	48	113	209	194	39
II	41	100	202	255	138	15
III	39	58	70	61	33	4
IV,V	17	13	22	10	10	1

TABLE 2.5. Intelligence versus Standard

	Intelligence					
Standard	B	C	D	E	F	G
I	17	27	45	50	27	1
II	23	34	61	66	36	1
III	42	42	69	117	72	10
IV	16	25	41	75	53	11
V	18	38	66	77	45	6
VI	10	32	73	80	98	18
VII	4	19	39	52	35	11
VIII	0	2	13	18	9	1

TABLE 2.6. Clothing versus Standard

	Clothing			
Standard	I	II	III	IV,V
I	20	87	56	4
II	71	88	42	20
III	157	134	41	20
IV	82	77	45	17
V	101	117	29	3
VI	127	145	32	7
VII	59	81	18	2
VIII	19	22	2	0

EXERCISE 2.7.4. *Partitioning Tables.*
The examination of odds ratios and residuals provide two ways to investigate lack of independence in a two-way table. The partitioning methods of Irwin (1949) and Lancaster (1949) provide another. Christensen (1996a, Section 8.6) gives extensive examples of the application of these methods. Table 2.7 gives data on the occupation of family heads for families of various religious groups. The occupations are A – Professions, B – Owners, Managers, and Officials, C – Clerical and Sales, D – Skilled, E – Semiskilled, F – Unskilled, G – Farmers, H – No Occupation. The data were extracted from Lazerwitz (1961). Although the data were collected using a complex sampling design (cf. Section 3.5), ignore this fact in your analysis. To establish the effect of the Protestant groups on the lack of independence, we can isolate the Protestant groups in a separate reduced table. We can also pool the Protestants together in a collapsed table that includes the non-Protestant groups. These are both given in Table 2.8. Test each of the three tables for independence. Note that G^2 for the full table equals the sum of the G^2's for the reduced table and the collapsed table. Continue the analysis of these data by using the partitioning procedure on the reduced and collapsed tables and on subsequent generations of reduced and collapsed tables. Note that tables can also be partitioned on their columns. At its logical extreme, this leads to a collection of 2×2 tables, each with one degree of freedom for testing independence. The Lancaster-Irwin partitioning provides a method of breaking the interaction (lack of independence) G^2 for the full table into one degree of freedom components that add up to the original G^2. This is similar to using orthogonal contrasts to break up the interaction sum of squares in a balanced analysis of variance. For a theoretical justification of the Lancaster-Irwin procedure, see Exercise 8.4.3.

TABLE 2.7. Occupation and Religion

Religion	A	B	C	D	E	F	G	H
White Baptist	43	78	64	135	135	57	86	114
Black Baptist	9	2	9	23	47	77	18	41
Methodist	73	80	80	117	102	58	66	153
Lutheran	23	36	43	59	46	26	49	46
Presbyterian	35	54	38	46	19	22	11	46
Episcopalian	27	27	20	14	7	5	2	15
Roman Catholic	102	140	127	279	254	127	51	190
Jewish	36	60	30	17	17	2	0	26
No Religion	19	12	6	12	25	9	14	28

EXERCISE 2.7.5. *Fisher's Exact Test.*
Consider the problem of testing whether the probability of success is the same for two independent binomials. Let $y_i \sim \text{Bin}(N_i, p_i)$, $i = 1, 2$. Write the 2×2 table as

TABLE 2.8. Partitioned Tables

Reduced Table

Religion	A	B	C	D	E	F	G	H
White Baptist	43	78	64	135	135	57	86	114
Black Baptist	9	2	9	23	47	77	18	41
Methodist	73	80	80	117	102	58	66	153
Lutheran	23	36	43	59	46	26	49	46
Presbyterian	35	54	38	46	19	22	11	46
Episcopalian	27	27	20	14	7	5	2	15

Collapsed Table

Religion	A	B	C	D	E	F	G	H
Protestant	210	277	254	394	356	245	232	415
Roman Catholic	102	140	127	279	254	127	51	190
Jewish	36	60	30	17	17	2	0	26
No Religion	19	12	6	12	25	9	14	28

y_1	$N_1 - y_1$	N_1
y_2	$N_2 - y_2$	N_2
t	$N_1 + N_2 - t$	$N_1 + N_2$

(a) Find $\Pr(y_1 = r_1 \text{ and } t = t_0)$ for arbitrary r_1 and t_0.

(b) Assuming $p_1 = p_2$, find $\Pr(y_1 = r_1 | t = t_0)$.

(c) Consider the following subset of the knee injury data of Example 2.3.1

	Result	
Injury	E	G
Direct	3	2
Twist	7	1

Using the conditional distribution of (b), find the probability of getting the observed value 3. The *P value* for Fisher's exact test is the sum of the $\Pr(y_1 = r_1 | t = 10)$'s for every r_1 value that satisfies

$$\Pr(y_1 = r_1 | t = 10) \leq \Pr(y_1 = 3 | t = 10).$$

Find the *P value* for the data given above. Note that this test does not depend on any large sample approximations, so it is exact even for small samples. On the other hand, the computations become difficult with large samples.

EXERCISE 2.7.6. *Yule's Q.*
For 2×2 tables, a measure of association similar to a correlation coefficient is Yule's Q, which is defined as

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}.$$

Find Q in terms of the odds ratio. Show that Q lies between -1 and 1.

EXERCISE 2.7.7. *Freeman-Tukey Residuals.*
Freeman and Tukey (1950) suggest a variance stabilizing transformation for Poisson data that leads to using the quantities

$$\sqrt{n_{ij}} + \sqrt{n_{ij}+1} - \sqrt{4\hat{m}_{ij}^{(0)}+1}$$

as residuals, cf. Bishop, Fienberg, and Holland (1975, Section 4.4). Reexamine the data of Example 2.3.1 using the Freeman-Tukey residuals.

EXERCISE 2.7.8. *Power Divergence Statistics.*
Cressie and Read (1984) and Read and Cressie (1988) have introduced the *power divergence* family of test statistics

$$2I^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{ij} n_{ij} \left[\left(\frac{n_{ij}}{\hat{m}_{ij}^{(0)}} \right)^{\lambda} - 1 \right],$$

where for $\lambda = -1, 0$ the statistics are defined by taking limits. They establish that for any λ, the large sample distribution under H_0 is χ^2 with the usual degrees of freedom. Show that $X^2 = 2I^1$ and $G^2 = 2I^0$. Find the relationship between $2I^{-1/2}$ and the Freeman-Tukey residuals discussed in Exercise 2.7.7.

EXERCISE 2.7.9. Compute the power divergence test statistics $2I^{-1/2}$ and $2I^{1/2}$ for the knee injury data of Example 2.3.1. Compare the results to G^2 and X^2. What conclusions can be reached about knee injuries?

EXERCISE 2.7.10. *Testing for Symmetry.*
Consider a multinomial sample arranged in an $I \times I$ table. In square tables with similar categories for the two factors, it is sometimes of interest to test

$$H_0 : \; p_{ij} = p_{ji}$$

for all i and j.

(a) Give a procedure for testing this hypothesis based on testing equality of probabilities (homogeneity of proportions) in a $2 \times I(I-1)/2$ table. If you think of the $I \times I$ table as a matrix, the rows indicate whether a cell is above or below the diagonal. The columns are corresponding off diagonal pairs. Illustrate the test for a 4×4 table.

(b) Give a justification for the procedure in terms of a (conditional) sampling model.

(c) The data in Table 2.9 were given by Fienberg (1980), Yule (1900), and earlier by Galton. They report the relative heights of 205 married couples.

TABLE 2.9. Heights of Married Couples

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Test for symmetry and do any other appropriate analysis for these data. Do the data display symmetry?

EXERCISE 2.7.11. *Correlated Data.*
There are actually 410 observations in Exercise 2.7.10 and Table 2.9. There are 205 men and 205 women. Why was Table 2.9 set up as a 3×3 table with only 205 observations rather than as Table 2.10, a 2×3, sex versus height table with 410 observations?

TABLE 2.10. Heights of Married Couples

	Height		
Sex	Tall	Medium	Short
Wife	50	104	51
Husband	60	99	46

EXERCISE 2.7.12. *McNemar's Test.*
McNemar (1947) proposes a method of testing for homogeneity of proportions among two binary populations when the data are correlated. (A binary population is one in which all members fall into one of two categories. Homogeneity means that the proportions in each category are the same for both groups.) If we restrict attention in Exercise 2.7.10 and Table 2.9 to the subpopulation of Tall and Medium people, we get an example of such data. The data on a husband and wife pair cannot be considered as independent, but this problem is avoided by treating each pair as a single response. The data from the subpopulation are given below.

	Wife	
Husband	Tall	Medium
Tall	18	28
Medium	20	51

Conditionally, these data are a multinomial sample of 117. The probability of a tall woman is $p_{11} + p_{21}$ and the probability of a medium woman is

one minus that. The probability of a tall man is $p_{11} + p_{12}$ and again the probability of a medium man can be found by subtraction. It follows that the probability of a tall woman is the same as the probability of a tall man if and only if $p_{21} = p_{12}$. Thus, for 2×2 tables, the problem of homogeneity of proportions is equivalent to testing for symmetry. McNemar's test is just the test for symmetry in Exercise 2.7.10 applied to 2×2 tables. Check for homogeneity of proportions in the subpopulation. For square tables that are larger than 2×2, the problem of testing for *marginal homogeneity* is more difficult and cannot, as yet, be addressed using log-linear models. Nonetheless, a test can be obtained from basic asymptotic results, cf. Exercise 10.8.6.

EXERCISE 2.7.13. Suppose the random variables n_{ij}, $i = 1, 2$, $j = 1, \ldots, N_i$, are independent Poisson(μ_i) random variables. Find the maximum likelihood estimates for μ_1 and μ_2 and find the generalized likelihood ratio test statistic for $H_0 : \mu_1 = \mu_2$.

EXERCISE 2.7.14. Yule's Q (cf. Exercise 2.7.6.) is one of many measures of association that have been proposed for 2×2 tables. Agresti (1984, Chapter 9) has a substantial discussion of measures of association. It has been suggested that measures of association for 2×2 multinomial tables should depend solely on the conditional probabilities of being in the first column given the row, i.e., $p_{11}/p_{1\cdot}$ and $p_{21}/p_{2\cdot}$, or, alternatively, on the conditional probabilities of being in the first row given the column, i.e., $p_{11}/p_{\cdot 1}$ and $p_{12}/p_{\cdot 2}$. Moreover, it has been suggested that the measure of association should not depend on which set of conditional probabilities are used. Show that any measure of association

$$f\left(\frac{p_{11}}{p_{1\cdot}}, \frac{p_{21}}{p_{2\cdot}}\right)$$

can be written as some function of the odds

$$g\left(\frac{p_{11}}{p_{12}}, \frac{p_{21}}{p_{22}}\right).$$

Show that if

$$f\left(\frac{p_{11}}{p_{1\cdot}}, \frac{p_{21}}{p_{2\cdot}}\right) = f\left(\frac{p_{11}}{p_{\cdot 1}}, \frac{p_{12}}{p_{\cdot 2}}\right)$$

for any sets of probabilities, then $g(x, y) = g(ax, ay)$ for any x, y, and a. Use this to conclude that any such measure of association is a function of the odds ratio.

3

Three-Dimensional Tables

Just as a multinomial sample can be classified by the levels of two factors, a multinomial sample can also be classified by the levels of three factors.

EXAMPLE 3.0.1. Everitt (1977) considers a sample of 97 ten-year-old school children who were classified using three factors: classroom behavior, risk of home conditions, and adversity of school conditions. Classroom behavior was judged by teachers to be either nondeviant or deviant. Risk of home conditions either identify the child as not at risk (N) or at risk (R). Adversity of school condition was judged as either low, medium, or high. The observations are denoted as n_{ijk}, $i = 1, 2$, $j = 1, 2$, $k = 1, 2, 3$. The three-dimensional table of n_{ijk}'s is

		Adversity of School (k)						
		Low		Medium		High		
	Risk (j)	N	R	N	R	N	R	Total
Classroom	Nondeviant	16	7	15	34	5	3	80
Behavior (i)	Deviant	1	1	3	8	1	3	17
	Total	17	8	18	42	6	6	97

The totals at the right-hand margin are $n_{1\cdot\cdot} = 80$ and $n_{2\cdot\cdot} = 17$. The totals along the bottom margin are $n_{\cdot 11} = 17$, $n_{\cdot 21} = 8$, $n_{\cdot 12} = 18$, $n_{\cdot 22} = 42$, $n_{\cdot 13} = 6$, $n_{\cdot 23} = 6$, and $n_{\cdots} = 97$.

In general, a three-dimensional table of counts is denoted n_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Marginal totals are denoted

$$n_{ij\cdot} = \sum_{k=1}^{K} n_{ijk}\,, \quad n_{i\cdot k} = \sum_{j=1}^{J} n_{ijk}\,, \quad n_{\cdot jk} = \sum_{i=1}^{I} n_{ijk}\,,$$

$$n_{i\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk} = \sum_{j=1}^{J} n_{ij\cdot} = \sum_{k=1}^{K} n_{i\cdot k}\,,$$

$$n_{\cdot j\cdot} = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ijk}\,, \quad n_{\cdot\cdot k} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ijk}\,,$$

and

$$n_{\cdots} = \sum_{ijk} n_{ijk}\,.$$

Similar notations are used for tables of probabilities p_{ijk}, tables of expected values m_{ijk}, and tables of estimates of the p_{ijk}'s and m_{ijk}'s.

Note that the values $n_{ij\cdot}$ define a two-dimensional $I \times J$ *marginal table*. The values $n_{i\cdot k}$ and $n_{\cdot jk}$ also define marginal tables.

Product-multinomial sampling is raised to a new level of complexity in three-dimensional tables. For example, we could have samples from I populations with each sample cross-classified into JK categories, or we could have samples from IJ (cross-classified) populations where each sample is classified into K categories.

Section 2 of this chapter discusses independence and odds ratio models for three-dimensional tables under multinomial sampling. Section 3 examines the iterative proportional fitting algorithm for finding estimates of expected cell counts. Section 4 introduces log-linear models for three-dimensional tables. Section 5 considers the modifications necessary for dealing with product-multinomial sampling and comments on other sampling schemes. Section 6 introduces model selection criteria and Section 7 introduces tables with four of more dimensions. We begin with a discussion of Simpson's paradox and the need for tables with more than two factors.

3.1 Simpson's Paradox and the Need for Higher-Dimensional Tables

It really is necessary to deal with three-dimensional tables; accurate information cannot generally be obtained by examining each of the three simpler two-dimensional tables. In fact, the conclusions from two-dimensional marginal tables can be contradicted by the accurate three-dimensional information. In this section, we demonstrate and examine the problem via an example.

EXAMPLE 3.1.1. Consider the outcome (success or failure) of two medical treatments classified by the sex of the patient. The data are given below.

		Patient Sex			
		Male		Female	
Outcome		Success	Failure	Success	Failure
Treatment	1	60	20	40	80
	2	100	50	10	30

Considering only the males, we have a two-way table of treatment versus outcome. The estimated probability of success under treatment 1 is $60/80 = .75$. For treatment 2, the estimated probability of success is $100/150 = .667$. Thus, for males, treatment 1 appears to be more successful.

Now consider the table of treatment versus outcome for females only. Under treatment 1, the estimated probability of success is $40/120 = .333$. Under treatment 2, the estimated probability of success is $10/40 = .25$. For women as for men, treatment 1 appears to be more successful.

Now examine the marginal table of treatment versus outcome. This is obtained by collapsing (summing) over the sexes. The table is given below.

		Outcome	
		Success	Failure
Treatment	1	100	100
	2	110	80

The estimated probability of success for treatment 1 is $100/200 = .50$, while the estimated probability of success for treatment 2 is $110/190 = .579$. The marginal table indicates that treatment 2 is better than treatment 1, whereas we know that treatment 1 is better than treatment 2 for both males and females! This contradiction is *Simpson's paradox*.

Simpson's paradox can occur because collapsing can lead to inappropriate weighting of the different populations. Treatment 1 was given to 80 males and 120 females, so the marginal table is indicating a success rate for treatment 1 that is a weighted average of the success rates for males and females with slightly more weight given to the females. Treatment 2 was given to 150 males and only 40 females, so the marginal success rate is a weighted average of the male and female success rates with most of the weight given to the male success rate. It is only a slight oversimplification to say that the marginal table is comparing a success rate for treatment 1 that is the mean of the male and female success rates, to a success rate for treatment 2 that is essentially the male success rate. Since the success rate for males is much higher than it is for females, the marginal table gives the illusion that treatment 2 is better.

The moral of all this is that one cannot necessarily trust conclusions drawn from marginal tables. It is generally necessary to consider all the dimensions of a table. Situations in which marginal (collapsed) tables yield valid conclusions are discussed in Section 5.3.

3.2 Independence and Odds Ratio Models

For multinomial sampling and a two-dimensional table, there was only one model of primary interest: independence of rows and columns. With three-dimensional tables, there are at least eight interesting models. Half of these are very easy to imagine. If we refer to the three dimensions of the table as rows, columns, and layers, we can have (0) rows, columns, and layers all independent, (1) rows independent of columns and layers (but columns and layers not necessarily independent), (2) columns independent of rows and layers, and (3) layers independent of rows and columns. Three of the remaining four models involve conditional independence: (4) given any particular layer, rows and columns are independent, (5) given any column, rows and layers are independent, and (6) given any row, columns and layers are independent. The last of the eight models is that certain odds ratios are equal. Section 4 discusses these models in relation to log-linear models.

We now examine the models in detail. The essential part of the log-likelihood for multinomial sampling is $\ell(p) \equiv \sum_{i,j,k} n_{ijk} \log(p_{ijk})$.

For all of the (conditional) independence models, the MLEs can be obtained using Lemma 2.4.1. The trick is to break $\ell(p)$ into a sum of terms, each of which can be maximized separately. The discussion below emphasizes a more general approach to finding MLEs.

3.2.1 The Model of Complete Independence

To put it briefly, the model of *complete independence* is that everything (rows, columns, and layers) is independent of everything else, cf. Example 1.1.3. Technically, the model is

$$M^{(0)}: p_{ijk} = p_{i\cdot\cdot} p_{\cdot j\cdot} p_{\cdot\cdot k}$$

where the superscript (0) is used to distinguish this model from the other models that will be considered.

The MLE of p_{ijk} under this model is

$$\begin{aligned} \hat{p}^{(0)}_{ijk} &= \hat{p}_{i\cdot\cdot} \hat{p}_{\cdot j\cdot} \hat{p}_{\cdot\cdot k} \\ &= (n_{i\cdot\cdot}/n_{\cdots})(n_{\cdot j\cdot}/n_{\cdots})(n_{\cdot\cdot k}/n_{\cdots})\,. \end{aligned}$$

Since $m_{ijk} = n_{...}p_{ijk}$, the MLE of m_{ijk} is

$$\begin{aligned}\hat{m}_{ijk}^{(0)} &= n_{...}\hat{p}_{ijk}^{(0)} \\ &= n_{i..}n_{.j.}n_{..k}/n_{...}^2 .\end{aligned}$$

This is another application of the general result, discussed in Section 2.4, that the MLE of a function of the parameters is just the function applied to the MLEs. The Pearson chi-square statistic for testing lack of fit of $M^{(0)}$ is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(n_{ijk} - \hat{m}_{ijk}^{(0)}\right)^2}{\hat{m}_{ijk}^{(0)}} .$$

The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(0)}\right) .$$

An α level test is rejected if the test statistic is greater than $\chi^2(1-\alpha, IJK - I - J - K + 2)$. As in Section 2.4, the maximum likelihood estimate of m_{ijk} without any restrictions is $\hat{m}_{ijk} = n_{ijk}$; thus, n_{ijk} is used in the formula for G^2. Degrees of freedom for the tests given in this section will be discussed in Section 4

Chapter 10 establishes that the MLEs for $M^{(0)}$ are characterized by

$$\begin{aligned}\hat{m}_{ijk} &= n_{...}\hat{p}_{i..}\hat{p}_{.j.}\hat{p}_{..k} \\ &= \frac{\hat{m}_{i..}\hat{m}_{.j.}\hat{m}_{..k}}{n_{...}^2}\end{aligned}$$

and the marginal constraints $\hat{m}_{i..} = n_{i..}$, $\hat{m}_{.j.} = n_{.j.}$, and $\hat{m}_{..k} = n_{..k}$. In other words, any set of values $\hat{m}_{ijk}$ that satisfy the marginal constraints and satisfy model $M^{(0)}$ must be the maximum likelihood estimates. The estimates $\hat{m}_{ijk}^{(0)}$ given above are, under weak restrictions on the n_{ijk}'s, the unique values that satisfy both sets of conditions.

Example 3.2.1. In the longitudinal study mentioned in Example 2.2.1, out of 3182 people without cardiovascular disease, 2121 neither exercised regularly nor developed cardiovascular disease during the $4\frac{1}{2}$-year study. We restrict our attention to these 2121 individuals. The subjects were cross-classified by three factors: Personality type (A,B), Cholesterol level (normal, high), and Diastolic Blood Pressure (normal, high). The data are

n_{ijk}		Diastolic Blood Pressure	
Personality	Cholesterol	Normal	High
A	Normal	716	79
	High	207	25
B	Normal	819	67
	High	186	22

The fitted values assuming complete independence are

$\hat{m}^{(0)}_{ijk}$		Diastolic Blood Pressure	
Personality	Cholesterol	Normal	High
A	Normal	739.9	74.07
	High	193.7	19.39
B	Normal	788.2	78.90
	High	206.3	20.65

Note that the $\hat{m}_{ijk}$'s satisfy the property that $n_{i..} = \hat{m}_{i..}$, $n_{.j.} = \hat{m}_{.j.}$, and $n_{..k} = \hat{m}_{..k}$ for all i, j, and k. For example, the Type A totals are $n_{1..} = 716+79+207+25 = 1027$ and $\hat{m}_{1..} = 739.9+74.07+193.7+19.39 = 1027.06$. The difference is roundoff error. Pearson's chi-square is

$$X^2 = \frac{(716 - 739.9)^2}{739.9} + \cdots + \frac{(22 - 20.65)^2}{20.65} = 8.730\,.$$

The likelihood ratio chi-square is

$$G^2 = 2\,[716\log(716/739.9) + \cdots + 22\log(22/20.65)] = 8.723\,.$$

The degrees of freedom for either chi-square test are

$$df = (2)(2)(2) - 2 - 2 - 2 + 2 = 4\,.$$

Since $\chi^2(.95, 4) = 9.49$, an $\alpha = .05$ level test will not reject the hypothesis of independence. In particular, the *P value* is .07. There is no clear evidence of any relationships among personality type, cholesterol level, and diastolic blood pressure level for these people who do not exercise regularly and do not have cardiovascular disease.

Although the test statistics give no clear evidence that complete independence does not hold, similarly they give no great confidence that complete independence is a good model. Deviations from independence can be examined using the Pearson residuals

$$\tilde{r}_{ijk} = \frac{n_{ijk} - \hat{m}_{ijk}}{\sqrt{\hat{m}_{ijk}}}\,.$$

The Pearson residuals for these data are

$\tilde{r}_{ijk}$		Diastolic Blood Pressure	
Personality	Cholesterol	Normal	High
A	Normal	−0.879	0.573
	High	0.956	1.274
B	Normal	1.097	−1.340
	High	−1.413	0.297

In particular, the residual for Type A, Normal, Normal is $-.879 = (716 - 739.9)/\sqrt{739.9}$. Relative to complete independence, high blood pressure and high cholesterol are overrepresented in Type A personalities (those showing signs of stress) and normal blood pressure and cholesterol are overrepresented in Type B personalities (relaxed individuals). These results agree well with conventional wisdom. Note also that for Type B personalities, individuals with only one high categorization are underrepresented.

The patterns in the residuals are interesting, but remember that there is no clear evidence (at this point) for rejecting the hypothesis of complete independence.

3.2.2 Models with One Factor Independent of the Other Two

With three factors, there are three ways in which one factor can be independent of the other two. For example, we can have rows independent of columns and layers, cf. Example 1.1.4. This model says nothing about the relationship between columns and layers. Columns and layers can either be independent or not independent. If they were not independent, typically we would be interested in examining how they differ from independence.

Specifically, the three models are rows independent of columns and layers,

$$M^{(1)}: p_{ijk} = p_{i\cdot\cdot}p_{\cdot jk} \ ,$$

columns independent of rows and layers,

$$M^{(2)}: p_{ijk} = p_{\cdot j\cdot}p_{i\cdot k} \ ,$$

and layers independent of rows and columns,

$$M^{(3)}: p_{ijk} = p_{\cdot\cdot k}p_{ij\cdot} \ .$$

All three of these models include the model of complete independence $M^{(0)}$ as a special case. If $M^{(0)}$ is true, then all three of these are true. The analyses for all three models are similar; we will consider only $M^{(1)}$ in detail.

Under $M^{(1)}$, no distinction is drawn between columns and layers. In fact, this model is equivalent to independence in an $I \times (JK)$ two-dimensional table where the columns of the two-dimensional table consist of all combinations of the columns and layers of the three-dimensional table.

EXAMPLE 3.2.2. Consider again the classroom behavior data of Example 3.0.1. The test of $M^{(1)}$ is simply a test of the independence of the two rows: nondeviant, deviant, and the six columns: Low-N, Low-R, Medium-N, Medium-R, High-N, High-R.

From our results in Chapter 2, the MLE of p_{ijk} under $M^{(1)}$ is

$$\begin{aligned} \hat{p}^{(1)}_{ijk} &= \hat{p}_{i\cdot\cdot}\hat{p}_{\cdot jk} \\ &= (n_{i\cdot\cdot}/n_{\cdots})(n_{\cdot jk}/n_{\cdots}) \ . \end{aligned}$$

The MLE of m_{ijk} is

$$\begin{aligned}\hat{m}_{ijk}^{(1)} &= n_{\cdots}\hat{p}_{ijk}^{(1)} \\ &= n_{i\cdot\cdot}n_{\cdot jk}/n_{\cdots}\,.\end{aligned}$$

The superscript (1) in $\hat{p}_{ijk}^{(1)}$ and $\hat{m}_{ijk}^{(1)}$ is used to indicate that these estimates are obtained assuming that $M^{(1)}$ is true. The Pearson chi-square test statistic for $M^{(1)}$ is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(n_{ijk}-\hat{m}_{ijk}^{(1)}\right)^2}{\hat{m}_{ijk}^{(1)}}\,.$$

The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(1)}\right)\,.$$

These are compared to percentage points of a chi-square distribution with degrees of freedom

$$\begin{aligned}df &= (I-1)(JK-1) \\ &= IJK - I - JK + 1\,.\end{aligned}$$

Similar to $M^{(0)}$, the MLEs for model $M^{(1)}$ are any values $\hat{m}_{ijk}$ that satisfy $M^{(1)}$ and the marginal constraints

$$\hat{m}_{i\cdot\cdot} = n_{i\cdot\cdot} \quad \text{and} \quad \hat{m}_{\cdot jk} = n_{\cdot jk}\,.$$

Again, the values $\hat{m}_{ijk}^{(1)}$ given above are the unique MLEs under mild restrictions on the n_{ijk}'s. It is interesting to note that choosing $\hat{m}_{ijk} = n_{ijk}$ satisfies the marginal constraints but, except in the most bizarre cases, does not satisfy model $M^{(1)}$. On the other hand, taking the estimated cell counts from the complete independence model $\hat{m}_{ijk} = n_{i\cdot\cdot}n_{\cdot j\cdot}n_{\cdot\cdot k}/n_{\cdots}^2$ satisfies $M^{(1)}$ but typically does not satisfy the marginal constraints.

EXAMPLE 3.2.2, CONTINUED. The table of $\hat{m}_{ijk}^{(1)}$'s is

$\hat{m}_{ijk}^{(1)}$		Adversity (k)						
		Low		Medium		High		
	Risk (j)	N	R	N	R	N	R	$\hat{m}_{i\cdot\cdot}$
Classroom	Non.	14.02	6.60	14.85	34.64	4.95	4.95	80
Behavior (i)	Dev.	2.98	1.40	3.15	7.36	1.05	1.05	17
	$\hat{m}_{\cdot jk}$	17	8	18	42	6	6	97

As displayed in the margins of the n_{ijk} and $\hat{m}^{(1)}_{ijk}$ tables, the MLEs satisfy the conditions $\hat{m}^{(1)}_{i\cdot\cdot} = n_{i\cdot\cdot}$ and $\hat{m}^{(1)}_{\cdot jk} = n_{\cdot jk}$.

The Pearson chi-square test statistic is

$$X^2 = 6.19 = \frac{(16 - 14.02)^2}{14.02} + \cdots + \frac{(3 - 1.05)^2}{1.05}\,.$$

The likelihood ratio test statistic is

$$G^2 = 5.56 = 2\left[16\log(16/14.02) + \cdots + 3\log(3/1.05)\right]\,.$$

The degrees of freedom for the chi-square test are

$$\begin{aligned} df &= (2-1)[(2)(3)-1] \\ &= 5\,. \end{aligned}$$

The 95th percentile of a chi-square with 5 degrees of freedom is

$$\chi^2(.95, 5) = 11.07\,.$$

Both X^2 and G^2 are less than $\chi^2(.95, 5)$, so an $\alpha = .05$ level test provides no evidence against $M^{(1)}$. In other words, we have no reason to doubt that classroom behavior is independent of risk and adversity.

It is quite possible in this study that our primary interest would be in explaining classroom behavior in terms of risk and adversity. Unfortunately, classroom behavior seems to be independent of both of the variables with which we were trying to explain it. On the other hand, examining the relationship between risk and adversity becomes very simple. If classroom behavior is independent of risk and adversity, we can study the marginal table of risk and adversity without worrying about Simpson's paradox. The marginal table of counts is

$n_{\cdot jk}$		Adversity (k)			
		Low	Medium	High	$n_{\cdot j\cdot}$
Risk (j)	N	17	18	6	41
	R	8	42	6	56
	$n_{\cdot\cdot k}$	25	60	12	97

The model of independence for this marginal table is

$$M:\ p_{\cdot jk} = p_{\cdot j\cdot}p_{\cdot\cdot k}, \qquad j = 1, \ldots, J\,,\ k = 1, \ldots, K\,.$$

The expected counts for the marginal table under M are

$$\begin{aligned} \hat{m}_{\cdot jk} &= n_{\cdots}\hat{p}_{\cdot j\cdot}\hat{p}_{\cdot\cdot k} \\ &= n_{\cdot j\cdot}n_{\cdot\cdot k}/n_{\cdots}\,. \end{aligned}$$

The table of estimated expected counts is

		Adversity (k)			
		Low	Medium	High	$\hat{m}_{\cdot j \cdot}$
Risk (j)	N	10.57	25.36	5.07	41
	R	14.43	34.64	6.93	56
	$\hat{m}_{\cdot\cdot k}$	25	60	12	97

yielding

$$X^2 = 10.78 \quad \text{and} \quad G^2 = 10.86$$

on

$$df = (2-1)(3-1) = 2\,.$$

Both X^2 and G^2 are significant at the $\alpha = .01$ level.

Either the residuals or the odds ratios can be used to explore the lack of independence. The table of residuals is

		Adversity (k)		
		Low	Medium	High
Risk (j)	N	1.98	−1.46	0.35
	R	−1.69	1.25	−0.35

For highly adverse schools, the residuals are near zero, so independence seems to hold. For schools with low adversity (i.e., good schools), the not-at-risk students are overrepresented and at-risk students are underrepresented. For schools with medium adversity, the at-risk students are overrepresented and the not-at-risk students are underrepresented. (I wonder if the criteria for determining whether a student is at risk may have been applied differently to students in high-adversity schools.)

Using odds ratios, we see that the odds of being not at risk for low-adversity schools (17/8) are about five times greater than for medium-adversity schools (18/42). In particular, the odds ratio is

$$\frac{(17)(42)}{(8)(18)} = 4.96\,.$$

The odds of being not at risk in a low-adversity school are only about twice as large as the odds of being not at risk in a high-adversity school [i.e., $(17)(6)/(8)(6) = 2.125$]. Finally, the odds of being not at risk in a medium-adversity school are only about half as large [$18(6)/42(6) = .429$] as the odds of being not at risk in a high-adversity school. Of course, the sample is small, so there is quite a bit of variability associated with these estimated odds ratios.

Before leaving this example, it is of interest to note a relationship between the two likelihood ratio test statistics that were considered. The models and statistics are

$$M^{(1)}\colon p_{ijk} \quad = \quad p_{i\cdot\cdot}p_{\cdot jk}\,, \qquad G^2 = 5.56\,, \qquad df = 5$$

$$M: p_{\cdot jk} = p_{\cdot j\cdot} p_{\cdot\cdot k}, \qquad G^2 = 10.86, \qquad df = 2.$$

Taken together, these models imply that

$$M^{(0)}: p_{ijk} = p_{i\cdot\cdot} p_{\cdot j\cdot} p_{\cdot\cdot k}$$

holds. The likelihood ratio test statistic for $M^{(0)}$ with these data is

$$G^2 = 16.42$$

with 7 degrees of freedom. As will be seen later, it is no accident that the test statistics G^2 satisfy

$$5.56 + 10.86 = 16.42$$

and that the degrees of freedom satisfy $5 + 2 = 7$.

EXERCISE 3.1. Examine the residuals from fitting $M^{(1)}$. Are any of them large enough to call in question the further analysis that was based on tentatively assuming that $M^{(1)}$ was true?

3.2.3 Models of Conditional Independence

Given that one is at a particular level of some factor, the other two factors could be independent. For example, for any given category in the levels, the rows and the columns may be independent, cf. Example 1.1.5. By the definition of conditional probability, the probability of row i and column j given that the layer is k is

$$\begin{aligned} &\Pr(\text{row} = i, \text{col} = j \mid \text{layer} = k) \\ &\quad = \Pr(\text{row} = i, \text{col} = j, \text{layer} = k)/\Pr(\text{layer} = k) \\ &\quad = p_{ijk}/p_{\cdot\cdot k}\,. \end{aligned} \tag{1}$$

Conditional independence of rows and columns for each layer means that for all i, j, and k

$$\begin{aligned} &\Pr(\text{row} = i, \text{col} = j \mid \text{layer} = k) \\ &\quad = \Pr(\text{row} = i | \text{layer} = k)\Pr(\text{col} = j | \text{layer} = k) \\ &\quad = (p_{i\cdot k}/p_{\cdot\cdot k})(p_{\cdot jk}/p_{\cdot\cdot k})\,. \end{aligned} \tag{2}$$

Assuming that every layer has a possibility of occurring (i.e., $p_{\cdot\cdot k} > 0$ for all k), then the model of conditional independence can be rewritten. Setting (1) and (2) equal and multiplying both sides by $p_{\cdot\cdot k}$ gives the requirement

$$p_{ijk} = p_{i\cdot k}\, p_{\cdot jk}/p_{\cdot\cdot k}$$

for independence of rows and columns given layers.

The nature of the conditional independence between rows and columns may or may not depend on the particular layer. For example, if rows, columns, and layers are all independent, then $M^{(0)}$ holds and for any layer k

$$\Pr(\text{row} = i,\ \text{col} = j \mid \text{layer} = k) = p_{i\cdot\cdot}p_{\cdot j\cdot}\,.$$

This does not depend on the layer. However, if rows are independent of columns and layers so that $M^{(1)}$ holds, then

$$\Pr(\text{row} = i,\ \text{col} = j \mid \text{layer} = k) = p_{i\cdot\cdot}(p_{\cdot jk}/p_{\cdot\cdot k})\,.$$

The column probabilities depend on the layer, but the row probabilities do not. Similarly, for $M^{(2)}$, the columns are independent of rows and layers; thus,

$$\Pr(\text{row} = i,\ \text{col} = j \mid \text{layer} = k) = (p_{i\cdot k}/p_{\cdot\cdot k})p_{\cdot j\cdot}\,.$$

The row structure depends on layers, but the column structure does not. Of course, the most interesting case of rows and columns independent given layers is when none of these simpler cases apply.

If two factors are to be independent given the third factor, there are three ways in which the conditioning factor can be chosen. This leads to three models: rows and columns independent given layers

$$M^{(4)}\colon p_{ijk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}\,,$$

rows and layers independent given columns

$$M^{(5)}\colon p_{ijk} = p_{ij\cdot}p_{\cdot jk}/p_{\cdot j\cdot}\,,$$

and columns and layers independent given rows

$$M^{(6)}\colon p_{ijk} = p_{ij\cdot}p_{i\cdot k}/p_{i\cdot\cdot}\,.$$

As in the previous subsection, the analyses for all three models are similar. We consider only $M^{(4)}$ in detail. The MLE for p_{ijk} is

$$\begin{aligned}\hat{p}^{(4)}_{ijk} &= \hat{p}_{i\cdot k}\hat{p}_{\cdot jk}/\hat{p}_{\cdot\cdot k}\\ &= (n_{i\cdot k}/n_{\cdots})(n_{\cdot jk}/n_{\cdots})\Big/(n_{\cdot\cdot k}/n_{\cdots})\\ &= n_{i\cdot k}n_{\cdot jk}/n_{\cdot\cdot k}n_{\cdots}\,.\end{aligned}$$

The MLE for $m_{ijk} = n_{\cdots}p_{ijk}$ is

$$\begin{aligned}\hat{m}^{(4)}_{ijk} &= n_{\cdots}\hat{p}_{i\cdot k}\hat{p}_{\cdot jk}/\hat{p}_{\cdot\cdot k}\\ &= n_{i\cdot k}n_{\cdot jk}/n_{\cdot\cdot k}\,.\end{aligned}$$

The MLEs are any numbers $\hat{m}_{ijk}$ that satisfy model $M^{(4)}$ and the marginal relations $\hat{m}_{i\cdot k} = n_{i\cdot k}$, $\hat{m}_{\cdot jk} = n_{\cdot jk}$, and (redundantly) $\hat{m}_{\cdot\cdot k} = n_{\cdot\cdot k}$. The test statistics are

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \hat{m}_{ijk}^{(4)}\right)^2}{\hat{m}_{ijk}^{(4)}}$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(4)}\right) .$$

The tests actually pool the K separate tests for independence of rows and columns which are computed for each individual layer. There are $(I-1)(J-1)$ degrees of freedom for the test at each layer. The degrees of freedom for the pooled test is

$$df = (I-1)(J-1)K .$$

EXAMPLE 3.2.3. Consider again the data of Example 3.2.1. In that example, we found that the *P value* for testing complete independence of personality type, cholesterol level, and diastolic blood pressure level was .067. Our residual analysis pointed out some interesting differences between personality types. We now examine the model $M^{(6)}$ that cholesterol level and diastolic blood pressure level are independent given personality type. The test is really a simultaneous test of whether independence holds in each of the tables given below.

	(Personality Type A)	
n_{1jk}	Diastolic Blood Pressure	
Cholesterol	Normal	High
Normal	716	79
High	207	25

	(Personality Type B)	
n_{2jk}	Diastolic Blood Pressure	
Cholesterol	Normal	High
Normal	819	67
High	186	22

Each table has $(2-1)(2-1) = 1$ degree of freedom, so the overall test has 2 degrees of freedom. The table of estimated cell counts under conditional independence is

$\hat{m}_{ijk}^{(6)}$			Diastolic Blood Pressure	
		Cholesterol	Normal	High
Personality	A	Normal	714.5	80.51
		High	208.5	23.49
	B	Normal	813.9	72.08
		High	191.1	16.92

giving

$$X^2 = 2.188 \quad \text{and} \quad G^2 = 2.062$$

and

$$df = 2\,.$$

This is a very good fit. Given the personality type, there seems to be no relationship between cholesterol level and diastolic blood pressure level. As in the previous examples, observe that the estimates $\hat{m}_{ijk}^{(6)}$ satisfy the *likelihood equations* $\hat{m}_{i\cdot k}^{(6)} = n_{i\cdot k}$ and $\hat{m}_{ij\cdot}^{(6)} = n_{ij\cdot}$. (The likelihood equations are just the marginal constraints.)

Note that the odds of being normal in either cholesterol or blood pressure is higher for Type B personalities than for Type A personalities. If for each personality type, cholesterol and blood pressure are independent, we can examine the relationship between either personality and cholesterol or between personality and blood pressure from the appropriate marginal table, cf. Section 5.3. For example, to examine personality and cholesterol, the marginal table is

	Cholesterol	
Personality	Normal	High
A	795	232
B	886	208

The odds of having normal cholesterol for Type A personalities is $795/232 = 3.427$. The odds of having normal cholesterol for Type B personalities is $886/208 = 4.260$. The odds ratio is

$$\frac{\hat{p}_{11\cdot}\hat{p}_{22\cdot}}{\hat{p}_{12\cdot}\hat{p}_{21\cdot}} = \frac{795(208)}{232(886)} = .804\,.$$

The odds of having a normal cholesterol level with personality Type A are only about 80% as large as the odds for personality Type B.

A similar analysis shows that the odds of having a normal diastolic blood pressure level with personality Type A is 78.6% of the odds for personality Type B. Although this odds ratio of .786 is further from one than the odds ratio for cholesterol, it turns out to be less significant. The variabilities of these point estimates depend on the sample sizes in all the cells. The personality–blood pressure marginal table has some smaller cells than the

cholesterol–blood pressure table; thus, the personality–blood pressure odds ratio is subject to more variability. We will see in Example 3.4.1 that one could reasonably take personality and blood pressure to be independent, but personality and cholesterol are not independent.

3.2.4 A Final Model for Three-Way Tables

The last of the standard models for three-way tables is due to Bartlett (1935) and must be stated in terms of odds ratios. To look at an odds ratio in a three-way table, one fixes a factor and looks at the odds ratio relating the other two factors. For example, we can fix layers and look at the odds ratio $p_{11k}p_{ijk}/p_{1jk}p_{i1k}$. The last of our models is that these odds ratios are the same for every layer. In particular, the model is

$$M^{(7)} : \frac{p_{111}p_{ij1}}{p_{i11}p_{1j1}} = \frac{p_{11k}p_{ijk}}{p_{i1k}p_{1jk}}$$

for all $i = 2, \ldots, I$, $j = 2, \ldots, J$, and $k = 2, \ldots, K$. $M^{(7)}$ is stated as if layers are fixed, but, in fact, it is easily shown that the model is unchanged if stated for rows fixed or columns fixed.

There are no simple formulae for $\hat{p}_{ijk}^{(7)}$ or $\hat{m}_{ijk}^{(7)}$. Iterative computing methods (cf. Section 3) must be used to obtain the MLEs. It can be shown that the MLEs must satisfy the marginal constraints $\hat{m}_{ij\cdot} = n_{ij\cdot}$, $\hat{m}_{i\cdot k} = n_{i\cdot k}$, $\hat{m}_{\cdot jk} = n_{\cdot jk}$ and also the model; i.e., we need $\hat{m}_{111}\hat{m}_{ij1}/\hat{m}_{i11}\hat{m}_{1j1} = \hat{m}_{11k}\hat{m}_{ijk}/\hat{m}_{i1k}\hat{m}_{1jk}$ for $i, j, k \geq 2$. Given the MLEs, the test statistics are computed as usual.

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \hat{m}_{ijk}^{(7)}\right)^2}{\hat{m}_{ijk}^{(7)}}$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(7)}\right) .$$

Although there is, at this point, no obvious reason for this figure, the degrees of freedom for the chi-square test is

$$df = (I-1)(J-1)(K-1) .$$

EXAMPLE 3.2.4. Fienberg (1980) and Kihlberg, Narragon and Campbell (1964) report data on severity of drivers' injuries in auto accidents along with the type of accident and whether or not the driver was ejected from the vehicle during the accident. We consider the results only for small cars. The data are listed below.

n_{ijk}		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver Ejected (i)	No	350	150	60	112
	Yes	26	23	19	80

Since $I = J = K = 2$, the model becomes

$$M^{(7)}: \frac{p_{111}p_{221}}{p_{121}p_{211}} = \frac{p_{112}p_{222}}{p_{122}p_{212}}.$$

Using the methods of Section 3, the MLEs of the m_{ijk}'s are found.

$\hat{m}_{ijk}^{(7)}$		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver Ejected (i)	No	350.5	149.5	59.51	112.5
	Yes	25.51	23.49	19.49	79.51

The test statistics have $(2-1)(2-1)(2-1) = 1$ degree of freedom and are

$$X^2 = .04323 \quad \text{and} \quad G^2 = .04334\,.$$

The model of equality of odds ratios fits the data remarkably well. The reader can verify that none of the models involving independence or conditional independence fit the data.

Another way to examine $M^{(7)}$ is to look at the estimated odds ratios and see if they are about equal. For this purpose, we use the unrestricted estimates of the p_{ijk}'s, i.e., $\hat{p}_{ijk} = n_{ijk}/n_{...}$. The estimated odds ratios are

$$\begin{aligned} \hat{p}_{111}\hat{p}_{221}/\hat{p}_{121}\hat{p}_{211} &= 350(23)/26(150) \\ &= 2.064 \end{aligned}$$

and

$$\begin{aligned} \hat{p}_{112}\hat{p}_{222}/\hat{p}_{122}\hat{p}_{212} &= 60(80)/19(112) \\ &= 2.256\,. \end{aligned}$$

These values are quite close.

In summary, for both collisions and rollovers, the odds of a severe injury are about twice as large if the driver is ejected from the vehicle than if not. Equivalently, the odds of having a nonsevere injury are about twice as great if the driver is not ejected from the vehicle than if the driver is ejected. It should be noted that the odds of being severely injured in a rollover are consistently much higher than in a collision. What we have concluded in our analysis of $M^{(7)}$ is that the *relative* effect of the driver being ejected is the same for both types of accident and that being ejected substantially increases one's chances of being severely injured. So you see, it really does pay to wear seat belts.

3.2.5 Odds Ratios and Independence Models

For a two-dimensional table, Proposition 2.3.3 established that the model of independence was equivalent to the model that all odds ratios equal one. Similarly, all eight of the models discussed for three-dimensional tables can be written in terms of odds ratios. So far, only our characterization of $M^{(7)}$ is in terms of odds ratios. Since models $M^{(1)}$, $M^{(2)}$, and $M^{(3)}$ are similar, we will examine only $M^{(1)}$. Similarly, we will consider $M^{(4)}$ as representative of $M^{(4)}$, $M^{(5)}$, and $M^{(6)}$.

To begin, consider $M^{(4)}$. If $M^{(4)}$ is true, then $p_{ijk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}$; $M^{(4)}$ is the model that has rows and columns independent given layers. Let us consider an odds ratio with layers fixed. We can write a typical odds ratio as $p_{ijk}p_{i'j'k}/p_{ij'k}p_{i'jk}$. Assuming $M^{(4)}$ with positive p_{ijk}'s, we get

$$\begin{aligned}\frac{p_{ijk}p_{i'j'k}}{p_{ij'k}p_{i'jk}} &= \frac{(p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k})(p_{i'\cdot k}p_{\cdot j'k}/p_{\cdot\cdot k})}{(p_{i\cdot k}p_{\cdot j'k}/p_{\cdot\cdot k})(p_{i'\cdot k}p_{\cdot jk}/p_{\cdot\cdot k})} \\ &= \frac{p_{i\cdot k}p_{\cdot jk}p_{i'\cdot k}p_{\cdot j'k}}{p_{i\cdot k}p_{\cdot j'k}p_{i'\cdot k}p_{\cdot jk}} \\ &= 1\,.\end{aligned}$$

Thus, for a fixed value of k, the odds ratio is one.

Conversely, if $p_{ijk}p_{i'j'k}/p_{ij'k}p_{i'jk} = 1$ for all i, i', j, and j', then

$$\begin{aligned}p_{ijk}p_{\cdot\cdot k} &= \sum_{i',j'} p_{ijk}p_{i'j'k} = \sum_{i'j'} p_{ij'k}p_{i'jk} = \sum_{j'} p_{ij'k}\sum_{i'} p_{i'jk} \\ &= p_{i\cdot k}p_{\cdot jk},\end{aligned}$$

so $M^{(4)}$ holds.

It is also of interest to note that, under $M^{(4)}$, odds ratios with the row (column) fixed are equal for all rows (columns). In particular,

$$\begin{aligned}\frac{p_{ijk}p_{ij'k'}}{p_{ijk'}p_{ij'k}} &= \frac{(p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k})(p_{i\cdot k'}p_{\cdot j'k'}/p_{\cdot\cdot k'})}{(p_{i\cdot k'}p_{\cdot jk'}/p_{\cdot\cdot k'})(p_{i\cdot k}p_{\cdot j'k}/p_{\cdot\cdot k})} \\ &= \frac{p_{i\cdot k}p_{\cdot jk}p_{i\cdot k'}p_{\cdot j'k'}}{p_{i\cdot k'}p_{\cdot jk'}p_{i\cdot k}p_{\cdot j'k}} \\ &= \frac{p_{\cdot jk}p_{\cdot j'k'}}{p_{\cdot jk'}p_{\cdot j'k}}\,.\end{aligned} \tag{3}$$

Thus, the odds ratio does not depend on i and must be the same for each row. These facts imply that $M^{(4)}$ is a special case of $M^{(7)}$.

Perhaps the simplest way to examine odds ratios in relation to $M^{(1)}$ is to use the results obtained for $M^{(4)}$. The following proposition allows this.

Proposition 3.2.5. $M^{(1)}$ is true if and only if both $M^{(4)}$ and $M^{(5)}$ are true.

Proof. If $M^{(1)}$ is true, then $p_{ijk} = p_{i\cdot\cdot}p_{\cdot jk}$. Summing over j gives

$p_{i\cdot k} = p_{i\cdot\cdot}p_{\cdot\cdot k}$; thus, $p_{i\cdot\cdot} = p_{i\cdot k}/p_{\cdot\cdot k}$. Substitution yields

$$p_{ijk} = p_{i\cdot\cdot}p_{\cdot jk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}\,;$$

thus, $M^{(4)}$ holds. A similar argument summing over k shows that $M^{(5)}$ holds.

Conversely, if both $M^{(4)}$ and $M^{(5)}$ are true, then

$$p_{ijk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}$$

and

$$p_{ijk} = p_{ij\cdot}p_{\cdot jk}/p_{\cdot j\cdot}\,.$$

It follows that

$$\frac{p_{ijk}}{p_{\cdot jk}} = \frac{p_{i\cdot k}}{p_{\cdot\cdot k}} = \frac{p_{ij\cdot}}{p_{\cdot j\cdot}}$$

and from the last equality,

$$p_{i\cdot k}p_{\cdot j\cdot} = p_{\cdot\cdot k}p_{ij\cdot}\,.$$

Summing over j gives

$$p_{i\cdot k}p_{\cdot\cdot\cdot} = p_{\cdot\cdot k}p_{i\cdot\cdot};$$

recalling that $p_{\cdot\cdot\cdot} = 1$ and rearranging terms gives

$$p_{i\cdot\cdot} = p_{i\cdot k}/p_{\cdot\cdot k}\,.$$

Substituting this into $M^{(4)}$ gives

$$\begin{aligned} p_{ijk} &= p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k} \\ &= p_{i\cdot\cdot}p_{\cdot jk} \end{aligned}$$

and $M^{(1)}$ holds. □

It follows from Proposition 3.2.5 and the discussion of $M^{(4)}$ that the model $M^{(1)}$ is equivalent to

$$p_{ijk}p_{i'j'k}/p_{ij'k}p_{i'jk} = 1 \qquad \text{(layers fixed)}$$

for all i, i', j, j', and k, and

$$p_{ijk}p_{i'jk'}/p_{ijk'}p_{i'jk} = 1 \qquad \text{(columns fixed)}$$

for all i, i', k, k', and j. In addition, all odds ratios with rows fixed will be equal (but not necessarily equal to one). $M^{(1)}$ is thus a special case of $M^{(7)}$.

Similar arguments establish that if $M^{(0)}$ is true, then all odds ratios equal one regardless of whether rows, columns, or layers have been fixed.

Finally, note that as in Chapter 2, odds ratios remain unchanged when the p_{ijk}'s are all replaced with m_{ijk}'s.

3.3 Iterative Computation of Estimates

To fit the model $M^{(7)}$ that all odds ratios are equal, we need to compute the $\hat{m}_{ijk}^{(7)}$'s. There are two standard algorithms for doing this: the *Newton-Raphson algorithm* and the *iterative proportional fitting algorithm*. Newton-Raphson amounts to doing a series of weighted regression analyses. It is commonly referred to as *iteratively reweighted least squares*. The Newton-Raphson algorithm is discussed in Section 10.5. Iterative proportional fitting was introduced by Deming and Stephan (1940) for purposes other than fitting models to discrete data, but the algorithm gives maximum likelihood estimates for the models discussed in the previous section and for the balanced ANOVA type log-linear models that will be discussed later. Meyer (1982) presents methods of transforming various other log-linear models so that iterative proportional fitting can be applied.

In this section, we describe the method of iterative proportional fitting for finding the $\hat{m}_{ijk}^{(7)}$'s. The method can be easily extended to find estimates for more complicated higher-dimensional tables. Under $M^{(7)}$, the $\hat{m}_{ijk}$'s are characterized by the model itself and the fitted margins $\hat{m}_{ij\cdot} = n_{ij\cdot}$, $\hat{m}_{i\cdot k} = n_{i\cdot k}$, and $\hat{m}_{\cdot jk} = n_{\cdot jk}$. The method is based on the fact that

$$1 = (n_{ij\cdot}/\hat{m}_{ij\cdot}) = (n_{i\cdot k}/\hat{m}_{i\cdot k}) = (n_{\cdot jk}/\hat{m}_{\cdot jk})$$

and, thus,

$$\begin{aligned} \hat{m}_{ijk} &= \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}}\hat{m}_{ijk}\,, \\ \hat{m}_{ijk} &= \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}}\hat{m}_{ijk}\,, \end{aligned}$$

and

$$\hat{m}_{ijk} = \frac{n_{\cdot jk}}{\hat{m}_{\cdot jk}}\hat{m}_{ijk}\,.$$

The iterative procedure begins with some initial guesses for the $\hat{m}_{ijk}$'s, say $\hat{m}_{ijk}^{[0]}$, and modifies the initial guess iteratively. Given estimates $\hat{m}_{ijk}^{[3t]}$, the modifications are

$$\begin{aligned} \hat{m}_{ijk}^{[3t+1]} &= \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{[3t]}}\hat{m}_{ijk}^{[3t]}\,, \\ \hat{m}_{ijk}^{[3t+2]} &= \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}^{[3t+1]}}\hat{m}_{ijk}^{[3t+1]}\,, \end{aligned}$$

and

$$\hat{m}_{ijk}^{[3(t+1)]} = \frac{n_{\cdot jk}}{\hat{m}_{\cdot jk}^{[3t+2]}}\hat{m}_{ijk}^{[3t+2]}\,.$$

The initial guesses can be any positive numbers that satisfy $M^{(7)}$. Typically, one takes

$$\hat{m}_{ijk}^{[0]} = 1 \qquad \text{for all} \qquad i, j, k\,.$$

The iterations continue until the estimates stop changing; i.e., for all i, j, k,

$$\hat{m}_{ijk}^{[3t]} \doteq \hat{m}_{ijk}^{[3t+1]} \doteq \hat{m}_{ijk}^{[3t+2]} \doteq \hat{m}_{ijk}^{[3(t+1)]} .$$

If convergence occurs to a set of values, say $\hat{m}_{ijk}$, then these must be the maximum likelihood estimates. To see this, we need to show two things: first, that $\hat{m}_{ij\cdot} = n_{ij\cdot}$, $\hat{m}_{i\cdot k} = n_{i\cdot k}$, and $\hat{m}_{\cdot jk} = n_{\cdot jk}$, and second, that the $\hat{m}_{ijk}$'s satisfy $M^{(7)}$.

Because of the nature of the iterative proportional fitting algorithm, at convergence we have

$$\hat{m}_{ijk} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}} \hat{m}_{ijk} ,$$

so

$$1 = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}}$$

and

$$\hat{m}_{ij\cdot} = n_{ij\cdot} .$$

Similarly, $\hat{m}_{i\cdot k} = n_{i\cdot k}$ and $\hat{m}_{\cdot jk} = n_{\cdot jk}$.

To see that $M^{(7)}$ is satisfied, we must show that

$$\hat{m}_{111}\hat{m}_{ij1}/\hat{m}_{i11}\hat{m}_{1j1} = \hat{m}_{11k}\hat{m}_{ijk}/\hat{m}_{i1k}\hat{m}_{1jk}$$

for any values of i, j, and k greater than one. The key point here is that if $\hat{m}_{ijk}^{[3t]}$ satisfies $M^{(7)}$, then the modifications also satisfy $M^{(7)}$. Thus, if the initial values satisfy $M^{(7)}$, the result of the iterative procedure also satisfies $M^{(7)}$.

Specifically, assume that the $\hat{m}_{ijk}^{[3t]}$'s satisfy $M^{(7)}$. We will show that the $\hat{m}_{ijk}^{[3t+1]}$'s satisfy $M^{(7)}$. Since

$$\hat{m}_{ijk}^{[3t+1]} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{[3t]}} \hat{m}_{ijk}^{[3t]},$$

we have

$$\frac{\hat{m}_{111}^{[3t+1]}\hat{m}_{ij1}^{[3t+1]}}{\hat{m}_{i11}^{[3t+1]}\hat{m}_{1j1}^{[3t+1]}} = \left[\frac{\left(n_{11\cdot}\Big/\hat{m}_{11\cdot}^{[3t]}\right)\left(n_{ij\cdot}\Big/\hat{m}_{ij\cdot}^{[3t]}\right)}{\left(n_{i1\cdot}\Big/\hat{m}_{i1\cdot}^{[3t]}\right)\left(n_{1j\cdot}\Big/\hat{m}_{1j\cdot}^{[3t]}\right)}\right] \frac{\hat{m}_{111}^{[3t]}\hat{m}_{ij1}^{[3t]}}{\hat{m}_{i11}^{[3t]}\hat{m}_{1j1}^{[3t]}}$$

and

$$\frac{\hat{m}_{11k}^{[3t+1]}\hat{m}_{ijk}^{[3t+1]}}{\hat{m}_{i1k}^{[3t+1]}\hat{m}_{1jk}^{[3t+1]}} = \left[\frac{\left(n_{11\cdot}\Big/\hat{m}_{11\cdot}^{[3t]}\right)\left(n_{ij\cdot}\Big/\hat{m}_{ij\cdot}^{[3t]}\right)}{\left(n_{i1\cdot}\Big/\hat{m}_{i1\cdot}^{[3t]}\right)\left(n_{1j\cdot}\Big/\hat{m}_{1j\cdot}^{[3t]}\right)}\right] \frac{\hat{m}_{11k}^{[3t]}\hat{m}_{ijk}^{[3t]}}{\hat{m}_{i1k}^{[3t]}\hat{m}_{1jk}^{[3t]}}.$$

Since $M^{(7)}$ is satisfied for the $\hat{m}_{ijk}^{[3t]}$'s and the multipliers do not depend on k, clearly $M^{(7)}$ is satisfied for the $\hat{m}_{ijk}^{[3t+1]}$'s. Similar arguments show that the $\hat{m}_{ijk}^{[3t+2]}$'s and $\hat{m}_{ijk}^{[3(t+1)]}$'s also satisfy $M^{(7)}$, cf. Exercise 3.8.11.

In fact, iterative proportional fitting can be used to fit any of the standard models. To fit, say $M^{(6)}$, the iterative procedure chooses $\hat{m}_{ijk}^{[0]}$ to satisfy $M^{(6)}$ and then modifies estimates $\hat{m}_{ijk}^{[2t]}$ using the marginal conditions $\hat{m}_{ij\cdot} = n_{ij\cdot}$ and $\hat{m}_{i\cdot k} = n_{i\cdot k}$. Specifically,

$$\hat{m}_{ijk}^{[2t+1]} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{[2t]}} \hat{m}_{ijk}^{[2t]}$$

and

$$\hat{m}_{ijk}^{[2(t+1)]} = \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}^{[2t+1]}} \hat{m}_{ijk}^{[2t+1]} .$$

The equations for modifying the $\hat{m}$'s are determined by the marginal conditions. It is easily checked that if the sequence converges, then the $\hat{m}$'s satisfy both the marginal conditions and $M^{(6)}$. In fact, since there are closed form estimates for the $\hat{m}$'s, it takes only one set of modifications to obtain the MLEs.

It was mentioned earlier that the initial guesses are typically taken as

$$\hat{m}_{ijk}^{[0]} = 1 \qquad \text{for all} \qquad ijk\,.$$

The reason is that this initial guess satisfies all of the standard models. Thus, to use iterative proportional fitting for any model, one need only specify the marginal conditions and the algorithm automatically provides the estimates.

Finally, since the method is based on multiplication, any initial guess of zero will always remain zero. In other words, if for any ijk, $\hat{m}_{ijk}^{[0]} = 0$, then $\hat{m}_{ijk}^{[t]} = 0$ for all t. Thus, if it is known ahead of time that $m_{ijk} = 0$, iterative proportional fitting can handle the situation by taking $\hat{m}_{ijk}^{[0]} = 0$. On the other hand, if it is not known that $m_{ijk} = 0$, then we must choose $\hat{m}_{ijk}^{[0]} > 0$ (cf. Section 8.1).

3.4 Log-Linear Models for Three-Dimensional Tables

In a three-dimensional table for continuous data, the basic model is a three-way analysis of variance model with all interactions, i.e., $y_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} + e_{ijk}$. For tables of counts, the same form model is used for the $\log(m_{ijk})$'s: A saturated model is written

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}.$$

We will be primarily interested in eight reduced versions of this model. Each of the eight reduced models corresponds to one of the eight independence – odds ratio models for three-dimensional tables. In particular, for tables of positive probabilities, an odds ratio model is true if and only if the corresponding log-linear model is valid.

The eight submodels are

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)}, \tag{0}$$

$$\begin{aligned}
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}, && (1)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)}, && (2)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}, && (3)
\end{aligned}$$

$$\begin{aligned}
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}, && (4)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}, && (5)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}, && (6)
\end{aligned}$$

and

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}\,. \tag{7}$$

For $i = 0, \ldots, 7$, the log-linear model (i) holds if and only if $M^{(i)}$ holds. One way to see this is to go through a series of arguments similar to those used in Section 2.4; however, the equivalence can most easily be seen by examining odds ratios. For example, model (7) holds if and only if the $u_{123(ijk)}$ terms can be dropped from the full model. As in standard analysis of variance, the three-factor interaction terms can be dropped if and only if any set of $(I-1)(J-1)(K-1)$ linearly independent three-factor interaction contrasts are all zero. In general, a three-factor interaction contrast is

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} q_{ijk} u_{123(ijk)}$$

or, equivalently,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} q_{ijk} \log(m_{ijk})$$

where $q_{ij\cdot} = q_{i\cdot k} = q_{\cdot jk} = 0$ for all i, j, k. The condition in $M^{(7)}$ is that

$$\frac{m_{111} m_{ij1}}{m_{i11} m_{1j1}} = \frac{m_{11k} m_{ijk}}{m_{i1k} m_{1jk}}$$

for all i, j, and k strictly greater than 1. Taking logs of both sides gives

$$\begin{aligned}
&\log(m_{111}) - \log(m_{i11}) - \log(m_{1j1}) + \log(m_{ij1})\\
&\qquad = \log(m_{11k}) - \log(m_{i1k}) - \log(m_{1jk}) + \log(m_{ijk}).
\end{aligned}$$

Rearranging terms, we see that $M^{(7)}$ is equivalent to

$$\log(m_{111}) - \log(m_{i11}) - \log(m_{1j1}) + \log(m_{ij1}) \\ - \log(m_{11k}) + \log(m_{i1k}) + \log(m_{1jk}) - \log(m_{ijk}) = 0$$

for each of the $(I-1)(J-1)(K-1)$ possible choices of i, j, and k greater than 1. All of these are three-factor interaction contrasts. Since these contrasts are linearly independent, $M^{(7)}$ holds if and only if the three-factor interaction terms can be dropped from the model, hence, if and only if model (7) holds.

Now consider $M^{(4)}$, that rows and columns are independent given layers. In terms of odds ratios, all odds ratios with any one factor fixed are equal, and all odds ratios with layers fixed equal one. To show that $M^{(4)}$ is equivalent to model (4), we need to show that $M^{(4)}$ is equivalent to having no three-factor interaction and no row-column interaction. As above, the fact that all odds ratios are equal is equivalent to having no three-factor interaction. We need to establish that all odds ratios with layers fixed equal one if and only if there is no row-column interaction. Under $M^{(4)}$, for all k, $m_{ijk}m_{i'j'k}/m_{ij'k}m_{i'jk} = 1$ or, taking logs,

$$\mu_{ijk} - \mu_{ij'k} - \mu_{i'jk} + \mu_{i'j'k} = 0$$

where $\mu_{ijk} \equiv \log m_{ijk}$. Recall that a contrast in the row-column interactions is an interaction contrast in the $\bar{\mu}_{ij\cdot}$'s. Averaging the odds ratio contrasts over k gives the row-column interaction contrast

$$\bar{\mu}_{ij\cdot} - \bar{\mu}_{ij'\cdot} - \bar{\mu}_{i'j\cdot} + \bar{\mu}_{i'j'\cdot} = 0\,.$$

If we take $i = 1$ and $j = 1$, we have $(I-1)(J-1)$ linearly independent contrasts in the row-column interaction equal to 0; the $u_{12(ij)}$ terms can be dropped from the full model. Conversely, if there is no three-factor interaction and no row-column interaction, then the contrasts $\mu_{ijk} - \mu_{ij'k} - \mu_{i'jk} + \mu_{i'j'k}$ are all equal and $\bar{\mu}_{ij\cdot} - \bar{\mu}_{ij'\cdot} - \bar{\mu}_{i'j\cdot} + \bar{\mu}_{ij\cdot} = 0$. Thus, all odds ratios are equal and those with layer fixed equal one.

Similarly, $M^{(1)}$ is true if and only if all odds ratios are equal and those with either columns or layers fixed equal one. All odds ratios equal is equivalent to no three-factor interaction; all odds ratios with layers fixed equal to one is equivalent to no row-column interaction (no $u_{12(ij)}$ terms); and all odds ratios with columns fixed equal to one is equivalent to no row-layer interaction (no $u_{13(ik)}$ terms).

Nearly all of models (0)-(7) are grossly overparametrized. For example, in model (1), the terms u, $u_{2(j)}$, and $u_{3(k)}$ are all totally redundant. The parameters $u_{1(i)}$ and $u_{23(jk)}$ are sufficient to explain everything. The u, $u_{2(j)}$, and $u_{3(k)}$ terms can take any values, yet by choosing the $u_{1(i)}$'s and $u_{23(jk)}$'s appropriately, model (1) holds. Rewriting the models in a less overparametrized fashion leads to a very convenient shorthand notation for the models

	MODEL	SHORTHAND
(0)	$\log(m_{ijk}) = u_{1(i)} + u_{2(j)} + u_{3(k)}$	[1][2][3]
(1)	$\log(m_{ijk}) = u_{1(i)} + u_{23(jk)}$	[1][23]
(2)	$\log(m_{ijk}) = u_{2(j)} + u_{13(ik)}$	[2][13]
(3)	$\log(m_{ijk}) = u_{3(k)} + u_{12(ij)}$	[3][12]
(4)	$\log(m_{ijk}) = u_{13(ik)} + u_{23(jk)}$	[13][23]
(5)	$\log(m_{ijk}) = u_{12(ij)} + u_{23(jk)}$	[12][23]
(6)	$\log(m_{ijk}) = u_{12(ij)} + u_{13(ik)}$	[12]][13]
(7)	$\log(m_{ijk}) = u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$	[12][13][23]

In addition, the unrestricted (saturated) model, $\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$ can be rewritten $\log(m_{ijk}) = u_{123(ijk)}$ and abbreviated as [123].

The shorthand can also be used to remember the conditional independence interpretations of the models. For example, [1][2][3] has everything in different brackets so everything is independent. In [1][23], the rows ([1]) are in a different bracket from columns and layers ([23]); thus rows are independent of columns and layers. In [13][23], rows 1 and columns 2 are in different brackets but layers 3 is in both brackets. Thus, given layers, rows and columns are independent. No such separation of factors works for [12][13][23], and there is no interpretation in terms of independence for the corresponding model.

The shorthand identifies both the model and the margins that must be fitted to obtain MLEs. Thus, the shorthand provides all the information necessary for fitting the models using iterative proportional fitting (or Newton-Raphson). For example, [1][23] requires that the margins $\hat{m}_{i..} = n_{i..}$ and $\hat{m}_{.jk} = n_{.jk}$ be fitted and the model [12][23] requires that $\hat{m}_{ij.} = n_{ij.}$ and $\hat{m}_{.jk} = n_{.jk}$ be fitted. As discussed in Chapter 10, under mild restrictions, any values $\hat{m}_{ijk}$ that satisfy the fitted margins and the log-linear model are the MLEs. In particular, for model (1) the unique MLEs are $\hat{m}^{(1)}_{ijk} = n_{i..}n_{.jk}/n_{...}$. These satisfy the marginal conditions and, because

$$\begin{aligned} \log\left(\hat{m}^{(1)}_{ijk}\right) &= \log(n_{i..}) + \log(n_{.jk}/n_{...}) \\ &= \hat{u}_{1(i)} + \hat{u}_{23(jk)}, \end{aligned}$$

they satisfy the log-linear model (1).

3.4.1 Estimation

Estimation of the expected cell counts m_{ijk} has already been considered. Given the $\hat{m}_{ijk}$'s, estimation of the model parameters can proceed in a manner similar to analysis of variance.

Consider a standard ANOVA model, say

$$y_{ijk} = \xi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + e_{ijk}$$

with $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, and the e_{ijk}'s independent $N(0, \sigma^2)$. A contrast in, for example, the $(\alpha\beta)$ interaction is determined by numbers q_{ij} $i = 1, \ldots, I$, $j = 1, \ldots, J$, that satisfy $q_{i\cdot} = q_{\cdot j} = 0$. The contrast is

$$\sum_{ij} q_{ij}(\alpha\beta)_{ij} \,.$$

The maximum likelihood estimate is

$$\sum_{ij} q_{ij}\bar{y}_{ij\cdot}$$

and

$$\text{Var}\left(\sum_{ij} q_{ij}\bar{y}_{ij\cdot}\right) = \frac{\sigma^2}{K}\sum_{ij} q_{ij}^2 \,.$$

The log-linear model

$$\log(m_{ijk}) = \xi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}$$

can be rewritten as

$$\mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \,,$$

where

$$\mu_{ijk} \equiv \log(m_{ijk}) \,.$$

Consider an interaction contrast

$$\sum_{ij} q_{ij} u_{12(ij)}$$

which is equivalent to

$$\sum_{ij} q_{ij}\bar{\mu}_{ij\cdot} \,.$$

The MLE of m_{ijk} is $\hat{m}_{ijk}$, so the MLE of μ_{ijk} is $\hat{\mu}_{ijk} = \log(\hat{m}_{ijk})$. Write

$$w_{ijk} \equiv \hat{\mu}_{ijk} = \log(\hat{m}_{ijk}) \,.$$

Then the estimated contrast is

$$\sum_{ij} q_{ij}\bar{w}_{ij\cdot} \,.$$

In general, the MLE of any function of the μ_{ijk}'s is just the same function applied to the $\hat{\mu}_{ijk}$'s. In particular, techniques from analysis of variance, when applied to the $\hat{\mu}_{ijk}$'s, give the estimates for contrasts in the corresponding log-linear models. In other words, whatever you would do to the

y_{ijk}'s in ANOVA to estimate a parameter, apply the same method to the $\hat{\mu}_{ijk}$'s to estimate the corresponding parameter in a log-linear model.

Unfortunately, computation of asymptotic variances is not straightforward. It requires the use of matrices and, even for contrasts, is similar in difficulty to finding the variance of a linear combination of regression coefficient estimates. Estimation is considered in detail in Section 10.2.

In the author's opinion, the most interesting aspects of estimation are those directly related to the m_{ijk}'s, odds, and odds ratios. Given the $\hat{m}_{ijk}$'s, estimates of odds and odds ratios are easy to obtain. Many examples of this have already been given. Again, the more difficult aspect of estimation is in obtaining asymptotic standard errors so that formal inferential procedures can be used.

3.4.2 Testing Models

In regression analysis it is well known that one can test a model against a larger model to see whether the smaller model is an inadequate explanation of the data. This technique is also used in analysis of variance but often it is not discussed explicitly because in balanced ANOVA it is possible to skirt the issue. For example, in a balanced ANOVA with two factors A and B and no interaction, the test for main effects in A does not depend on whether the main effects for B are included in the model. *The technique of testing models against larger models is fundamental in log-linear model analysis.* The sense in which one model is larger than another is illustrated below.

All of the tests discussed in Section 2 can be viewed as testing models against the saturated model. The test of $M^{(r)}$ was based on $\hat{m}_{ijk}^{(r)}$ and the n_{ijk}'s. The n_{ijk}'s are used because $n_{ijk} = \hat{m}_{ijk}$, where $\hat{m}_{ijk}$ is the unrestricted MLE of m_{ijk}. The unrestricted MLE of m_{ijk} is obtained by using a model that puts no restrictions on the m_{ijk}'s, namely, the saturated model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \,. \quad (8)$$

Again, this model is grossly overparametrized; an equivalent model is

$$\log(m_{ijk}) = u_{123(ijk)} \,.$$

A saturated model has at least one parameter for every cell in the table, so the model always fits the data perfectly.

More generally, one can test any model against a strictly larger model. For instance, model (1)

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

can be tested against model (4)

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)} \,,$$

because model (4) contains all of the terms in model (1) plus additional terms, the $u_{13(ik)}$'s. In other words, model (1) is a special case of model (4). The test is simply a test of whether the u_{13}'s are needed in model (4) (or equivalently a test of $M^{(1)}$ versus $M^{(4)}$).

To test [1][23] (model (1)) against [13][23] (model (4)), we use the $\hat{m}_{ijk}^{(1)}$'s, the $\hat{m}_{ijk}^{(4)}$'s, and the Pearson or likelihood ratio chi-squares. The Pearson chi-square is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(\hat{m}_{ijk}^{(4)} - \hat{m}_{ijk}^{(1)}\right)^2}{\hat{m}_{ijk}^{(1)}} .$$

The likelihood ratio chi-square is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{ijk}^{(4)} \log\left(\hat{m}_{ijk}^{(4)}/\hat{m}_{ijk}^{(1)}\right) .$$

Since this is a test of no row-layer interactions, the degrees of freedom are the degrees of freedom for row-layer interactions, i.e., $(I-1)(K-1)$, exactly as in analysis of variance.

Similarly, model (1): [1][23] can be tested against model (5): [12][23] because model (5) contains model (1). On the other hand, [1][23] cannot be tested against [12][13] because [1][23] contains $u_{23(jk)}$ terms, but [12][13] does not contain $u_{23(jk)}$ terms; thus, [12][13] is not strictly larger than [1][23]. In this case, we say that [1][23] and [12][13] are not comparable.

To perform tests, we need to be able to identify the degrees of freedom associated with each model. For standard analysis of variance type models, the degrees of freedom for a model are just the sum of the degrees of freedom for each term in the model. The degrees of freedom for terms are the same as in standard analysis of variance.

Term	Degrees of Freedom
u	1
u_1	$I-1$
u_2	$J-1$
u_3	$K-1$
u_{12}	$(I-1)(J-1)$
u_{13}	$(I-1)(K-1)$
u_{23}	$(J-1)(K-1)$
u_{123}	$(I-1)(J-1)(K-1)$

The degrees of freedom for testing [1][23] versus [13][23] are the degrees of freedom for [13][23] minus the degrees of freedom for [1][23]. Adding up the degrees of freedom for individual u terms, the degrees of freedom for [13][23] are $1+(I-1)+(J-1)+(K-1)+(I-1)(K-1)+(J-1)(K-1)$. The degrees

of freedom for [1][23] are $I + (I-1) + (J-1) + (K-1) + (J-1)(K-1)$. The degrees of freedom for the test are the difference in the model degrees of freedom, which is $(I-1)(K-1)$. As mentioned before, this is just the degrees of freedom for the terms that are in [13][23] but not in [1][23], i.e., the $u_{13(ik)}$'s.

In general, to test model (r) against model (s), where model (s) is strictly larger than model (r),

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(\hat{m}_{ijk}^{(s)} - \hat{m}_{ijk}^{(r)}\right)^2}{\hat{m}_{ijk}^{(r)}}$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{ijk}^{(s)} \log\left(\hat{m}_{ijk}^{(s)}/\hat{m}_{ijk}^{(r)}\right) . \tag{9}$$

These values can be compared to a chi-square distribution. The degrees of freedom for the chi-square is the difference in the degrees of freedom for models (s) and (r).

One of the advantages of using G^2 instead of X^2 is that it simplifies the process of testing models against each other. Any of the usual tests of models can be obtained easily from the eight tests in Section 2. In fact, this is the standard way of performing tests on models. The tests in Section 2 are all tests of a reduced log-linear model against the saturated model (8). (Note that the saturated model is strictly larger than any of the other eight models.)

EXAMPLE 3.4.1. In Example 3.2.2 on classroom behavior, it was remarked that there were relationships between various likelihood ratio test statistics. Specifically, the following results were given:

Model	G^2	*df*
$M^{(0)}$: [1][2][3]	16.42	7
$M^{(1)}$: [1][23]	5.56	5
$p_{\cdot jk} = p_{\cdot j\cdot}p_{\cdot\cdot k}$	10.86	2

The test statistics for all the models are for testing against the saturated model. (In the third case, it is the saturated model for the two-dimensional table.)

The model [1][23] includes u_{23} terms in addition to the terms in [1][2][3]. Because the model [1][23] is strictly larger than [1][2][3], a test for the adequacy of the smaller model can be performed. Rather than using equation (9) directly, the test of the adequacy of the smaller log-linear model can be obtained by subtraction from the saturated model test statistics given above. Specifically, for testing [1][2][3] versus [1][23],

$$G^2 = 16.42 - 5.56 = 10.86$$

with $7 - 5 = 2$ degrees of freedom. *Not only can this be viewed as a test of the log-linear models but also as a test of the independence models, i.e.,* $M^{(0)}$ *versus* $M^{(1)}$. Moreover, it is precisely the test given in Example 3.2.2 for

$$M: \quad p_{\cdot jk} = p_{\cdot j\cdot} p_{\cdot\cdot k}\,.$$

EXERCISE 3.2. Using the data of Example 3.2.2, compute the $\hat{m}_{ijk}^{(0)}$'s and use (9) to verify that $G^2 = 10.86$ for testing [1][2][3] versus [1][23].

We now develop these results in general. Consider testing each of models (r) and (s) against the saturated model. Recalling that for the saturated model $\hat{m}_{ijk} = n_{ijk}$, we get likelihood ratio test statistics of

$$G^2(\text{r vs. 8}) = 2\sum_{ijk} n_{ijk} \log\left(n_{ijk}/\hat{m}_{ijk}^{(r)}\right)$$

and

$$G^2(\text{s vs. 8}) = 2\sum_{ijk} n_{ijk} \log\left(n_{ijk}/\hat{m}_{ijk}^{(s)}\right)\,.$$

As shown at the end of Section 10.1, maximum likelihood estimates for log-linear models satisfy

$$\begin{aligned} G^2(\text{r vs. s}) &= 2\sum_{ijk} \hat{m}_{ijk}^{(s)} \log\left(\hat{m}_{ijk}^{(s)}/\hat{m}_{ijk}^{(r)}\right) \\ &= 2\sum_{ijk} n_{ijk} \log\left(\hat{m}_{ijk}^{(s)}/\hat{m}_{ijk}^{(r)}\right)\,. \end{aligned}$$

Given this property, it is a simple matter to check that

$$G^2(\text{r vs. s}) = G^2(\text{r vs. 8}) - G^2(\text{s vs. 8})\,.$$

Moreover, the degrees of freedom for the tests satisfy

$$df(\text{r vs. s}) = df(\text{r vs. 8}) - df(\text{s vs. 8})\,. \tag{10}$$

To see (10), note that (a) the degrees of freedom for the saturated model (8) are IJK, (b) $df(\text{r vs. s}) = df(\text{model (s)}) - df(\text{model (r)})$, (c) $df(\text{r vs. 8}) = IJK - df(\text{model (r)})$, and (d) $df(\text{s vs. 8}) = IJK - df(\text{model (s)})$. Substitution into (10) gives the correct result. *The methods of obtaining* G^2*(r vs. s) and df (r vs. s) from* G^2*'s and df 's for testing against saturated models are basic to log-linear model practice.* Typically, computer programs only provide G^2's and *df*'s for testing against saturated models, so reduced model tests must be constructed using this method.

In fact, this approach to testing (r) versus (s) using G^2's for the saturated model is not restricted to G^2's for the saturated model. It can be used with

any model (t) that is larger than both (r) and (s). The use of the saturated model is a convenience because it is strictly larger than *all* other models.

EXAMPLE 3.4.2. Once again, we consider the data on personality (1), cholesterol (2), and diastolic blood pressure (3) of Example 3.2.3. In the table below are given the degrees of freedom, values of X^2 and G^2, and the *P value* associated with G^2 for testing all eight of the standard models against the saturated model.

Model	*df*	X^2	G^2	*P*
[12][13][23]	1	0.617	0.613	.434
[12][13]	2	2.188	2.062	.358
[12][23]	2	2.985	2.980	.224
[13][23]	2	4.566	4.563	.100
[1][23]	3	7.102	7.101	.067
[2][13]	3	6.189	6.184	.102
[3][12]	3	4.543	4.601	.207
[1][2][3]	4	8.730	8.723	.067

Using a criterion of $\alpha = .05$, all of these models fit the data; however, consider testing [1][2][3] against [12][13]. The test statistic is

$$G^2 = 8.723 - 2.062 = 6.661$$

with

$$df = 4 - 2 = 2\,.$$

Because

$$\chi^2(.95, 2) = 5.99\,,$$

the model [12][13] fits significantly better than [1][2][3]. In other words, the model with cholesterol level and diastolic blood pressure level independent given personality type fits significantly better than the model of complete independence. The reader can also verify that the models [3][12] and [12][13][23] also fit significantly better than [1][2][3]. (The model [12][23] is almost significantly better than [1][2][3].) We are left with a sequence of hierarchical models [3][12], [12][13], and [12][13][23] that all fit better than complete independence. Testing [3][12] against [12][13] gives

$$\begin{aligned} G^2 &= 4.601 - 2.062 = 2.539\,, \\ df &= 3 - 2 = 1\,, \\ \chi^2(.95, 1) &= 3.84\,, \end{aligned}$$

so there seems to be no reason to take the larger model. Similarly, testing [3][12] against [12][13][23] gives

$$\begin{aligned} G^2 &= 4.601 - 0.613 = 3.988 \\ df &= 3 - 1 = 2 \\ \chi^2(.95, 2) &= 5.99\,, \end{aligned}$$

so again, the model [3][12] seems adequate. The model that posits blood pressure being independent of personality type and cholesterol is the smallest model that adequately fits the data. (It is interesting to note that based on Akaike's information criterion, cf. Section 6, model [12][13] is the best model.)

To complete the analysis, we need to examine the nature of the relationship between personality and cholesterol. This was done in Example 3.2.3. That analysis remains valid.

3.5 Product-Multinomial and Other Sampling Plans

In this section, we consider the implications of product-multinomial sampling and give a brief discussion of the effect of complex sampling plans involving stratified sampling and cluster sampling. In addition, the use of conditional distributions as a basis for statistical inference is mentioned.

Recall the data from Example 2.1.1 on opinions about legalized abortion.

	Support	Do Not Support	Total
Female	309	191	500
Male	319	281	600
Total	628	472	1100

This is product-multinomial sampling. A sample of 500 females was taken. An independent sample of 600 males was also taken. The results were combined into a 2×2 table. We consider two extensions of these data to illustrate product-multinomial sampling in three-dimensional tables.

EXAMPLE 3.5.1. Suppose that each population is further classified according to political affiliation. We might then get the table

		Opinion (k)		
Sex (i)	Party (j)	Support	Do Not Support	Total
Female	Republican	79	40	119
	Democrat	132	71	203
	Independent	98	80	178
	Total	309	191	500
Male	Republican	65	94	159
	Democrat	141	95	236
	Independent	113	92	205
	Total	319	281	600

The totals for females and males are fixed. We know the female total is 500, so we must expect the female total to be 500. This means that

$$m_{1\cdot\cdot} = n_{1\cdot\cdot} = 500\,.$$

Similarly, for male totals,

$$m_{2\cdot\cdot} = n_{2\cdot\cdot} = 600\,.$$

More briefly, we write simply

$$m_{i\cdot\cdot} = n_{i\cdot\cdot}\,,$$

$i = 1, 2$. Any model that we fit must accommodate these facts. In other words, our estimates must satisfy the constraints

$$\hat{m}_{i\cdot\cdot} = n_{i\cdot\cdot}\,, \tag{1}$$

$i = 1, 2$. Fortunately, the estimates for all of the ANOVA type models that we have discussed satisfy this condition. Any model that includes the $u_{1(i)}$ terms (or their equivalents) will satisfy (1).

We now consider a slightly more complex sampling scheme.

EXAMPLE 3.5.2. Consider a three-factor table based on sex, socioeconomic status, and opinion about legalized abortion. Socioeconomic status has two categories: low and not low. The table of counts is

		Opinion (k)		
Sex (i)	Status (j)	Support	Do Not Support	Total
	Low	171	79	250
Female	Not Low	138	112	250
	Total	309	191	500
	Low	152	148	300
Male	Not Low	167	133	300
	Total	319	281	600

In this table, four independent samples have been incorporated into the table. The samples are (1) a sample of 250 low-status females, (2) a sample of 250 females not of low status, (3) a sample of 300 low status males, and (4) a sample of 300 males not of low status. The sampling design has fixed the sex-status marginal totals, so the expected sex-status totals equal the observed totals, i.e., $m_{ij\cdot} = n_{ij\cdot}$. Any model that estimates expected cell counts must also incorporate the condition that

$$\hat{m}_{ij\cdot} = n_{ij\cdot}$$

for all i and j. In particular, any model that includes the $u_{12(ij)}$ terms will have these margins fixed. If we restrict attention to models that include the $u_{12(ij)}$ terms, we do not have to concern ourselves further with the product-multinomial nature of the sampling design.

The restriction that $u_{12(ij)}$ terms must be in the model reduces the possible number of models. The possible models are listed below.

Possible Models with $m_{ij.}$ Fixed by the Sampling Design
[123]
[12][13][23]
[12][13]
[12][23]
[12][3]

Finally, these ideas extend easily to higher-dimensional tables. Suppose we have a four-dimensional table with indices h, i, j, k. If the sampling design fixes the margins

$$m_{h\cdot jk} = n_{h\cdot jk}\,,$$

then we restrict attention to log-linear models that include the $u_{134(hjk)}$ terms. If the sampling design fixes the margins

$$m_{\cdot i\cdot k} = n_{\cdot i\cdot k}\,,$$

then we consider only models with $u_{24(ik)}$ terms. Note that if the model includes, say, the $u_{234(ijk)}$ terms, then the $u_{24(ik)}$ terms are implicitly in the model. With the $u_{234(ijk)}$ terms in the model, the $u_{24(ik)}$ terms are redundant and it is irrelevant whether the $u_{24(ik)}$'s are explicitly stated as part of the model or not.

Examples 3.5.1 and 3.5.2 illustrate the two primary sampling schemes for response factors that were discussed in Section 2.3. In both examples, Opinion can be viewed as a response factor. In Example 3.5.2, both Sex and Status are explanatory factors. For every combination of the levels of the explanatory factors, there is an independent multinomial sample. The categories for each multinomial are the levels of the response factor Opinion. This is the first of the sampling schemes discussed in Section 2.3. In Example 3.5.1, only the two levels of the explanatory factor Sex are used to define the independent multinomial populations. The categories of the multinomials are defined by all combinations of the levels of the response factor Opinion and the levels of the explanatory factor Party. This is the generalized sampling scheme discussed in Section 2.3. If Opinion is regarded as a response, it is not unusual to condition on all of the explanatory factors, i.e., Sex and Party, in the analysis. Thus, the data may be treated as if they were product-multinomial with an independent multinomial for each

combination of the explanatory factors. These issues are discussed again in Chapter 4.

3.5.1 Other Sampling Models

As mentioned in Section 1.5, the other commonly used sampling scheme for log-linear models is Poisson sampling. In Poisson sampling, an independent Poisson random variable is observed for each cell in the table. It is easily seen that Poisson sampling leads to the same methods of analysis that are used for multinomial sampling, cf. Chapter 12.

Although Poisson, product-multinomial, and multinomial sampling are the only sampling models considered in this book, it should not be concluded that these are the only sampling schemes used to generate and analyze categorical data. The hypergeometric distribution is often taken as the appropriate sampling model. Also, most large sample surveys are not conducted so as to generate multinomial or Poisson data. Of course, these alternative sampling models may require substantial changes in the statistical analysis.

Hypergeometric sampling arises quite naturally in discrete data problems. For example, 2×2 tables in which both the row totals and the column totals are fixed can be generated by hypergeometric sampling. Hypergeometric sampling is easily generalized to $I \times J$ tables. The hypergeometric distribution is also appropriate if one conditions on the row and column totals. The reason for conditioning on the row and column totals is that they are sufficient statistics under the model of independence.

Tests for model adequacy based on the conditional distribution, given the sufficient statistics of the model, play a major role in Statistics generally and have a particularly important history in the analysis of categorical data. *Fisher's exact conditional test* for 2×2 tables, cf. Exercise 2.7.5 or Plackett (1981), may be the most famous single methodology in categorical data analysis. A key reason for using *conditional tests* is that they are appropriate even for small samples. The main problem with conditional tests is that for log-linear models, they are generally difficult to compute. McCullagh (1986) and Hirji, Mehta, and Patel (1987) use alternative approaches to examine conditional tests. McCullagh concentrates on the use of *asymptotic conditional distributions* with the idea that conditional asymptotics are more appropriate for small and moderate sample sizes than unconditional asymptotics. Hirji et al. *enumerate* all of the tables that have the same values for the sufficient statistics. This work also applies to logistic regression. While enumeration is a very demanding computational task, with modern algorithms and computing equipment it has become a realistic alternative to the methods discussed here. Haberman (1974a, p. 14-33) details conditional inference for categorical data. Balmer (1988), Bedrick and Hill (1990), Mehta and Patel (1980, 1983), Mehta, Patel, and Gray (1985), Mehta, Patel, and Tsiatis (1984), and Pagano and Taylor-Halvorson

(1981) all address issues of conditional inference and table enumeration. Recent reviews of these methods have been given by Agresti (1992) and Mehta (1994). The computer programs StatXact and/or LogXact perform the necessary computations. Davison (1988) uses *saddlepoint methods* to approximate conditional distributions. In some cases, random samples of the tables can be used rather than enumerating all of the tables. Kreiner (1987) discusses model selection using conditional tests and, specifically, the use of random samples from the conditional distribution. The generation of such random samples is based on the work of Agresti, Wackerly and Boyett (1979), Boyett (1979), and Patefield (1981). Berkson (1978) and Kempthorne (1979) give alternative views of Fisher's exact test. For a general discussion of conditional tests see Lehmann (1986).

Large sample surveys typically involve the use of *stratification* and *cluster sampling*, cf. Kish (1965). Multinomial sampling corresponds to simple random sampling with replacement. Product-multinomial sampling involves independent samples on a number of subpopulations. This is just stratified sampling. As we have seen, if strata are included as a factor in the table, then stratified sampling causes no problems in a log-linear model analysis of the data. The difficulty with stratified sampling is that often the individual strata are not of interest in the analysis. The desired conclusions are for the population as a whole and must be arrived at by weighting results from the separate strata. In sampling theory, the point of selecting strata is to reduce the variability of the overall results. If this is accomplished and if data from a stratified sample are analyzed as though they are multinomial, the variability is overestimated and results appear to be less significant than they actually are.

The incorporation of cluster sampling is fundamentally more difficult to deal with. The whole point of cluster sampling is that the observations within a cluster are not independent. Typically, they display a positive correlation. All of our standard sampling plans assume independence between individual observations, so the standard plans are clearly inappropriate for cluster sampling. Inferences based on independence will underestimate the variability of data with a positive correlation. Thus, analyzing cluster sampling data as if they were multinomial will typically overstate the significance of results.

In a *complex survey* involving both stratification and cluster sampling, the tendencies to understate significance and to overstate significance will, to some extent, offset each other. While this is a positive sign for the standard multinomial analysis, it by no means ensures that any particular set of survey data can be analyzed accurately when the complex sampling structure is ignored. In all likelihood, one or the other tendency to misstate the variability will dominate. Any serious analysis of complex survey data must involve an evaluation of the effect of the survey design on the analysis. Some of the key references on the analysis of complex survey data are Koch, Freeman, and Freeman (1975), Fienberg (1979), Brier (1980), Holt, Scott,

and Ewings (1980), Rao and Scott (1981, 1984, 1987), Bedrick (1983), Binder et al. (1984), Gross (1984), Fay (1985), and Thomas and Rao (1987). The collection of papers edited by Skinner, Holt, and Smith (1989) provides a useful summary of methods for analyzing data from complex surveys, including categorical data.

3.6 Model Selection Criteria

In analysis of variance and regression, the three measures most commonly used to evaluate the fit of models are R^2, Adjusted R^2, and Mallows' C_p. Each of these measures has natural analogues in log-linear models. R^2 measures how much of the total variation is being explained by the model. R^2 has the property that models must explain as much or more of the variation than their submodels (models in which some terms have been deleted). Adjusted R^2 modifies the definition of R^2 so that larger models are penalized for being larger. Mallows' C_p statistic is related to Akaike's information criterion. We will apply Akaike's information criterion to model selection for log-linear and logistic regression models and discuss the relation of Akaike's criterion to Mallows' C_p.

Discussions of model selection criteria in general settings are given by Akaike (1973) and Schwarz (1978). A good review and comparison is presented in Clayton, Geisser, and Jennings (1986).

3.6.1 R^2

In standard regression analysis, R^2 is defined as

$$R^2 = \frac{\text{SSReg}}{\text{SSTot} - \text{C}}$$

where SSReg is the sum of squares for regression and SSTot−C is the sum of squares total corrected for the grand mean. In fact, SSTot−C is just the error sum of squares for the model that includes only an intercept. If we denote SSE(X) as the error sum of squares for an arbitrary model called X (e.g., with design matrix X) and SSE(X_0) as the error sum of squares for a model with only an intercept, then

$$R^2 = \frac{\text{SSE}(X_0) - \text{SSE}(X)}{\text{SSE}(X_0)} \, .$$

SSE(X_0) is the total variation and SSE(X_0) − SSE(X) is the variability explained by the model X. The ratio of these two, R^2, is the proportion of the total variation explained by the model.

In general, there is no reason that SSE(X_0) has to be the sum of squares for a model with just an intercept. In general, SSE(X_0) could be the error

sum of squares for the smallest interesting model. In regression analysis, the smallest interesting model is almost always the model with only an intercept. In log-linear models, the smallest interesting model may well be the model of complete independence. (Recall from Exercise 2.4 that the independence model for a two-way table is also the intercept-only model for logistic regression.)

In log-linear models, G^2 plays a role similar to that of SSE in regression. If X_0 indicates the smallest interesting model and X indicates the log-linear model of interest, we define

$$R^2 = \frac{G^2(X_0) - G^2(X)}{G^2(X_0)}$$

where $G^2(X)$ and $G^2(X_0)$ are the likelihood ratio test statistics for testing models X and X_0 against the saturated model.

If the X_0 model is the smallest interesting model, then $G^2(X_0)$ is a measure of the total variability in the data. (It tests X_0 against a model that fits the data perfectly.) It follows that $G^2(X_0) - G^2(X)$ measures the variability explained by the X model. R^2 is the proportion of the total variability explained by the X model. Alternative definitions of R^2 are available.

As in standard regression analysis, R^2 cannot be used to compare models that have different numbers of degrees of freedom. In regression analysis, this is caused by the fact that larger models have larger R^2's. Exactly the same phenomenon occurs with log-linear models. In fact, R^2 for the saturated model will always equal one because G^2 for the saturated model is zero.

3.6.2 Adjusted R^2

Having defined R^2 for log-linear models, the same adjustment for model size used in standard regression analysis can be used for log-linear models. The *adjusted* R^2 is

$$\text{Adj. } R^2 = 1 - \frac{q - r_0}{q - r}[1 - R^2]$$

where q is the number of cells in the table and r and r_0 are the degrees of freedom for the models X and X_0. Note that there are $q - r$ degrees of freedom for testing X against the saturated model and $q - r_0$ degrees of freedom for testing X_0.

A little algebra shows that

$$\text{Adj. } R^2 = 1 - \frac{G^2(X)/(q - r)}{G^2(X_0)/(q - r_0)} .$$

A large value of Adj. R^2 indicates that the model X fits well. The largest value of Adj. R^2 will occur for the model X with the smallest value of

$G^2(X)/(q-r)$. Just as in regression analysis, the Adj. R^2 criterion suggests the inclusion of many (probably too many) explanatory terms.

3.6.3 Akaike's Information Criterion

We now consider Akaike's information criterion as a method for selecting log-linear models. After describing Akaike's method, we demonstrate its close relationship to standard regression model selection based on Mallow's C_p statistic.

Akaike (1973) proposed a criterion of the information contained in a statistical model. He advocated choosing the model that maximizes this information. For log-linear models, maximizing *Akaike's information criterion* (AIC) amounts to choosing the model X that *minimizes*

$$A_X = G^2(X) - [q - 2r] \,, \qquad \text{(log-linear)}$$

where $G^2(X)$ is the likelihood ratio test statistic for testing the X model against the saturated model, r is the number of degrees of freedom for the X model, and there are q degrees of freedom for the saturated model, i.e., q cells in the table.

Given a list of models to be compared along with their G^2 statistics and the degrees of freedom for the tests, a slight modification of A_X is easier to compute by hand.

$$\begin{aligned} A_X - q &= G^2(X) - 2[q - r] \\ &= G^2(X) - 2 \text{ (test degrees of freedom)} \,. \end{aligned}$$

Because q does not depend on the model X, minimizing $A_X - q$ is equivalent to minimizing A_X. Note that for the saturated model, $A - q = 0$.

Before continuing our discussion of the AIC, we give an example of the use of A, R^2, and Adj. R^2.

EXAMPLE 3.6.1. For the personality (1), cholesterol (2), blood pressure (3) data of Examples 3.2.1 and 3.2.3, testing models against the saturated model gives

Model	*df*	G^2	$A - q$	R^2	Adj. R^2
[12][13][23]	1	0.613	−1.387	.885	.719
[12][13]	2	2.062	−1.938	.764	.527
[12][23]	2	2.980	−1.020	.658	.318
[13][23]	2	4.563	0.563	.477	−.046
[1][23]	3	7.101	1.101	.186	−.085
[2][13]	3	6.184	0.184	.291	.055
[3][12]	3	4.602	−1.398	.472	.297
[1][2][3]	4	8.723	0.723	0	0

In Example 3.4.2, we established that there were three eligible models: [3][12], [12][13], and [12][13][23] and that model [12][23] was almost eligible. The AIC criterion $A - q$ easily picks out all four of these models, with [12][13] the best of them. The adjusted R^2 criterion also identifies these four models, but the values seem rather strange. For example, [12][23], which is not significantly better than [1][2][3], has a higher Adj. R^2 than [3][12], which is significantly better than [1][2][3]. The R^2 values seem like reasonable measures. The author's inclination is to use the AIC, cf. Clayton, Geisser, and Jennings (1986).

With only three factors, it is easy to look at all possible models. Model selection criteria become more important when dealing with tables having more factors.

Relation to Mallows' C_p

The approach to using Akaike's information criterion outlined above is quite general. If we are considering a collection of models indexed by ξ, we can denote individual models as M_ξ. For any ξ, let p_ξ be the dimension of the parameter space of the model. This is the number of independent parameters in the model. For models with linear structure, this is the degrees of freedom for the model (rank of the design matrix) plus the number of any independent nonlinear parameters. Log-linear models do not involve any nonlinear parameters. Standard regression and ANOVA involve one nonlinear parameter, the variance σ^2.

Suppose that there exists a most general model M with s independent parameters. In other words, any model M_ξ is just a special case of M. For log-linear models, M is the saturated model and s is the number of cells in the table. For selecting variables in standard regression analysis, M is the full model that includes an intercept plus all $s - 1$ available variables. Finally, let $\Lambda(\xi)$ be the likelihood ratio test statistic for testing M_ξ against the larger model M. Maximizing Akaike's information criterion is equivalent to choosing ξ to minimize

$$A_\xi = \Lambda(\xi) - (s - 2p_\xi) \, .$$

Consider applying this to the problem of variable selection in regression analysis. Let $\mathrm{SSE}(F)$ be the sum of squares for error of the full model and $\mathrm{SSE}(X)$ be the error sum of squares for a reduced model with design matrix X and $\mathrm{rank}(X) = p$. Let s be the degrees of freedom for the full model. *If we assume that the variance σ^2 is known*, then

$$A_X = \frac{\mathrm{SSE}(X) - \mathrm{SSE}(F)}{\sigma^2} - (s - 2p) \, . \qquad \text{(regression)}$$

Because σ^2 will not really be known, an ad hoc procedure would be to estimate σ^2 with $\hat{\sigma}^2 = \mathrm{SSE}(F)/(n - s)$, the mean squared error for the full

model where n is the regression sample size. An estimate of A_X is

$$\begin{aligned}\hat{A}_X &= \frac{\text{SSE}(X)}{\hat{\sigma}^2} - \frac{\text{SSE}(F)}{\hat{\sigma}^2} - (s-2p) \\ &= \frac{\text{SSE}(X)}{\hat{\sigma}^2} - (n-s) - (s-2p) \\ &= \frac{\text{SSE}(X)}{\hat{\sigma}^2} - (n-2p) \\ &= C_p\end{aligned}$$

where C_p is Mallows' well-known criterion for selecting regression models, cf. Christensen (1996b, Section 14.1).

3.7 Higher-Dimensional Tables

Log-linear models can easily be extended to tables with more than three factors. All of the basic principles from three-dimensional tables continue to apply. However the models, as well as independence and odds ratio relationships, become more complex. Independence relationships for high-dimensional tables are discussed in Chapter 5. With more factors, there are many more models to consider. Systematic methods of model selection are discussed in Chapter 6. In this section, we just illustrate some examples.

EXAMPLE 3.7.1. A study was performed on mice to examine the relationship between two drugs and muscle tension. For each mouse, a muscle was identified and its tension was measured. A randomly chosen drug was given to the mouse and the muscle tension was measured again. The muscle was then tested to identify which type of muscle it was. The weight of the muscle was also measured. Factors and levels are tabulated below.

Factor	Abbreviation	Levels
Change in Muscle Tension	T	High, Low
Weight of Muscle	W	High, Low
Muscle	M	Type 1, Type 2
Drug	D	Drug 1, Drug 2

The sampling is product-multinomial with the total count for each muscle type fixed. The data are

Tension(h)	Weight(i)	Muscle(j)	Drug(k) Drug 1	Drug 2
	High	Type 1	3	21
		Type 2	23	11
High				
	Low	Type 1	22	32
		Type 2	4	12
	High	Type 1	3	10
		Type 2	41	21
Low				
	Low	Type 1	45	23
		Type 2	6	22

For illustration, we fit three log-linear models to this four-factor table: the model of all main effects

$$\log(m_{hijk}) = \gamma + \tau_h + \omega_i + \mu_j + \delta_k, \tag{1}$$

the model of all two-factor interactions

$$\begin{aligned}\log(m_{hijk}) &= \gamma + \tau_h + \omega_i + \mu_j + \delta_k + (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} \\ &\quad + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk},\end{aligned} \tag{2}$$

and the model of all three-factor interactions

$$\begin{aligned}\log(m_{hijk}) &= \gamma + \tau_h + \omega_i + \mu_j + \delta_k + (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} \\ &\quad + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk} \\ &\quad + (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik} + (\tau\mu\delta)_{hjk} + (\omega\mu\delta)_{ijk}.\end{aligned} \tag{3}$$

Getting rid of some of the redundant parameters, these can be rewritten as

$$\log(m_{hijk}) = \tau_h + \omega_i + \mu_j + \delta_k, \tag{1}$$

$$\log(m_{hijk}) = (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk}, \tag{2}$$

and

$$\log(m_{hijk}) = (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik} + (\tau\mu\delta)_{hjk} + (\omega\mu\delta)_{ijk} \tag{3}$$

respectively, leading to the shorthand notations

$$[\mathrm{T}][\mathrm{W}][\mathrm{M}][\mathrm{D}], \tag{1}$$

$$[\mathrm{TW}][\mathrm{TM}][\mathrm{WM}][\mathrm{TD}][\mathrm{WD}][\mathrm{MD}], \tag{2}$$

$$[\mathrm{TWM}][\mathrm{TWD}][\mathrm{TMD}][\mathrm{WMD}]. \tag{3}$$

As discussed in Section 4, the shorthand provides all the information necessary for fitting the model (other than the actual cell counts).

The test statistics for testing these models against the saturated model are given below. Clearly, the only model that fits the data is the model of all three-factor interactions.

Model	*df*	G^2	*P*
[TWM][TWD][TMD][WMD]	1	0.11	.74
[TW][TM][WM][TD][WD][MD]	5	47.67	.00
[T][W][M][D]	11	127.4	.00

As before, we can also test any model against reduced models. For example the test of [TWM][TWD][TMD][WMD] versus the reduced model [TW][TM][WM][TD][WD][MD] has $G^2 = 47.67 - 0.11 = 47.56$ on $df = 5 - 1 = 4$.

The data of Example 3.7.1 and the following data will be used to illustrate techniques in subsequent chapters.

EXAMPLE 3.7.2. Consider a data set in which there are four factors defining a $2 \times 2 \times 3 \times 6$ table. The factors are

Factor	Abbreviation	Levels
Race	R	White, Nonwhite
Sex	S	Male, Female
Opinion	O	Yes = Supports Legalized Abortion No = Opposed to Legalized Abortion Und = Undecided
Age	A	18-25, 26-35, 36-45, 46-55, 56-65, 66+ years

The sex and opinion factors are reminiscent of Example 2.1.1, but the data are distinct. The data are given in Table 3.1. See also Exercise 3.8.10.

3.7.1 Computer Commands

The muscle tension data are listed in file 'tension.dat' as counts for each cell with indices for tension, weight, muscle type, and drug, respectively. The file is as given below.

```
 3 1 1 1 1
21 1 1 1 2
23 1 1 2 1
11 1 1 2 2
22 1 2 1 1
32 1 2 1 2
 4 1 2 2 1
```

TABLE 3.1. Abortion Opinion Data

Race	Sex	Opinion	Age					
			18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Und	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Und	1	4	9	9	10	16
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Und	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Und	1	2	1	1	1	1

```
12 1 2 2 2
 3 2 1 1 1
10 2 1 1 2
41 2 1 2 1
21 2 1 2 2
45 2 2 1 1
23 2 2 1 2
 6 2 2 2 1
22 2 2 2 2
```

We can fit the log-linear model [WMD][TWM][TWD][TMD] using SAS PROC GENMOD.

```
options ps=60 ls=72 nodate;
data tension;
   infile 'tension.dat';
   input n T W M D;
proc genmod data=tension;
   class T W M D;
   model n = W*M*T T*W*M T*W*D T*M*D / link=log
                                       dist=poisson;
run;
```

The main differences between these commands and those given in Subsection 2.6.1 for logistic regression are that now "link=log" and

"dist=poisson". These change GENMOD from fitting logistic regression to fitting log-linear models. To fit other specific models such as [TM][WM][MD] or [T][WM][D], the model statement uses T*M W*M M*D or T W*M D, respectively. The "class" command used above specifies that a variable is not acting like a predictor variable in regression but rather that it gives indices for specifying the levels of an analysis of variance type factor.

Similarly, we can use GENMOD to fit the abortion data. The data file 'abort.dat' has five columns, the first four are indices for race, sex, age, and opinion. The last column has the counts for each cell. The SAS commands for fitting the model [RSO][RSA][ROA][SOA] are

```
options ps=60 ls=72 nodate;
data abort;
   infile 'abort.dat';
   input R S A O N;
proc genmod data=abort;
   class R S A O;
   model N = R*S*O R*S*A R*O*A S*O*A / link=log
                                       dist=poisson;
run;
```

GLIM uses commands that are similar to GENMOD. Interactions are specified with a period rather than an asterisk. GLIM begins by specifying the number of cells in the table, i.e., the "units."

```
$units 72$
$data r s a o n$
$factor r 2 s 2 a 6 o 3$
$dinput 6$
$yvar n$
$error poisson$
$fit r.s.o + r.s.a + r.o.a + s.o.a$
$display e$
$stop$
```

The "factor" command specifies that a variable is not acting like a predictor variable in regression but rather that it gives indices for specifying the levels of an analysis of variance type factor. For GLIM, the user needs to specify the number of levels for each factor. After the 'dinput 6' command, the DOS version of GLIM prompts the user for the name of the data file.

GENMOD and GLIM use the Newton-Raphson algorithm. BMDP-4F uses iterative proportional fitting. For the model [RSO][RSA][ROA][SOA], the BMDP-4F commands are as follows:

```
/ INPUT     FILE = 'C:\LOGLIN\ABORT.DAT'.
          FORMAT = FREE.
```

```
         VARIABLES = 5.
/ VARIABLE  NAMES = R, S, A, O, N.
/ TABLE     INDICES = R, S, A, O.
            COUNT = N.
/ STAT      ALL.
/ FIT       MODEL = RSO, RSA, ROA, SOA.
/ PRINT     LINE = 79.
/ END
```

BMDP-4F is the most powerful program I am aware of for fitting analysis of variance type log-linear models. In addition to GENMOD, SAS has a procedure called CATMOD. CATMOD will not be discussed. (I'll leave the reasons to your imagination.)

3.8 Exercises

EXERCISE 3.8.1. Complete an analysis similar to that of Example 3.4.2 for the classroom behavior data of Example 3.0.1.

EXERCISE 3.8.2. Complete an analysis similar to that of Example 3.4.2 for the auto accident data of Example 3.2.4.

EXERCISE 3.8.3. Radelet (1981) gives data on the relationship between race and the imposition of the death penalty. The data are given in Table 3.2. Analyze the data.

TABLE 3.2. Race and the Death Penalty

Defendant's Race	Victim's Race	Death Penalty Yes	No
Black	Black	6	97
	White	11	52
White	Black	0	9
	White	19	132

EXERCISE 3.8.4. The data on graduate admissions at Berkeley given in Exercise 2.6.1 was actually collapsed over the six largest departments within the university. The possibility exists that the data may display Simpson's paradox. The full data are given in Table 3.3. Analyze the three-dimensional table and comment on Simpson's paradox relative to these data.

EXERCISE 3.8.5. Discuss Simpson's paradox in terms of the following probability inequalities.

$$\Pr(A|B \text{ and } C) < \Pr(A|\text{ not } B \text{ and } C),$$

TABLE 3.3. Graduate Admissions at Berkeley

	Male		Female	
Dept.	Admitted	Rejected	Admitted	Rejected
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

$$\Pr(A|B \text{ and not } C) < \Pr(A|\text{ not } B \text{ and not } C),$$

and

$$\Pr(A|B) > \Pr(A|\text{ not } B).$$

EXERCISE 3.8.6. Reevaluate your analysis of the data discussed in Exercise 2.6.3 in light of Simpson's paradox. Are there other factors that need to be accounted for in a correct analysis of these data?

EXERCISE 3.8.7. For the data of Example 3.2.4, do the first step of the iterative proportional fitting algorithm for $\hat{m}_{ijk}^{(7)}$ using a hand calculator. Use starting values of $\hat{m}_{ijk}^{[0]} = 1$. Compare the results after one step to the fully iterated estimates.

EXERCISE 3.8.8. Consider the model $\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$.

(a) Show that the maximum likelihood estimate of $u_{3(1)} - u_{3(2)}$ is $\log(n_{..1}) - \log(n_{..2})$.

(b) Show that maximum likelihood estimation gives

$$\log\left[\frac{\hat{u}_{12(11)}\hat{u}_{12(22)}}{\hat{u}_{12(12)}\hat{u}_{12(21)}}\right] = \log(n_{11.}) - \log(n_{12.}) - \log(n_{21.}) + \log(n_{22.}).$$

EXERCISE 3.8.9. *The Mantel-Haenszel Statistic.*
In biological and medical applications, it is not uncommon to be confronted with a series of 2×2 tables that examine the same effect under different conditions. If there are K such tables, the data can be combined to form a $2 \times 2 \times K$ table. Because each 2×2 table examines the same effect, it is often assumed that the odds ratio for the effect is constant over tables. This is equivalent to assuming the no three-factor interaction model. To test for the existence of the effect, one tests whether the common log odds ratio is zero while adjusting for the various circumstances under which data were collected. In terms of log-linear models, this is a one degree of freedom test

of conditional independence given the layer k. Prior to the development of log-linear model theory, Mantel and Haenszel (1959) proposed a statistic for testing this hypothesis. The statistic, apart from a continuity correction factor, is

$$\frac{\left[\sum_k (n_{11k} - \hat{m}_{11k})\right]^2}{\sum_k \left[\hat{m}_{11k}\hat{m}_{22k}\right] \Big/ \left[n_{..k} - 1\right]},$$

where the $\hat{m}$'s are obtained from the conditional independence model. This statistic has an asymptotic $\chi^2(1)$ distribution under the conditional independence model.

The Berkeley graduate admission data of Exercise 3.8.4 and Table 3.3 is a set of six 2×2 tables. In each table we are interested in the effect of sex on admission; the six departments constitute various conditions under which this effect is being investigated.

a) Give a justification for whether or not use of the Mantel-Haenszel statistic is appropriate for these data.

b) If appropriate, use both G^2 and the Mantel-Haenszel statistic to test whether there is an effect of sex on admission.

c) Show that the denominator of the Mantel-Haenszel statistic can be written as $\sum_k \left[\hat{m}_{12k}\hat{m}_{21k}\right] \Big/ \left[n_{..k} - 1\right]$.

Exercise 3.8.10. Using the data of Table 3.1 fit the all main effects model, the all two-factor effects model, and the all three-factor effects model. Perform all of the tests possible among these three models. Discuss your results.

Exercise 3.8.11. With regard to Section 3, show that the $\hat{m}_{ijk}^{[3t+2]}$'s and $\hat{m}_{ijk}^{[3(t+1)]}$'s also satisfy $M^{(7)}$.

Exercise 3.8.12. As can be seen from the iterative proportional fitting algorithm, the $\hat{m}$'s for the model of no three-factor interaction depend only on the 3 two-dimensional marginal tables. Discuss how this fact can be used to develop a more complete analysis for the Gilby data of Exercise 2.7.3. What assumptions must be made and what techniques should be used? What problems will Standard VIII cause?

Project 3.8.13. Write a computer program to fit the model of no three-factor interaction to the Gilby data of Exercise 2.6.3. This can be done using iterative proportional fitting. Assuming that this model fits, do any submodels fit adequately?

4

Logistic Regression, Logit Models, and Logistic Discrimination

In logistic regression, there is a (binary) response of interest, and predictor variables are used to model the probability of that response. More generally, in a table of counts, primary interest is frequently centered on one factor that constitutes a response (dependent) variable. The other factors in the table are only of interest for their ability to help explain the response variable. Special kinds of models have been developed to handle these situations. In particular, rather than modeling log expected cell counts or log probabilities (as in log-linear models), when there is a response variable, various log *odds* related to the response variable are modeled.

The special case in which the response variable has only two categories is of particular interest and lends itself to an especially nice treatment. This is because, with only two categories, there is essentially only one way to define the odds. If p_1 is the probability in the first category and p_2 is the probability in the second category, then the odds of getting category one are p_1/p_2. The odds of getting category two are p_2/p_1. The important point is that *either* of these numbers, together with the fact that $p_1 + p_2 = 1$, completely determine both p_1 and p_2. So with two categories, the two choices for the odds lead to the same results. (In the last section of this chapter, we look at ratios p_1/p_2 but *without* $p_1 + p_2 = 1$; this causes complications.)

With a two-category response variable, we will examine models for $\log(p_1/p_2)$. When these models are regression type models, they are called *logistic regression models*. When these models are ANOVA type models, they are often referred to as *logit models*. The two terms "logit" and "logistic regression," as applied to models, are essentially two names for the

same idea. Technically, the terms logit and logistic are names for transformations. The logit transformation takes a number p between 0 and 1 and transforms it to $\log[p/(1-p)]$. The logistic transformation takes a number x on the real line and transforms it to $e^x/(1+e^x)$. Note that the logit transformation and the logistic transformation are inverses of each other. In other words, the logistic transformation applied to $\log[p/(1-p)]$ gives p and the logit transformation applied to $e^x/(1+e^x)$ gives x. Doing an analysis of data requires both of these transformations. It is largely a matter of personal preference as to which name is associated with the model.

The situation when there are more than two categories in the response variable is considerably more complicated because it is not clear which sets of odds to model. Several choices have been suggested; some of these are discussed in Section 6. As will be discussed in this chapter and shown in Chapter 11, the log odds models turn out to be equivalent to a log-linear model. *It is important to remember that log odds models are for use when relationships between the nonresponse factors (explanatory variables) are not of interest.* It is implicit in the definition of a logit model that no structure between the explanatory variables is taken into account. It *is* possible to use models for log odds that incorporate explanatory factor structure, but such models are not what are generally known as logistic regression or logit models.

Goodman (1973) proposes a multistep modeling procedure for response factors. His procedure involves collapsing on the response factor and fitting a log-linear model to the marginal table of the explanatory factors. This is followed by fitting a logit model for the response factor. Taken together, these give the probability of any cell in the table as the product of the marginal probability of its explanatory categories and the conditional probability of the cell given its explanatory categories. These recursive response models may or may not be log-linear models. Asmussen and Edwards (1983) give conditions for the equivalence of log-linear models and these multistep response models. Fienberg (1980, Chapter 7) gives a good brief discussion of response models and their limitations. More recently, graphical response models have been discussed by Asmussen and Edwards (1983), Edwards and Kreiner (1983), Kiiveri, Speed, and Carlin (1984), Kiiveri and Speed (1982), and Wermuth and Lauritzen (1983). Graphical models are discussed in the next chapter. Holland (1986) discusses statistics and causal inference, as do Glymour et al. (1987). The latter authors seem to have a substantially different perspective than that presented here. Goodman's basic procedure can be applied with more than one response factor and with responses involving more than two categories. In this chapter, attention is concentrated on models for one response factor that condition on all the explanatory factors. If more than one response factor is present, one simple approach just restricts the fitted models to log-linear models that condition on the explanatory factors.

Section 1 examines regression models for two category responses. Sections 2, 3, and 4 discuss measuring the fit of models, logistic regression diagnostics, and variable selection, respectively. Analysis of variance type models are examined in Section 5 for responses with two categories and in Section 6 for responses with more than two categories. Section 7 examines the analysis of retrospective studies via logistic discrimination; the distinction between retrospective and prospective studies is discussed in the next subsection.

Retrospective Versus Prospective Studies

It is important to distinguish between prospective and retrospective studies. It is important because aspects of the material in this chapter do not apply to retrospective studies. For our purposes, the distinction is based on the nature of the sampling scheme.

If the sampling is independent Poisson or if the categories of the response factor are in the same multinomial for every combination of the explanatory factors, the study is *prospective*. This is probably the most common way of thinking about data collection. For example, a prospective study of heart attack victims might take 250 people and examine each to determine whether they have had a heart attack and their levels of various explanatory factors. The explanatory factors may be such things as age, blood cholesterol, and blood pressure. In the prospective study, each combination of explanatory factors can be used to determine a population and an individual randomly falls into a response category. So, typically, there are many populations, and often each individual in the study is sampled from a different population. Prospective studies are (or can be thought of) as product-multinomials in which *the multinomial categories are the categories of the response factor*.

Obviously, in a random sample of 250 people from the general population, very few would have had heart attacks. An alternative sampling method is often used in the study of such rare events as heart attacks. One might sample 100 people who are known to have had heart attacks and 150 people who have not had heart attacks. Again, each individual is characterized by their levels of the explanatory factors. There are only two populations here: the heart attack victims and the subjects without heart attacks. The categories of the multinomials for the two populations are the different categories of explanatory factors. The analysis of such data involves describing the characteristics of the two groups in terms of the explanatory factors. The key fact in this second example is that *the response factor categories define different multinomial populations and the multinomial categories are the different combinations of the explanatory factors*. Generally, if, for all combinations of the explanatory factors, the various categories of the response factor occur in different multinomials, then the study is *retrospective*. In medical research, retrospective data col-

lection corresponds to a *case-control* study. Clearly, for rare events, this procedure has advantages over selecting a simple random sample. In a multinomial sample, so few of the rare events will occur as to give little power for determining their likelihood.

In the hypothetical retrospective study of heart attacks, let the index i denote a set of explanatory characteristics; let p_{1i} be the probability of that set for the heart attack population and p_{2i} be the probability for the population who have not had heart attacks. A parameter of interest is p_{1i}/p_{2i}, the ratio of the probabilities. (Note that, in general, $p_{1i} + p_{2i} \neq 1$, so these are *not* odds.) The parameter addresses the question of whether the explanatory characteristics i are more or less likely among heart attack victims than among subjects who have not had heart attacks. Unfortunately, this does not address the issue of cause and effect. It simply describes characteristics of the two populations. As will be seen in Section 7, inferences about $\log(\hat{p}_{1i}/\hat{p}_{2i})$ will be complicated by the fact that the probabilities apply to different multinomials. The asymptotic covariance of $\log(\hat{p}_{1i}/\hat{p}_{2i})$ does not simplify like it would if the probabilities were from the same multinomial. Moreover, one does not typically have the simplification that p_{1i}/p_{2i} is equal to m_{1i}/m_{2i}; thus, inferences about $\log(p_{1i}/p_{2i})$ cannot be made directly by examining $\log(\hat{m}_{1i}/\hat{m}_{2i})$.

Prospective studies do not directly address the issue of cause and effect either, but they come closer than retrospective studies. As discussed earlier, both multinomial sampling and some forms of product-multinomial sampling generate prospective studies. An example of a product-multinomial prospective study is to independently sample people for each set of explanatory characteristics and see how many have had heart attacks. In medical research, the data given by this sampling scheme are called *cohort* data, and *cross-sectional* data are used to indicate the results of simple multinomial sampling.

For cohort data, take p_{1i} to be the probability of a heart attack in the population defined by the ith set of explanatory variables. Obviously, $p_{2i} = 1 - p_{1i}$ is the probability of no heart attack in that population. The ratio p_{1i}/p_{2i} is the odds of having a heart attack for that population. (Recall that describing p_{1i}/p_{2i} as an odds is not appropriate in a retrospective study.) In a cohort study, one can argue that the ith population is a cause and that the ratio p_{1i}/p_{2i} is an effect. Unfortunately, populations usually involve more than the explanatory factors used to define them. Aspects of the ith population other than the values of the explanatory factors may be the true cause for p_{1i}/p_{2i}. We hope that random sampling within the population of people minimizes these effects.

For cross-sectional data, one can use the mental device of conditioning on the number of people who fall in each explanatory category to get an independent multinomial sample for each set of explanatory characteristics. Thus, we can treat cross-sectional prospective data as if they were cohort prospective data. In fact, any prospective study can be thought of

as product-multinomial sampling with an independent sample for each set of explanatory characteristics. This results from the fact that a prospective study is defined to be one in which all of the categories of the response factor are contained in the same multinomial.

Except for the last section, we restrict attention in this chapter to prospective studies. A convenience of dealing with prospective studies is that log odds defined for the response factor are the same using either probabilities or expected cell counts, i.e., $\log(p_{1i}/p_{2i}) = \log(m_{1i}/m_{2i})$.

4.1 Multiple Logistic Regression

This section is devoted to regression models for the log odds of a two-category response variable. The difference between this section and Section 2.6 is that here we consider the use of multiple predictor variables. The discussion will be centered around an example.

EXAMPLE 4.1.1. *Chapman Data.*
Dixon and Massey (1983) present data on 200 men taken from the Los Angeles Heart Study conducted under the supervision of John M. Chapman, UCLA. The data consist of seven variables:

Abbreviation	Variable	Units
Ag	Age:	in years
S	Systolic Blood Pressure:	millimeters of mercury
D	Diastolic Blood Pressure:	millimeters of mercury
Ch	Cholesterol:	milligrams per DL
H	Height:	inches
W	Weight:	pounds
CNT	Coronary incident:	1 if an incident had occurred in the previous ten years; 0 otherwise

Of the 200 cases, 26 had coronary incidents. The data are available electronically from STATLIB as well as through my web homepage:

`http://stat.unm.edu/~fletcher`

Additional information is given in the Preface.

As discussed in Section 2.6, such data can be viewed as a 200×2 contingency table in which the columns indicate presence or absence of a coronary incident and the rows indicate the 200 combinations of the explanatory variables Ag, S, D, Ch, H, and W associated with the men in the study. Each row is considered an independent binomial involving one trial. (If more than one person has the same combination of explanatory variables, it is

irrelevant whether they are treated as binomials with one trial or grouped together yielding a table with less than 200 rows.) The table has 200 counts and 400 cells, so the data are very sparse. As discussed in Section 2.6, when testing a logistic regression model against the saturated log-linear model (i.e., testing the logistic model for lack of fit), the asymptotic χ^2 approximation is notoriously bad. The test statistics are reasonable things to look at, but formal χ^2 tests are generally inappropriate because of the sparse data. Somewhat surprisingly, asymptotic χ^2 approximations do work for testing one logistic regression model (a full model) against another logistic regression (a reduced model).

Let p_i be the probability of a coronary incident for the ith man. We begin with the logistic regression model

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + \beta_3 D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i, \quad (1)$$

$i = 1, \ldots, 200$. As discussed in Section 2.6 and Chapter 11, this is equivalent to a log-linear model for a two-way table in which the predictor variables are used to model the interaction, cf. Exercise 4.8.15. The model can be fitted using methods for log-linear models or the methods can be specialized for fitting logistic regression models, cf. Subsection 4.4.2 for SAS, BMDP, and GLIM commands. The actual methods for fitting logistic models will be examined in later chapters. The maximum likelihood fit of this model is given below.

Variable	Estimate	Std. Error	z
Intercept	−4.5173	7.451	−0.61
Ag	0.04590	0.02344	1.96
S	0.00686	0.02013	0.34
D	−0.00694	0.03821	−0.18
Ch	0.00631	0.00362	1.74
H	−0.07400	0.1058	−0.70
W	0.02014	0.00984	2.05

$$G^2 = 134.9, \qquad df = 193$$

The formula for G^2 is as in Section 2.6. The df is the number of cases, 200, minus the number of fitted parameters, 7. Based on the z values, none of the variables really stand out. There are suggestions of age, cholesterol, and weight effects. The G^2 would look good except that, as discussed earlier, there is no basis for comparing it to a standard.

Prediction follows much the same form as in Section 2.6,

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \hat{\beta}_0 + \hat{\beta}_1 Ag_i + \hat{\beta}_2 S_i + \hat{\beta}_3 D_i + \hat{\beta}_4 Ch_i + \hat{\beta}_5 H_i + \hat{\beta}_6 W_i\,.$$

For a 60-year-old man with blood pressure of 140 over 90, a cholesterol reading of 200, who is 69 inches tall and weighs 200 pounds, the estimated

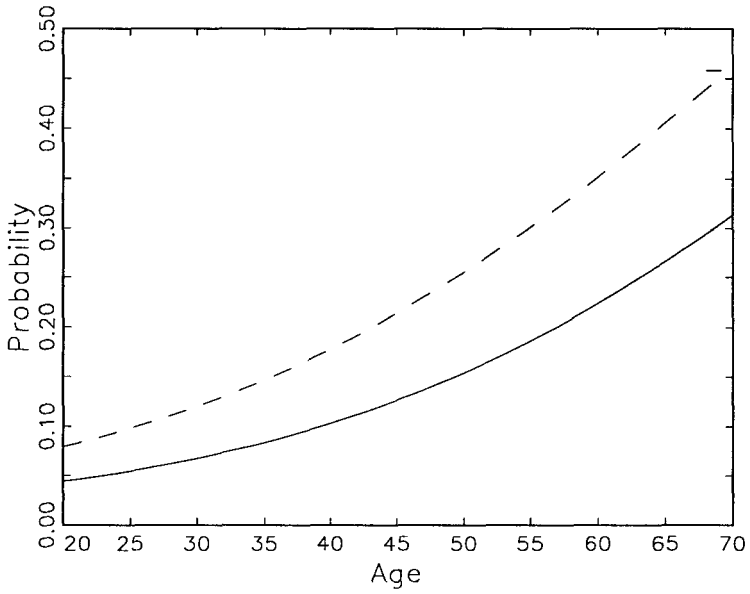


FIGURE 4.1. Coronary incident probabilities as a function of age. Solid curve–$Ch = 200$; dashed curve–$Ch = 300$.

log odds of a coronary incident are

$$\log[\hat{p}/(1-\hat{p})] = -4.5173 + .04590(60) + .00686(140) - .00694(90) + .00631(200) - 0.07400(69) + 0.02014(200) = -1.2435.$$

The probability of a coronary incident is estimated as

$$\hat{p} = \frac{e^{-1.2435}}{1 + e^{-1.2435}} = .224\,.$$

Figure 4.1 gives plots of the estimated probability of a coronary incident as a function of age for people with $S = 140$, $D = 90$, $H = 69$, $W = 200$, and either $Ch = 200$ (solid line) or $Ch = 300$ (dashed line).

4.1.1 Informal Model Selection

We now consider fitting some reduced models. Simple linear logistic regressions were fitted for each of the variables with high z values, i.e., Ag, Ch, and W. Regressions with variables that seem naturally paired were also fitted, i.e., S,D and H,W. Listed below are the models, df, G^2, $A - q$, and A^*. The first two of these are the degrees of freedom and the likelihood ratio test statistic for testing against the saturated model. No P *values* are given because the asymptotic χ^2 approximation does not hold. Also given are two analogues of Mallow's C_p statistic, $A - q$ and A^*. $A - q$ was

discussed in detail in Section 3.6. A^* is a modification of $A - q$ for logistic regression. $A - q \equiv G^2 - 2(df)$ and is the Akaike information criterion less the number of cells (200×2) in the table. A^* is a version of the Akaike information criterion defined for comparing submodels of model (1) to the full model. It is defined by

$$A^* = (G^2 - 134.9) - (7 - 2p)\,.$$

Here, 134.9 is G^2 for the full model (1), 7 comes from the degrees of freedom for the full model (6 explanatory variables plus an intercept), and p comes from the degrees of freedom for the submodel ($p = 1 +$ number of explanatory variables). The information in $A - q$ and A^* is identical: $A^* = 258.1 + (A - q)$. (The value 258.1 = number of cells $- G^2$[full model] $- p$[full model] $= 400 - 134.9 - 7$.) A^* is listed because it is a little easier to look at and takes values similar to C_p.

Model Variables	*df*	G^2	$A - q$	A^*
Ag,S,D,Ch,H,W	193	134.9	−251.1	7
Ag	198	142.7	−253.3	4.8
W	198	150.1	−245.9	12.2
H,W	197	146.8	−247.2	10.9
Ch	198	146.9	−249.1	9.0
S,D	197	147.9	−246.1	12.0
Intercept	199	154.6	−243.4	14.7

Of the models listed,

$$\log[p_i/(1 - p_i)] = \gamma_0 + \gamma_1 Ag_i \tag{2}$$

is the only model that is better than the full model based on the information criterion; i.e., A^* is 4.8 for this model, less than the 7 for model (1).

Asymptotically valid tests of submodels against model (1) are available. These are performed in the usual way; i.e., the differences in degrees of freedom and G^2's give the appropriate values for the tests. For example, the test of model (2) versus model (1) has $G^2 = 142.7 - 134.9 = 7.8$ with $df = 198 - 193 = 5$. Other tests are given below.

Tests against Model (1)

Model	*df*	G^2
Ag	5	7.8
W	5	15.2**
H,W	4	11.9*
Ch	5	12.0*
S,D	4	13.0*
Intercept	6	19.7**

All of the test statistics are significant at the .05 level, except for that associated with model (2). This indicates that none of the models other than (2) is an adequate substitute for the full model (1). In the table above, one asterisk indicates significance at the .05 level and two asterisks denotes significance at the .01 level.

Our next step is to investigate models that include Ag and some other variables. If we can find one or two variables that account for most of the value $G^2 = 7.8$, we may have an improvement over model (2). If it takes three or more variables to explain the 7.8, model (2) will continue to be the best-looking model. [Note that $\chi^2(.95, 3) = 7.81$, so a model with three more variables than model (2) and the same fit as model (1) would still not demonstrate a significant lack of fit in model (2).]

Below are fits for all models that involve Ag and either one or two other explanatory variables.

Model Variables	*df*	G^2	A^*
Ag,S,D,Ch,H,W	193	134.9	7.0
Ag,S,D	196	141.4	7.5
Ag,S,Ch	196	139.3	5.4
Ag,S,H	196	141.9	8.0
Ag,S,W	196	138.4	4.5
Ag,D,Ch	196	139.0	5.1
Ag,D,H	196	141.4	7.5
Ag,D,W	196	138.5	4.6
Ag,Ch,H	196	139.9	6.0
Ag,Ch,W	196	135.5	1.6
Ag,H,W	196	138.1	4.2
Ag,S	197	141.9	6.0
Ag,D	197	141.4	5.5
Ag,Ch	197	139.9	4.0
Ag,H	197	142.7	6.8
Ag,W	197	138.8	2.9
Ag	198	142.7	4.8

Based on the A^* values, two models stand out:

$$\log[p_i/(1-p_i)] = \gamma_0 + \gamma_1 Ag_i + \gamma_2 W_i \tag{3}$$

with $A^* = 2.9$ and

$$\log[p_i/(1-p_i)] = \eta_0 + \eta_1 Ag_i + \eta_2 W_i + \eta_3 Ch_i \tag{4}$$

with $A^* = 1.6$.

The estimated parameters and standard errors for model (3) are

Variable	Parameter	Estimate	SE
Intercept	γ_0	−7.513	1.706
Ag	γ_1	0.06358	0.01963
W	γ_2	0.01600	0.00794

For model (4), these are

Variable	Parameter	Estimate	SE
Intercept	η_0	−9.255	2.061
Ag	η_1	0.05300	0.02074
W	η_2	0.01754	0.003575
Ch	η_3	0.006517	0.008243

The coefficients for Ag and W are quite stable in the two models. The coefficients of Ag, W, and Ch are all positive, so that a small increase in age, weight, or cholesterol is associated with a small increase in the odds of having a coronary incident. Note that we are establishing association, not causation.

As in regular regression, interpreting regression coefficients can be very tricky. The fact that the regression coefficients are all positive conforms with the conventional wisdom that high values for any of these factors increases one's chance of heart trouble. However, as in standard regression analysis, correlations between predictor variables can make interpretations of individual regression coefficients almost impossible.

It is interesting to note that from fitting model (1), the estimated regression coefficient for D, diastolic blood pressure, is negative. A naive interpretation would be that as diastolic blood pressure goes up, the probability of a coronary incident goes down. (If the log odds go down, the probability goes down.) This is contrary to common opinion about how these things work. Actually, this is really just an example of the fallacy of trying to interpret regression coefficients. The regression coefficients have been determined so that the fitted model explains these particular data as well as possible. As mentioned, correlations between the predictor variables can have a huge effect on the estimated regression coefficients. The sample correlation between S and D is .802, so estimated regression coefficients for these variables are unreliable. Moreover, it is not even enough just to check pairwise correlations between variables; any large partial correlations will also adversely affect interpretations. Fortunately, such correlations should not normally have an adverse affect on the predictive ability of the model; they only adversely affect attempts to interpret regression coefficients. In Chapter 13, we will see that the regression coefficients also depend on the precise form of the logit model. Other methods for modeling the probabilities that are both reasonable and very similar to logistic regression can have very different regression coefficients while giving very similar probabilities. Finally, in this particular example, another excuse for the D coefficient $\hat{\beta}_3$

being negative is that from the z value, β_3 is not significantly different from zero.

The estimated blood pressure coefficients from model (1) also suggest an interesting hypothesis. (The hypothesis would be much more interesting if the individual coefficients were significant, but we wish to demonstrate a modeling technique.) The coefficient for D is $-.00694$, which is approximately the negative of the coefficient for S, .00686. This suggests that perhaps the difference $S - D$ would be just as valuable a predictor as the individual predictors S and D. We can evaluate this by fitting

$$\log[p_i/(1-p_i)] = \gamma_0 + \gamma_1 Ag_i + \gamma_2(S_i - D_i) + \gamma_3 Ch_i + \gamma_4 H_i + \gamma_5 W_i,$$

which gives $G^2 = 134.9$ on $df = 194$. This model is a special case of model (1), so a test of it against model (1) has

$$G^2 = 134.9 - 134.9 = 0$$

with $df = 194 - 193 = 1$. The G^2 is essentially zero, so the data are consistent with the reduced model. Of course, this reduced model was suggested by the fitted full model, so any formal test would be biased — but then one does not accept null hypotheses anyway, and the whole point of choosing this reduced model was that it seemed likely to give a G^2 close to that of model (1). We note that the new variable, $S - D$, is still not significant; it only has a z value of $.006834/.01877 = .36$.

Another way to view the procedure of the previous paragraph would be as a test of $H_0 : \beta_3 = -\beta_2$ in model (1). If we incorporate this hypothesis into model (1), we get

$$\begin{aligned}
&\log[p_i/(1-p_i)] \\
&\quad = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + (-\beta_2) D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i \\
&\quad = \beta_0 + \beta_1 Ag_i + \beta_2(S_i - D_i) + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i
\end{aligned}$$

as displayed above. (Whether we call the parameters β's or γ's is irrelevant.)

We learned earlier that, relative to model (1), either model (3) or (4) does an adequate job of explaining the data. This conclusion was based on looking at A^* values, but would also be obtained by doing formal tests of models. Thus, we know that age and weight are important variables in explaining coronary incidents. Moreover, cholesterol may also be an important variable. However, we have not explained most of the variability in coronary incidents.

Consider a measure analogous to R^2. The smallest interesting logistic regression model is $\log[p_i/(1-p_i)] = \gamma_0$. As seen earlier, this has $G^2 = 154.6$ on 199 degrees of freedom. The percent of variability explained by model (3) is

$$R^2(Ag, W) = \frac{154.6 - 138.8}{154.6} = .10\,,$$

which seems pretty pathetic. In fact, the R^2 from the full model is not much better

$$R^2(Ag, S, D, Ch, H, W) = \frac{154.6 - 134.9}{154.6} = .13\,.$$

We are a very long way from fitting the data as well as the saturated model fits. In fact, if we fit a 28-parameter model including all variables, all squares of variables, and all two-factor cross-product terms, G^2 is 108.8, so R^2 is still only .30.

Granted, we have not explained most of the variation in the data, but it was probably not reasonable to think that we could. In standard regressions, a perfect fitting model can have a low R^2. This happens when there is substantial pure error in the model. The same thing happens in logistic regression. That fact will be illustrated in the next section.

4.2 Measuring Model Fit

In regression and ANOVA, R^2 is large when the pure error $\text{Var}(y_i) = \sigma^2$ is small. When σ^2 is unknown, there is always the hope that it will be small (if we can find the correct model). In logistic regression, $\text{Var}(y_i) = N_i p_i(1-p_i)$. There isn't any hope of making this universally small. You can only make it truly small by looking at uninteresting data — those with p_i near 0 or 1. Cases with a realistic chance of going either way make for large variability.

We begin our examination of how well models fit by looking at the likelihood ratio and Pearson test statistics as applied to logistic regression. As before, we have I independent binomials, each consisting of one trial. Thus, we have a logistic regression with I cases and the dependent variable y is either 0 or 1. Equivalently, we have an $I \times 2$ table with counts n_{ij}, where $n_{i1} + n_{i2} = 1$. Let the variable y correspond to counts in the first column of the table so that $y_i = n_{i1}$ and $1 - y_i = n_{i2}$. If a logistic regression model for $\log[p_i/(1-p_i)]$ is fitted, we obtain estimates $\hat{p}_i$ of the p_i's. For the corresponding log-linear model, $\hat{p}_i = \hat{m}_{i1}$ and $(1 - \hat{p}_i) = \hat{m}_{i2}$. The likelihood ratio and Pearson test statistics against the saturated model were given in Section 2.6.

EXAMPLE 4.2.1. Suppose that in a true and very accurate model, there are 30 observations (i values) each with $p_i \doteq .1 \doteq \hat{p}_i$ and 30 with $p_i \doteq .9 \doteq \hat{p}_i$. From the first 30, we could then expect to get about three observations with $y_i = 1$. From (2.6.6), each of these observations has a crude standardized residual of about

$$\frac{(1 - .1)}{\sqrt{.1(1 - .1)}} = 3\,.$$

The other 27 observations will have residuals of

$$\frac{(0-.1)}{\sqrt{.1(1-.1)}} = -.333\,.$$

Similarly, from the second 30 observations, three would be about $(0-.9)/\sqrt{.9(1-.9)} = -3$ and 27 would be about $(1-.9)/\sqrt{.9(1-.9)} = .333$. It is disturbing that this perfect model with a perfect fit has what usually would be considered large residuals.

Using the formula for X^2 from Section 2.6, the Pearson statistic will be about

$$X^2 \doteq 6(3^2) + 54(.333^2) = 60\,,$$

which is the number of cases. No matter how accurate the model is, the Pearson statistic for these observations will still be about 60. It will never get small. A similar phenomenon holds for the likelihood ratio test statistic. Under the same circumstances as discussed above,

$$G^2 \doteq 2[6\log(1/.1) + 54\log(1/(1-.1))] = 33.32\,.$$

(Asymptotic theory does not hold for these tests, so there is no reason to expect X^2 and G^2 to be about the same.) Note that a constant model for these data would have $\hat{p}_i \doteq .5$ and

$$G^2 = 2(60)\log(1/.5) = 83.18,$$

so for this essentially perfect model which is fit perfectly,

$$R^2 = \frac{83.18 - 33.32}{83.18} = .599;$$

not very high under the circumstances. In fact, since the probabilities in this example were chosen to be quite extreme (near 0 and 1), the observations have unusually low variability, which should actually inflate R^2. The moral is simply that one should not expect to see the very high R^2's that one sometimes gets in standard regression.

No matter how accurate the fitted model, the test statistics will not become arbitrarily small nor will R^2 approach 1. The likelihood ratio statistic G^2 will become small only as the intrinsic variability of the true model decreases, i.e., as all probabilities approach 0 or 1. The Pearson statistic evaluated at the true model will remain near I, the number of cases.

There are two morals to all of this. First, R^2 type measures can be used to measure relative goodness of fit but may be misleading if used to measure absolute goodness of fit. Models with low R^2's can fit great. Models with high R^2's can exhibit lack of fit. Second, residuals from logistic regression cannot be used without special consideration given to the 0-1 nature of the data (cf. Jennings, 1986).

EXAMPLE 4.2.2. Using model (4.1.4), one finds that of the 200 cases in the Chapman data, 26 cases had a crude standardized residual in excess of .97. The 26 cases were precisely the 26 cases that had coronary incidents. A method for identifying unusual cases that indicates that every case with a coronary event is unusual leaves something to be desired.

4.2.1 Checking Lack of Fit

Methods of checking for lack of fit in logistic regression have been discussed by Tsiatis (1980), Landwehr, Pregibon, and Shoemaker (1984), and Fienberg and Gong (1984). Their approaches are based on clustering near replicates of the regression variables so that something akin to pure error can be identified. We present here a method of evaluation inspired by standard residual analysis. Rather than clustering near replicates, it clusters cases with similar $\hat{p}_i$'s.

A standard method for identifying lack of fit in regression analysis is to plot the residuals against the predicted values. This plot should form a structureless horizontal band about zero (cf. Christensen, 1996b, Section 13.4). An equivalent plot would be the observations versus the predicted values. This plot should form a structureless band about the line with slope 1 and intercept 0. For logistic regression, a plot of observations versus predicted values should show predicted values near 0 having most observations equal to 0, and predicted values near 1 having most observations equal to 1; predicted values near .5 should have about equal numbers of observations that are 0s and 1s, etc. Such a plot could be difficult to interpret visually, so let's get cruder.

Break the predicted values into, say, 10 intervals: [0,.1), [.1,.2), [.2,.3), ..., [.9,1]. For each interval, find the number of cases that have $\hat{p}$ in the interval and the number of those cases that have $y = 1$. The midpoint of the interval multiplied by the number of cases should be close to the number of cases with $y = 1$. The fit within each interval can be summarized by looking at components of a Pearson-like statistic

$$\frac{[(\text{cases with } y = 1) - (\text{total interval cases})(\text{midpoint})]^2}{(\text{total interval cases})(\text{midpoint})}.$$

These case values can be added to obtain a summary measure of the goodness of fit.

EXAMPLE 4.2.3. For the Chapman data using model (4.1.4), no $\hat{p}_i$ values are greater than .6. Intervals, total cases, coronary incidents, expected values (cases times midpoints), and components are listed below.

$\hat{p}$ Interval	Number of Cases	Coronary Incidents	Cases × Midpoint	Components
[0,.1)	99	5	4.95	0.0005
[.1,.2)	60	10	9	0.1111
[.2,.3)	22	2	5.5	2.2273
[.3,.4)	10	5	3.5	0.6429
[.4,.5)	7	2	3.15	0.4198
[.5,.6)	2	2	1.1	0.7364
				Total = 4.138

The component for the interval [.2,.3) is much larger than the others. If there is a lack of fit in evidence, it is most likely that people with estimated probabilities of a coronary incident between .2 and .3 actually have a considerably lower chance of having a coronary. On the other hand, if the components are at all analogous to χ^2's, even the interval [.2,.3) is not clear evidence of lack of fit. Based on this rather questionable comparison, neither the individual value 2.2273 nor the total 4.138 are unreasonable.

A possible improvement to this technique is, rather than taking the midpoint of the interval, to average the $\hat{p}$'s in the interval. For the interval [.2,.3), the average of the 22 $\hat{p}$'s is .24045 with a corresponding cell component of 2.0461. This indicates even less lack of fit.

4.3 Logistic Regression Diagnostics

We assume that the reader is familiar with diagnostics for standard regression. The diagnostics for logistic regression to be discussed are analogues of common methods used for standard regression. In standard regression, some of the usual diagnostic statistics are the residuals, standardized (studentized) residuals, standardized predicted (deleted) residuals (t residuals), Cook's distances, and the leverages, i.e., the diagonal elements of the projection operator ("hat matrix").

A primary use of residuals is in detecting outliers. However, as we have seen, for data consisting of 0s and 1s, the detection of outliers presents some unusual problems. When there are only two outcomes, it is difficult to claim that seeing either of them constitutes an outlier. If we have too many 0s or 1s in situations where we would not expect them (e.g., too many 1s in situations that we think have a small probability of yielding a 1), then we have a problem, but the problem is best thought of as a lack of fit. Moreover, we have seen that with 0-1 data, perfectly reasonable observations can have "unusually large" residuals.

Another use for residuals is in checking normality. For log-linear models, this can be thought of as checking how well the asymptotic theory holds.

Unlike ANOVA type log-linear models, for 0-1 data the residuals are not asymptotically normal, so, again, the usual residual analysis is not appropriate.

All in all, the residuals (and modified residuals) do not seem very useful in and of themselves. We will concern ourselves with examining leverages and influential observations. In particular, we will examine the logistic regression analogue of Cook's distance that was discussed in Pregibon (1981) and Johnson (1985).

There are two questions frequently asked about influential observations. One is, "In what sense is this observation influential?" The question is crucial. Observations are not "influential" in a vacuum. They may be influential to the estimated regression parameters; they may be influential to the fitted probabilities. They may be influential to just about anything. When examining influential observations, one first decides on the important aspects of the model and then examines influence measures appropriate to those aspects. The author agrees with Johnson (1985) that, typically, the primary concern should be about influence on the fitted probabilities. In logistic regression, Cook's distance is a direct influence measure relative to the fitted regression coefficients, but, as Johnson has shown, it can be viewed as an approximation to his Kullback-Leibler (K-L) divergence measure and Cook and Weisberg's (1982) likelihood distance measure. Although the author's inclination is toward the K-L divergence measure, the absence of readily available computer software dictates that the discussion be focused on Cook's distance.

The second frequently asked question about influential observations is, "Given some influential observations, what do you do about them?" My answer is that you should worry about them. Primarily, you should worry about whether it is more appropriate to ignore the fact that they are influential or eliminate their influence by deleting them from the data and then refitting the model. Of course, all of this is complicated by the fact that whether or not a case is influential depends on what model is being fitted. In the end, the answer to this question must depend on the data and the purpose of the analysis.

Many standard logistic regression programs provide diagnostics. For example, SAS PROC LOGISTIC and BMDP-LR both provide them and they can also be obtained from GLIM. Some sample commands are given in Subsection 4.4.2. In addition, many standard regression programs routinely provide diagnostics and these can also be used to obtain logistic regression diagnostics because the Newton-Raphson method of fitting logistic regression models amounts to doing a series of weighted regressions.

Standard diagnostic quantities for each case are the log odds

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \hat{\beta}_0 + \hat{\beta}_1 Ag_i + \hat{\beta}_2 Ch_i + \hat{\beta}_3 W_i,$$

the predictive probability $\hat{p}_i$, the leverage $\hat{a}_{ii}$, the large sample standard error of $\hat{p}_i$, which is $\sqrt{\hat{p}_i(1-\hat{p}_i)(1-\hat{a}_{ii})}$, the standardized residual

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-\hat{a}_{ii})}},$$

the Pearson residual

$$\tilde{r}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}},$$

the square of which is the ith component of Pearson's chi-square, the deviance residual

$$\pm\sqrt{2[y_i \log(y_i/\hat{p}_i) + (1-y_i)\log((1-y_i)/(1-\hat{p}_i))]}$$

where the sign is taken to be the same as the sign of $y_i - \hat{p}_i$, and a version of Cook's distance for logistic regression. These formulae are for y_i either 0 or 1 and, as discussed earlier, residuals are not very interesting in this case. When y_i is a binomial count between 0 and N_i, the residuals can be useful for reasonably large N_i. The standardized residuals then become

$$r_i = \frac{y_i - N_i\hat{p}_i}{\sqrt{N_i\hat{p}_i(1-\hat{p}_i)(1-\hat{a}_{ii})}}$$

with similar adjustments to other quantities, cf. Subsection 4.4.1.

It should be mentioned that, properly defined, the logistic regression version of Cook's distance for the ith case requires computation of both the estimated logistic regression coefficients and the estimated coefficients when the ith case is deleted from the data. Both sets of estimates require iterative computations and it will be desired to investigate Cook's distance for every case in the data. This can be expensive. To reduce costs, it is common practice to use estimates when the ith case is deleted that are the result of one iteration of Newton-Raphson with starting values taken as the estimates from the full data set. In other words, this involves doing only one weighted regression. These one-step procedures will be the subject of our discussion.

We now present the procedure for obtaining diagnostic values in the context of fitting the model

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 Ch_i + \beta_3 W_i$$

to the Chapman data. Specifically, we discuss how to use the diagnostic procedures in standard regression to obtain diagnostics for logistic regression.

Using a logistic regression program, we get the fit

Variable	Parameter	Estimate	SE
Intercept	β_0	−9.255	2.061
Ag	β_1	0.05300	0.02074
Ch	β_2	0.006517	0.003575
W	β_3	0.01754	0.008243

Having fit the model, create a data file that contains Ag, Ch, W, y, and $\hat{p}$, where y consists of the 0-1 counts y_i and $\hat{p}$ is a new variable that consists of the 200 values for $\hat{p}_i$. This data file is used as input into a regression program that allows (a) transformations of variables, (b) weighted regression, (c) computation of leverages, and (d) computation of Cook's distances. Using the transformation capability, define weights for the regression, say RWT, with $RWT_i = \hat{p}_i(1 - \hat{p}_i)$. Also, define two variables Y_0 and Y with

$$Y_{0i} = \log[\hat{p}_i/(1 - \hat{p}_i)]$$

and

$$Y_i = Y_{0i} + (y_i - \hat{p}_i)/RWT_i\,.$$

See Subsection 4.4.1 for computing methods when the data are not binary.

The variable Y_{0i} can be used to help verify that things are working as they should. Use the regression program to fit

$$Y_{0i} = \beta_0 + \beta_1 Ag_i + \beta_2 Ch_i + \beta_3 W_i + e_i$$

with the weights RWT_i. The regression coefficients from this fit should be identical to those obtained from the logistic regression program.

Now fit

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_2 Ch_i + \beta_3 W_i + e_i$$

with weights RWT_i. This gives estimated regression coefficients that are one additional step of the Newton-Raphson algorithm beyond those obtained by the logistic regression program. In this example, the regression gives

Variable	Parameter	Estimate	SE
Intercept	β_0	−9.256	2.066
Ag	β_1	0.05300	0.02077
Ch	β_2	0.006518	0.003578
W	β_3	0.017539	0.008248

$MSE = .9942$

The parameter estimates are very close to the original logistic regression estimates, but need not be identical to them. (They are close enough that the logistic regression program concluded that the results had converged.) The MSE is close to one, so the reported standard errors are also very close in the regression and the logistic regression. In logistic regression, there is no variance parameter to be estimated, as there is in standard regression, so for logistic regression, anything from a standard regression that involves the MSE must have that involvement eliminated. Appropriate standard errors are the reported standard errors divided by $\sqrt{MSE}$.

From this standard regression fit, we can also obtain leverages, standardized residuals, and Cook's distances. The leverages are precisely those suggested by Pregibon (1981). The standardized residuals and Cook's distances

reported by a standard regression program also involve adjustments for the MSE. For logistic regression, those adjustments must be eliminated. The reported Cook's distances from the standard regression are essentially the one-step logistic regression Cook's distances. The difference in the Cook's distances is that the reported values are the true one-step estimates divided by the MSE. In the discussion below, the reported Cook's distances have been multiplied by MSE to give the appropriate values. In any case, because the values from the program are all being divided by the same number, to make comparisons among cases the values could be used without modification. Another difference is that in standard regression, Cook's distance involves dividing by the number of regression parameters. Often in logistic regression programs, this division is not used. So in this example, to get the Cook's distances given by, say, BMDP-LR, the Cook's distances reported by the regression program have to be multiplied by $4\,MSE$. For what they are worth with 0-1 data, the standardized residuals reported by the regression program times $\sqrt{MSE}$ are the correct standardized residuals.

The nine cases with the highest leverages are

Case	19	38	41	84	108	111	116	153	157
Leverage	.104	.081	.149	.079	.147	.062	.067	.090	.093

The two that really stand out are Cases 41 and 108. Case 41 has $Ag = 40$, $W = 169$, and $Ch = 520$. This is an exceptionally high cholesterol value. Of the 200 cases, only 9 have Ch values over 400 and only 3 have Ch values over 428. These are Case 116 with $Ch = 453$, Case 38 with $Ch = 474$, and Case 41. A similar phenomenon occurs with Case 108. It has $Ag = 51$, $W = 262$, and $Ch = 269$. The weight of 262 pounds is extremely high within the data set. Of the nine cases with high leverage, only Cases 19, 41, and 111 correspond to men that had coronary incidents.

Denote the Cook's distances as C_i's. There are 32 cases with $C_i \geq .01$. These include all 26 of the individuals who had coronary incidents. Of the other six cases, four were also among the highest leverage cases and the remaining two also had reasonably high leverages.

Only four cases had $C_i \geq .05$. These are

Case	C_i	Leverage
41	.112	.149
86	.078	.008
126	.079	.042
192	.064	.022

All of these correspond to individuals with coronary incidents. If we compare the values $4\,C_i$ to a $\chi^2(4)$ distribution as suggested by Johnson (1985), we can get some global idea of the amount of influence each case is having.

The conclusion is that none of the cases has much effect on the fitted model. The multiplier and *df* of 4 for calibrating C_i were the number of regression coefficients in the logistic regression. Of course, to compare these values to a chi-squared distribution, it is vital that the C_i's be computed properly; i.e., values from a standard regression program have to be multiplied by MSE.

Case 41 is easily the most influential, so it is of interest to examine what happens if this case is deleted from the data. For the most part, the fitted p_i's are similar. The estimated coefficients with Case 41, without Case 41, and standard errors without Case 41 are given below.

Variable	Estimate with Case 41	Estimate without Case 41	SE without Case 41
Intercept	−9.255	−8.813	2.058
Ag	0.05300	0.05924	0.02138
Ch	0.006517	0.004208	0.003881
W	0.01754	0.01714	0.008216

The estimates have changed but not dramatically. Perhaps the most striking aspect is the change in the evidence for the effect of cholesterol. With Case 41 deleted, the estimate divided by the standard error is .004208/.003881 = 1.08. With Case 41 included, this value is .006517/.003575 = 1.82. The inclusion of cholesterol in the model was questionable before; without Case 41, there seems little need for it.

With Case 41 deleted, $G^2(Ag, Ch, W) = 132.8$ on 195 degrees of freedom. $G^2(Ag, W) = 134.0$ on 196 degrees of freedom. The difference in G^2's is $134.0 - 132.8 = 1.2$ with 1 degree of freedom. There is virtually no evidence for including cholesterol in the model. The estimated coefficients without Case 41 using only Ag and W are

Variable	Estimate	SE
GM	−7.756	1.745
Ag	0.06675	0.02013
W	0.01625	0.008042

Thus, we have found that almost all of the evidence for a cholesterol effect is based on the fact that one individual with a very high cholesterol level had a coronary incident. We can just drop that individual and state that for more moderate levels of cholesterol, the numerical level of cholesterol does not enhance our predictive ability. But this conclusion is already indicating that qualitatively different things happen at different cholesterol levels. Why not try to incorporate that idea into the models being considered? One could group cases based on cholesterol levels and fit different models to different groups. Rather than forming groups, perhaps cholesterol levels should be transformed before being used in the model. Whatever course

the eventual analysis takes, Case 41 has directed our attention to the role of cholesterol. We now must question whether the current forms of our models are adequate for approximating the effect of cholesterol or whether the effect of cholesterol may be an oddity caused by one individual who just happened to have a coronary incident and very high cholesterol.

4.4 Model Selection Methods

Formal model selection methods can be based either on stepwise methods or finding best subsets of variables based on some criterion (e.g., Akaike's information). Fitting lots of models can be very expensive because each fit requires an iterative procedure. Stepwise methods are sequential, hence cheaper than best subset methods.

Standard computer programs are available for doing stepwise logistic regression, e.g., BMDP-LR and SAS PROC LOGISTIC. These operate in a fashion similar to standard regression (cf. Christensen, 1996a, 1996b; Draper and Smith, 1981; Weisberg, 1985). They are also very similar to the methods discussed in Sections 6.1 and 6.3. We will not give a detailed discussion.

To the best of the author's knowledge, the only standard computer program available for doing best subset logistic regression is SAS PROC LOGISTIC. This procedure is based on doing score tests, a subject that will be discussed below. In addition, programs for doing standard best subset selection can be used with one-step estimates of logistic regression parameters to identify good candidate models. To do this, the best subset program must allow weighted regression.

EXAMPLE 4.4.1. Model (4.1.1) was fitted to the Chapman data to obtain $\hat{p}_i$'s. We then defined two variables: a weight variable

$$RWT_i = \hat{p}_i(1-\hat{p}_i)$$

and a dependent variable

$$Y_i = \log[\hat{p}_i/(1-\hat{p}_i)] + (y_i - \hat{p}_i)/RWT_i\,.$$

The best subset regression program BMDP-9R was used employing the weights RWT to get best subsets of

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + \beta_3 D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i + e_i\,.$$

(Note the similarities to the procedure for getting diagnostic statistics.) The fits used in comparing various models are not fully iterated maximum likelihood fits. They involve one-step of the Newton-Raphson algorithm starting at the maximum likelihood fit for model (4.1.1). Determinations

of best-fitting models are based on residual sums of squares rather than G^2's.

Based on the C_p statistic, the five best-fitting models are

Variables	C_p
Ag, Ch, W	1.66
Ag, W	2.88
Ag, Ch, H, W	3.13
Ag, S, Ch, W	3.49
Ag, D, Ch, W	3.59

The last three models are among the best because adding a worthless variable to the good model based on Ag, Ch, and W cannot do too much harm. The two most interesting models are precisely those identified earlier by less systematic means. In fact, in this example, the C_p statistics are very similar to the corresponding A^* values.

Of course, the C_p statistics are based on one-step fits. Below, we compare the MLEs for the model with Ag, Ch, and W to the one-step fit.

Variable	MLE	One-Step
Intercept	−9.2559	−9.21822
Ag	0.053004	0.0529624
Ch	0.0065179	0.00647380
W	0.017539	0.0174699

(Note that the MLEs differ slightly from the values given previously. The earlier values were obtained from the program GLIM. These values were obtained from BMDP-LR. It is normal for different [correct] programs to give *slightly* different answers.) The C_p's are not based on fully iterated fits, so it is probably a good idea to consider a larger number of models than one ordinarily would in standard regression. One hopes that the best fitting fully iterated models will be among the best fitting one-step models, but the relationship need not be exact.

This method of obtaining best subsets using one-step approximations is very natural, so it is not surprising that it has been discovered independently several times. The earliest references of which I am aware are Nordberg (1981, 1982).

As mentioned earlier, to the best of my knowledge, the only method for best subset regression that appears in a standard computer package is a method in SAS PROC LOGISTIC that gives the models with the best score tests. Score tests are arrived at in a similar fashion to the procedures discussed above. With regard to Example 4.4.1, to get the score test for dropping all of Ag, S, D, Ch, H, and W, fit the full regression model as indicated in the example but with one exception. The exception is that

RWT and Y are defined as indicated using the $\hat{p}_i$'s, but in Example 4.4.1, the $\hat{p}_i$'s were obtained from a maximum likelihood fit of the full model, whereas for a score test, the $\hat{p}_i$'s are obtained from a maximum likelihood fit of the model that contains only an intercept. The score test statistic for whether the six variables can be dropped is just the sum of squares for regression in the six-variable weighted regression model. The statistic is compared to a $\chi^2(6)$ distribution.

One nice thing about score tests is that the $\hat{p}_i$'s depend just on the intercept-only model, so getting the score statistics for testing any model against the intercept-only model is merely a matter of fitting a regression on that model. In other words, with the same definitions for RWT_i and Y_i, one can test the full model as well as models such as

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_4 Ch_i + e_i$$

and

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_4 Ch_i + \beta_6 W_i + e_i$$

simply by fitting the regressions and evaluating the sums of squares for regression.

Of course, the method presented in Example 4.4.1 is essentially the same except that the $\hat{p}_i$'s are taken from the (presumably more accurate) full model rather than the no-intercept model (which nobody takes seriously as a model). Also, some account of model size is being taken by looking at C_p statistics. If one specified best subset selection using the R^2 criterion in the regression program, *the only difference in the Nordbert and score procedures for choosing the best models would be in the choice of* $\hat{p}_i$*'s.*

4.4.1 Computations for Nonbinary Data

In this section and Section 3, we have considered only the case where the counts are either 0s or 1s. In Section 2.6 and in later chapters, we consider logistic models and/or theory for data involving counts that may be greater than 1. The computing methods discussed here are easily adapted to those situations. If the ith case has N_i trials (i.e., the possible values for y_i are $0, 1, \ldots, N_i$), then the appropriate weights are

$$RWT_i = N_i \hat{p}_i (1 - \hat{p}_i)$$

and the dependent variable in the regressions is

$$Y_i = \log[\hat{p}_i/(1 - \hat{p}_i)] + (y_i - N_i \hat{p}_i)/RWT_i \,.$$

Note that

$$\hat{p}_i/(1 - \hat{p}_i) = N_i \hat{p}_i/(N_i - N_i \hat{p}_i) \,.$$

Often logistic regression computer programs will provide the values $N_i \hat{p}_i$ rather than $\hat{p}_i$ as diagnostics. Finally, if all of the N_i's are large, then looking

at the standardized residuals becomes reasonable. Also, when all the N_i's are large, tests against the saturated model can be validly compared to a chi-squared distribution.

4.4.2 Computer Commands

Below are SAS, BMDP, and GLIM commands for obtaining a logistic regression. The data are in a file 'chapman.dat' with eight columns: the case index, Ag, S, D, Ch, H, W, and Cnt. The file looks like this.

```
  1 44 124  80 254 70 190 0
  2 35 110  70 240 73 216 0
  3 41 114  80 279 68 178 0
  4 31 100  80 284 68 149 0
       data continue
199 50 128  92 264 70 176 0
200 31 105  68 193 67 141 0
```

We begin with SAS commands.

Perhaps the simplest way to fit the logistic regression model (4.1.4) in SAS is to use PROC GENMOD. The first line controls printing. The next four lines involve defining and reading the data and creating a variable "n" that gives the total number of possible successes for each case. The remaining lines specify the model and that a logistic regression is to be performed.

```
options ps=60 ls=72 nodate;
data chapman;
   infile 'chapman.dat';
   input ID Ag S D Ch H W Cnt;
   n = 1;
proc genmod data=chapman ;
   model Cnt/n = Ag Ch W / link=logit
                           dist=binomial;
run;
```

A more powerful program for logistic regression is PROC LOGISTIC.

```
options ps=60 ls=72 nodate;
data chapman;
   infile 'chapman.dat';
   input ID Ag S D Ch H W Cnt;
proc logistic data=chapman descending;
   model Cnt=Ag Ch W / waldcl waldrl plcl
                       influence iplots lackfit rsq;
   output out=chdiag predicted=phat;
run;
```

```
proc print data=chdiag;
run;
proc logistic data=chapman descending;
   model Cnt=Ag S D Ch H W / selection = score
                             best = 3 details;
run;
```

This program includes two calls of PROC LOGISTIC. The first is a standard procedure for obtaining a logistic regression. The second involves model selection. On the line with "proc logistic", one specifies the data being used and the command "descending". The command "descending" is used so that the program models the probabilities of events coded as 1 rather than events coded as 0. In other words, it makes the program model the probability of a coronary incident rather than the probability of no coronary incident. Standard output includes the estimated regression coefficients, standard errors, values of z^2, P *values*, and $e^{\hat{\beta}_k}$'s. The model statement is straightforward, specifying the dependent variable and the predictor variables. After the / on the model line, options are specified. "waldcl" causes the program to give the confidence intervals $\hat{\beta}_k \pm 1.96\,\mathrm{SE}(\hat{\beta}_k)$; call the interval (a, b). "waldrl" causes the program to give the values $e^{\hat{\beta}_k}$ and intervals (e^a, e^b). "plcl" gives alternative confidence intervals for the β_k's based on profile likelihoods. The command "influence" causes diagnostics to be presented, basically everything discussed by Pregibon (1981). This includes leverages, Cook's distance C_i (the same version as BMDP presents), and something called Cbar, which is $(1 - \hat{a}_{ii})C_i$. Index plots are given by specifying "iplots". For binary data, G^2 does not give a valid lack of fit test, "lackfit" gives a test similar in spirit to that discussed in Section 2. "rsq" gives values for R^2 and Adj. R^2, but R^2 is defined differently than it is here. The "output" command creates a SAS data set containing the diagnostics, so they can then be printed or manipulated in other ways.

The second proc logistic line was set up to do model selection. It uses the "selection" option. This can be set to "forward", "backwards", "stepwise", or "score". With the score option and "best = 3", the three one-variable models with the best score statistics, the three best two-variable models, the three best three-variable models, and so on, are all presented.

For BMDP, the commands are similar. You actually run the program BMDP-LR, so no statement of this being a logistic regression procedure is needed. Again the data are specified. The variables to be used are specified along with the dependent variable and the model. Interval and categorical variables must be specified prior to specifying the model.

```
/ INPUT     FILE = 'CHAPMAN.DAT'.
          FORMAT = FREE.
          VARIABLES = 8.
/ VARIABLE  NAMES = Index, Ag, S, D, Ch, H, W, Cnt.
```

```
         USE = Ag TO Cnt.
/ REGRESS   DEPENDENT = Cnt.
         INTERVAL = Ag, S, D, Ch, H, W.
            MODEL = Ag, Ch, W.
            MOVE =   0, 0, 0.
            METHOD = MLR.
/ PRINT     CELLS = MODEL.
/ END
```

The program is actually set up to do forward, backward, or stepwise regression. The "move" command was used to make the program fit only the model desired. Diagnostics are obtained by the "cells = model" specification.

Finally, for anyone who might want to use GLIM (still one of my favorites):

```
$units 200$
$data I Ag S D Ch H W Cnt $
$dinput 6$
$calc n = 1$
$yvar Cnt$
$error binomial n$
$fit Ag+Ch+W$
$display e$
$extract %vl$
$calc ahat=%vl*%wt/%sc$                    (leverages)
$calc r=(Cnt-%fv)/%sqrt(%sc*(1-ahat))$   (std. resids)
$calc C=(ahat/(1-ahat))*(r**2)$          (Cook's distances)
$look ahat r C$
```

"units" specifies the number of cases in the regression. After "dinput 6", DOS versions of GLIM prompt you for a file name, i.e., "chapman.dat". The Cook's distances are the same as those used in SAS and BMDP and 4 times those defined here. GLIM is similar in spirit to PROC GENMOD.

4.5 ANOVA Type Logit Models

In this section, analysis of variance type models for the log odds of a two-category response variable are discussed. We begin with a standard example.

EXAMPLE 4.5.1. Consider the muscle tension data of Example 3.7.1. Recall that the factors and levels are

Factor	Abbreviation	Levels
Change in muscle tension	T	High, Low
Weight of muscle	W	High, Low
Muscle type	M	Type 1, Type 2
Drug	D	Drug 1, Drug 2

and the data are

			Drug (k)	
Tension (h)	Weight (i)	Muscle (j)	Drug 1	Drug 2
High	High	Type 1	3	21
		Type 2	23	11
	Low	Type 1	22	32
		Type 2	4	12
Low	High	Type 1	3	10
		Type 2	41	21
	Low	Type 1	45	23
		Type 2	6	22

Change in tension can be viewed as a response factor. Weight, muscle type, and drug are all explanatory variables. Thus, it is appropriate to model the log odds of having a high change in muscle tension. The three explanatory factors affect the log odds for high tension change. The most general model available is to use a model that includes all main effects and all interactions between the explanatory factors, i.e.,

$$\begin{aligned}\log(p_{1ijk}/p_{2ijk}) &= G + W_i + M_j + D_k \\ &\quad + (WM)_{ij} + (WD)_{ik} + (MD)_{jk} \\ &\quad + (WMD)_{ijk}\,. \end{aligned} \tag{1}$$

As usual, this is equivalent to a model with just the highest-order interactions; in this case,

$$\log(p_{1ijk}/p_{2ijk}) = (WMD)_{ijk}\,.$$

Model (1) can be fit by maximum likelihood. Reduced models can be tested. Estimates and asymptotic standard errors can be obtained. In other words, the analysis of model (1) is similar to that of an (unbalanced) standard ANOVA model or a log-linear model.

Of course, the analysis of model (1) should be similar to that of a log-linear model analysis because in a profound sense (alluded to in Section 2.6

and discussed in detail in Chapter 11), model (1) is precisely the same model as the saturated log-linear model, i.e.,

$$\log(m_{hijk}) = (\tau\omega\mu\delta)_{hijk}\,, \tag{2}$$

where we have used Greek equivalents of T, W, M, and D to emphasize that the parameters in (1) and (2) are different. Note that in both models (1) and (2), there is at least one parameter on the right-hand side for every term on the left-hand side. In both models, the data are fitted perfectly. In examining the correspondence between logit models and log-linear models, it is crucial to keep in mind the fact that this is a prospective study, so

$$p_{1ijk}/p_{2ijk} = m_{1ijk}/m_{2ijk}\,.$$

Now consider a more interesting logit model than the saturated logit model (1). Consider, say,

$$\log(p_{1ijk}/p_{2ijk}) = W_i + (MD)_{jk} \tag{3}$$

where we have eliminated the redundant terms G, M_j, and D_k and assumed that the terms $(WM)_{ij}$, $(WD)_{ik}$, and $(WMD)_{ijk}$ add nothing to model (1). We wish to find the corresponding log-linear model. Model (3) is a model that explains tension change odds, so an effect, say W_i, alters the odds of high tension change. The odds cannot be altered without altering both the probability of high tension change and the probability of low tension change; thus, W_i affects both of these probabilities. In other words, the probabilities (and the expected cell counts) depend on both the tension change level T and the weight W. It follows that the logit effect W_i corresponds to a log-linear model interaction, say $(\tau\omega)_{hi}$. Similarly, the logit effect $(MD)_{jk}$ corresponds to the interaction $(\tau\mu\delta)_{hjk}$. As shown in Section 11.1, the log-linear model must contain $(\omega\mu\delta)_{ijk}$ terms, so model (3) is equivalent to

$$\log(m_{hijk}) = (\tau\omega)_{hi} + (\tau\mu\delta)_{hjk} + (\omega\mu\delta)_{ijk}\,. \tag{4}$$

Inclusion of the terms $(\omega\mu\delta)_{ijk}$ is required to deal with the sampling scheme when thinking of the sampling as product-multinomial (i.e., independent binomials) for every combination of the explanatory factors. This mental device was discussed in the subsection of the introduction on retrospective versus prospective studies.

In fact, these ideas extend to all logit and logistic regression models. Note that the shorthand notation used for ANOVA type log-linear models is easily adapted to ANOVA type logit models. Using this shorthand, the correspondence between logit models and log-linear models is illustrated in Table 4.1.

In each case, the effects in the logit model correspond to log-linear model effects that are the interaction between T and the logit model terms. In

addition, the log-linear models always include the three-way interaction between the explanatory factors. Note that models (3) and (4) correspond to line 7 of Table 4.1. As another example, line 3 of the table indicates that the model

$$\log(p_{1ijk}/p_{2ijk}) = G + W_i + M_j + D_k + (WM)_{ij} + (MD)_{jk}$$

is equivalent to the model

$$\begin{aligned}\log(m_{hijk}) \quad = \quad & \gamma + \omega_i + \mu_j + \delta_k + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk} + (\omega\mu\delta)_{ijk} \\ & + \tau_h + (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} \\ & + (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik}.\end{aligned}$$

Of course, the log-linear model can be written much more simply as

$$\log(m_{hijk}) = (\omega\mu\delta)_{ijk} + (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik}\,.$$

TABLE 4.1. Correspondence Between Some Logit and Log-Linear Models

	Logit Model	Log-Linear Model
1)	{WM}{WD}{MD}	[WMD][TWM][TWD][TMD]
2)	{WM}{WD}	[WMD][TWM][TWD]
3)	{WM}{MD}	[WMD][TWM][TMD]
4)	{WD}{MD}	[WMD][TWD][TMD]
5)	{WM}{D}	[WMD][TWM][TD]
6)	{WD}{M}	[WMD][TWD][TM]
7)	{MD}{W}	[WMD][TMD][TW]
8)	{W}{M}{D}	[WMD][TW][TM][TD]
9)	{W}{M}	[WMD][TW][TM]
10)	{W}{D}	[WMD][TW][TD]
11)	{M}{D}	[WMD][TM][TD]

Given the log-linear models, the logit models can be obtained by subtraction. Using model (4), observe that

$$\begin{aligned}\log(p_{1ijk}/p_{2ijk}) \quad = \quad & \log(m_{1ijk}/m_{2ijk}) \\ = \quad & \log(m_{1ijk}) - \log(m_{2ijk}) \\ = \quad & (\tau\omega)_{1i} + (\tau\mu\delta)_{1jk} + (\omega\mu\delta)_{ijk} \\ & - (\tau\omega)_{2i} - (\tau\mu\delta)_{2jk} - (\omega\mu\delta)_{ijk} \\ = \quad & [(\tau\omega)_{1i} - (\tau\omega)_{2i}] + [(\tau\mu\delta)_{1jk} - (\tau\mu\delta)_{2jk}] \\ = \quad & W_i + (MD)_{jk}\end{aligned}$$

where $W_i \equiv [(\tau\omega)_{1i} - (\tau\omega)_{2i}]$ and $(MD)_{jk} \equiv [(\tau\mu\delta)_{1jk} - (\tau\mu\delta)_{2jk}]$. Thus, model (4) implies model (3). Conversely, as will be seen in Chapter 11, the logit model (3) implies the log-linear model (4).

It is interesting to note that models other than (4) will also imply a logit structure of $\log(p_{1ijk}/p_{2ijk}) = W_i + (MD)_{jk}$. Any reduced model relative to (4) where the reduction involves only the $(\omega\mu\delta)_{ijk}$ terms also implies that $\log(p_{1ijk}/p_{2ijk}) = W_i + (MD)_{jk}$. For example, if $\log(m_{hijk}) = (\tau\omega)_{hi} + (\tau\omega\delta)_{hjk} + (\omega\mu)_{ij} + (\delta)_k$, then $\log(p_{1ijk}/p_{2ijk}) = \log(m_{1ijk}) - \log(m_{2ijk}) = W_i + (MD)_{jk}$. Such models imply that the logit structure holds *plus some additional conditions on the explanatory factors.* The logit structure of model (3) without any other conditions is equivalent to model (4), so if one is fitting logit models, the corresponding log-linear model must contain the three-factor effects $(\omega\mu\delta)_{ijk}$. It will be recalled that our original argument for including the $(\omega\mu\delta)_{ijk}$ effects was based on the existence of product-binomial sampling. This condition is not necessary; a logit model implies the existence of the $(\omega\mu\delta)_{ijk}$ terms regardless of the sampling structure, cf. Section 11.1. Of course, in the absence of product-binomial sampling, it would seem to be difficult to interpret the terms $\log(p_{1ijk}/p_{2ijk})$ because we no longer have $p_{1ijk} + p_{2ijk} = 1$. Fortunately, if we condition on explanatory variables (i.e., if we condition on the marginal totals $n_{\cdot ijk}$), then for any of the standard prospective sampling schemes, the *conditional* sampling scheme is product-binomial and the standard interpretations can be used for the conditional distribution.

Before examining the actual analysis of the muscle tension data, we make one final comment about the logit model—log-linear model relationship. *A logit model can be thought of as a model fitted to a two-factor table, where one factor is tension and the other factor consists of all combinations of weight, muscle type, and drug.* The smallest interesting log-linear model is the model of independence:

$$\log(m_{hijk}) = \tau_h + (\omega\mu\delta)_{ijk}\,.$$

Looking at $\log(p_{1ijk}/p_{2ijk}) = \log(m_{1ijk}) - \log(m_{2ijk})$, we see that this model corresponds to a model $\log(p_{1ijk}/p_{2ijk}) = \tau_1 - \tau_2 \equiv G$, i.e., just fitting a grand mean. Intuitively, this would be the smallest interesting logit model. The saturated model for the two-factor table is the interaction model

$$\log(m_{hijk}) = \tau_h + (\omega\mu\delta)_{ijk} + (\tau\omega\mu\delta)_{hijk}\,,$$

which is the logit model $\log(p_{1ijk}/p_{2ijk}) = G + (WMD)_{ijk}$. The more interesting logit models correspond to modeling the interaction in this two-way table. They posit more interaction than complete independence, but less interaction than the saturated model. Note that thinking of this as a two-way table is also consistent with the idea of product-binomial sampling. The muscle tension data corresponds to a 8×2 table. Each row is a distinct set of explanatory variables, indexed by ijk. The columns are the two categories of the response, indexed by h. Each row is thought of as an independent binomial, so the row totals should be fixed by inclusion of a main effect for rows, i.e., the W, M, D three-way interaction.

We now return to the data analysis. Table 4.2 gives a list of logit models, df, G^2, P *values*, and $A-q$ values. The df's, G^2's, P's, and $A-q$'s were actually obtained by fitting the corresponding log-linear models. Clearly, the best fitting logit models are the models {MD}{W} and {WM}{MD}. Both involve the muscle type—drug interaction and a main effect for weight. One of the models includes the muscle type—weight interaction.

TABLE 4.2. Statistics for Logit Models

Logit Model	df	G^2	P	$A-q$
{WM}{WD}{MD}	1	0.111	.7389	−1.889
{WM}{WD}	2	2.810	.2440	−1.190
{WM}{MD}	2	0.1195	.9417	−3.8805
{WD}{MD}	2	1.059	.5948	−2.941
{WM}{D}	3	4.669	.1966	−1.331
{WD}{M}	3	3.726	.2919	−2.274
{MD}{W}	3	1.060	.7898	−4.940
{W}{M}{D}	4	5.311	.2559	−2.689
{W}{M}	5	11.35	.0443	1.35
{W}{D}	5	12.29	.0307	2.29
{M}{D}	5	7.698	.1727	−2.302

We now take a closer look at the logit model {MD}{W}. As mentioned earlier, Table 4.2 was obtained by fitting the log-linear models corresponding to the logit model. The log-linear model corresponding to {MD}{W} is [WMD][TMD][TW]. Each logit model term becomes an interaction with the response factor T and there is an interaction between all of the explanatory factors. The estimated expected cell counts for [WMD][TMD][TW] are given in Table 4.3.

TABLE 4.3. Estimated Expected Cell Counts for the Log-Linear Model [WMD][TMD][TW]

			Drug (k)	
Tension (h)	Weight (i)	Muscle (j)	Drug 1	Drug 2
High	High	Type 1	2.31	20.04
		Type 2	23.75	11.90
	Low	Type 1	22.68	32.96
		Type 2	3.26	11.10
Low	High	Type 1	3.69	10.97
		Type 2	40.24	20.10
	Low	Type 1	44.32	22.03
		Type 2	6.74	22.90

By taking the ratio of the high tension change estimates to the low tension change estimates, we obtain the estimated odds from the logit model. For example, as in Table 4.3, the high-tension, high-weight, type 1, drug 1 estimate is 2.308; the low-tension, high-weight, type 1, drug 1 estimate is 3.693. The ratio is $2.308/3.693 = .625$. This is the logit model estimate of the odds of a high-tension change for high-weight, type 1, drug 1. The estimated odds for all cells are given in Table 4.4.

TABLE 4.4. Estimated Odds of High Tension Change for the Logit Model {MD}{W}

		Drug	
Weight	Muscle	Drug 1	Drug 2
High	Type 1	.625	1.827
	Type 2	.590	.592
Low	Type 1	.512	1.496
	Type 2	.483	.485

The estimated odds of having a high tension change are 1.22 times greater for high-weight muscles than for low-weight muscles. For example, in Table 4.4, $.625/.512 = 1.22$ but also $1.22 = .590/.483 = 1.827/1.495 = .592/.485$. To put it another way, $\hat{m}_{11jk}\hat{m}_{22jk}/\hat{m}_{12jk}\hat{m}_{21jk} = 1.22$. This corresponds to the main effect for weight in the logit model. The odds also involve a muscle type—drug interaction. The nature of this interaction is easily established. Consider the four estimated odds for high weights, $\hat{m}_{11jk}/\hat{m}_{21jk}$. These are the four values at the top of Table 4.4; e.g., for muscle type 1, drug 1, this is .625. In every muscle type—drug combination other than type 1, drug 2, the estimated odds of having a high tension change are about .6. The estimated probability of having a high tension change is about $.6/(1 + .6) = .375$. However, for type 1, drug 2, the estimated odds are 1.827 and the estimated probability of a high tension change is $1.827/(1 + 1.827) = .646$. The chance of having a high tension change is much greater for the combination muscle type 1, drug 2 than for any other muscle type—drug combination. A similar analysis holds for the low-weight odds $\hat{m}_{12jk}/\hat{m}_{22jk}$ but the actual values of the odds are smaller by a factor of 1.22 because of the main effect for weight.

The other logit model that fits quite well is {WM}{MD}. Tables 4.5 and 4.6 contain the estimated odds of high tension change for this model. The difference between Tables 4.5 and 4.6 is that the rows of Table 4.5 have been rearranged in Table 4.6. This sounds like a trivial change, but examination of the tables shows that Table 4.6 is easier to interpret.

Looking at the type 2 muscles, the high-weight odds are .919 times the low-weight odds. Also, the drug 1 odds are 1.111 times the drug 2 odds.

TABLE 4.5. Estimated Odds for the Logit Model {WM}{MD}

		Drug	
Weight	Muscle	Drug 1	Drug 2
High	Type 1	.809	2.202
	Type 2	.569	.512
Low	Type 1	.499	1.358
	Type 2	.619	.557

TABLE 4.6. Estimated Odds for the Logit Model {WM}{MD}

		Drug	
Muscle	Weight	Drug 1	Drug 2
Type 1	High	.809	2.202
	Low	.499	1.358
Type 2	High	.569	.512
	Low	.619	.557

Neither of these are really very striking differences. For muscle type 2, the odds of a high tension change are about the same regardless of weight and drug. Contrary to our previous model, they do not seem to depend much on weight, and to the extent that they do depend on weight, the odds go down rather than up for higher weights.

Looking at the type 1 muscles, we see the dominant features of the previous model reproduced. The odds of high tension change are 1.622 times greater for high weights than for low weights. The odds of high tension change are 2.722 times greater for drug 2 than for drug 1.

Both models indicate that for type 1 muscles, high weight increases the odds and drug 2 increases the odds. Both models indicate that for type 2 muscles, drug 2 does not substantially change the odds. The difference between the models {MD}{W} and {WM}{MD} is that {MD}{W} indicates that for type 2 muscles, high weight should increase the odds, but {WM}{MD} indicates little change for high weight and, in fact, what change there is indicates a decrease in the odds.

Incidentally, the reason for changing from Table 4.5 to Table 4.6 was the nature of the logit model. The model {WM}{MD} has M in both terms, so it is easiest to interpret when fixing the level of M. For a fixed level of M, the effects of W and D are additive, although the size of those effects change with the level of M.

This analysis of the data on change in muscle tension was intentionally performed at the lowest level of *technical* sophistication. The estimated expected cell counts were obtained by iterative proportional fitting. The

entire analysis was based on these fitted values and the associated likelihood ratio test statistics. For example, conclusions about the importance of estimates were drawn without the benefit of standard errors for those estimates. Obtaining standard errors requires more computational sophistication. In particular, it requires fitting an auxiliary regression as discussed in Sections 6.7 and 10.2. However, it is interesting to see how much can be obtained from such a small computational investment.

There are two ways to fit an analysis of variance type logit (logistic) model. One way is to fit the corresponding log-linear model. The second way is to fit the logit model directly. This section has dealt exclusively with fitting the corresponding log-linear models. Section 1 deals exclusively with fitting the logistic (logit) model directly. Although Section 1 deals specifically with regression models, the procedures for a direct fit of an ANOVA model are similar.

To reiterate, there are two principles that define the correspondence between logit models and log-linear models. Recall that effects in the logit model only involve the explanatory factors; e.g., logit effects for the tension change data never involve T, the response factor, only the explanatory factors. The first principle is that any effect in the logit model corresponds in the log-linear model to an interaction between the response factor and the logit effect. For example, a logit effect {MD} corresponds to a log-linear effect [TMD]. The second principle is that the log-linear model always includes the full interaction between the explanatory factors; e.g., all log-linear models include the [WMD] interaction. These principles also hold for logistic regression models. If g is an index or group of indexes that identify all levels of the predictor variables (i.e., explanatory factors), the log-linear model will have a term $u_{(g)}$ which is essentially the full interaction between the explanatory factors. Also, any linear logistic effect βx_g becomes a log-linear interaction $\eta_h x_g$ where h indexes the two levels of the response factor.

4.5.1 Computer Commands

The muscle tension data are listed in the file 'tenslr.dat' with one column for the number of high tension scores, one column for the low tension scores, and three columns of indices that specify the level of weight (high is 1), muscle type, and drug, respectively.

```
 3  3 1 1 1
21 10 1 1 2
23 41 1 2 1
11 21 1 2 2
22 45 2 1 1
32 23 2 1 2
 4  6 2 2 1
```

```
12 22 2 2 2
```

The following commands fit the model {WM}{WD}{MD} using SAS PROC GENMOD. This procedure works very much like GLIM. Note that the variable "n" is the total number of individuals with each level of weight, muscle type, and drug. As in Subsection 3.7.1, the "class" command is used to distinguish ANOVA type factors from regression predictors.

```
options ps=60 ls=72 nodate;
data tension;
   infile 'tenslr.dat';
   input H L W M D;
   n = H+L;
proc genmod data=tension;
   class W M D;
   model H/n = W*M W*D M*D / link=logit
                             dist=binomial;
run;
proc print data=chdiag;
run;
```

Alternatively, the log-linear model for [WMD][TWM][TWD][TMD] can be fitted as in Subsection 3.7.1. To fit other models such as {WM}{MD} or {WM}{D} using GENMOD, the model statement uses W∗M M∗D or W∗M D, respectively.

4.6 Logit Models for a Multinomial Response

The basic method for dealing with a response variable (factor) with more than two levels is to arrange things so that only two things are compared at a time. One way of doing this is to identify pairs of levels to be compared. For example, if the response factor has R levels, comparing each level to the next level leads to modeling

$$\log(m_i/m_{i+1}), \qquad i = 1, \ldots, R-1\,, \tag{1}$$

or, equivalently,

$$\log(p_i/p_{i+1}), \qquad i = 1, \ldots, R-1\,.$$

These are the odds of getting level i relative to getting level $i+1$. They can be viewed as conditional odds given that either level i or $i+1$ occurs. To illustrate multinomial response models, consider the data of Example 3.7.2 presented in Table 3.1. The data involve factors of which we will treat abortion opinion as a response. The levels of abortion opinion are Yes, No, and Undecided. These indicate levels of support for legalized abortion. The

model scheme indicated by equation (1) dictates looking at a series of odds: the odds of Yes to No and the odds of No to Undecided. In this case, the nominal levels of the response can be rearranged to suit us. For example, we could choose to look at the odds of No to Yes and of Yes to Undecided. These latter odds can be viewed as the conditional odds of No to Yes for people who have a clear opinion, and the odds of Yes to Undecided for people who are not opposed.

An alternative modeling scheme is for each level to be compared to a particular level; e.g., models can be formed for

$$\log(m_i/m_R), \qquad i = 1, \ldots, R-1\,. \tag{2}$$

With abortion opinions given in the order Yes, No, Undecided, these models involve the odds of Yes to Undecided and of No to Undecided. Again, one could (and in this case probably would) rearrange the order of the levels so that the level everything is compared to is a particularly interesting category.

If the same form model is used for each value of i, these methods are equivalent and both are equivalent to fitting a log-linear model. For example, the models

$$\log(m_{ijk}/m_{i+1\ jk}) = w_{2(j)} + w_{3(k)}, \qquad i = 1, \ldots, R-1,$$

and

$$\log(m_{ijk}/m_{Rjk}) = v_{2(j)} + v_{3(k)}, \qquad i = 1, \ldots, R-1,$$

are equivalent. (Note that the w and v parameters will also depend on i.) Both of these models are equivalent to

$$\log(m_{ijk}) = u_{23(jk)} + u_{12(ij)} + u_{13(ik)}\,.$$

Just as in two-category logit models, the interaction between all explanatory factors, $u_{23(jk)}$, is included in the model. The logit effects correspond in the log-linear model to interactions with the response factor. Given the log-linear model, the various logit models can be obtained by looking at differences. For example,

$$\log(m_{ijk}/m_{i+1\ jk}) = \log(m_{ijk}) - \log(m_{i+1\ jk})$$

and

$$\log(m_{ijk}/m_{Rik}) = \log(m_{ijk}) - \log(m_{Rjk})$$

lead to parametrizations such as

$$w_{2(j)} = u_{12(ij)} - u_{12(i+1\ j)}$$

and

$$v_{3(k)} = u_{13(ik)} - u_{13(Rk)}\,.$$

(Note that, as mentioned above, w_2 and v_3 depend on the category i that is being examined.) Fits for all of the models in (1) and (2) can be obtained by fitting one log-linear model.

Another way of reducing several response levels to binary comparisons is to pool response levels. One way to do this is to compare each level to the total of all other levels, e.g., model

$$\log\left(\frac{m_i}{\sum_{h\neq i} m_h}\right), \qquad i = 1, \ldots, R. \tag{3}$$

These are the odds of getting category i relative to not getting level i. Fitting these models requires fitting at least $R-1$ logit models. One log-linear model will not do. With the abortion opinion data, these are models for the odds of Yes to not Yes, the odds of No to not No, and the odds of Undecided to not Undecided. These models focus on one category of response and ignore the structure of all other categories.

If the response levels have a natural ordering, say from smallest to largest, then it may be appropriate to look at *continuation ratios*

$$\log\left(\frac{m_i}{\sum_{h=i+1}^{R} m_h}\right), \qquad i = 1, \ldots, R-1. \tag{4}$$

These are the odds of getting level i relative to getting a category higher than level i. As always, we can rearrange the ordering of the response categories if it suits us. This method works very nicely for the abortion data, even though the response levels have no natural ordering. Think of the categories as being ordered as Undecided, Yes, No. Then the first model here has the odds of undecided to everything else, i.e., the odds of undecided to being decided. The second model has the odds of Yes to No, i.e., the odds of supporting legalized abortion relative to opposing it.

The odds in (4) are actually conditional odds. The probability of level i divided by the probability of a higher level is the odds of getting level i given that level i or higher is obtained. For example, the odds of Support to Oppose are actually conditional on being decided. As is seen in Exercise 4.8.14, fitting continuation ratio models for all i is equivalent to fitting a series of log-linear models.

Yet another possibility is to fit *cumulative logits*,

$$\log\left(\frac{\sum_{h=1}^{i} m_h}{\sum_{h=i+1}^{R} m_h}\right), \qquad i = 1, \ldots, R-1.$$

For abortion opinions ordered as Undecided, Yes, No, these models describe the odds of undecided to decided and the odds of not opposed to opposed.

EXAMPLE 4.6.1. We now examine fitting models to the data on race, sex, opinions on abortion, and age from Section 3.7. In a log-linear model,

the variables are treated symmetrically. The analysis looks for relationships among any of the variables. Here, we consider opinions as a response variable. This changes the analysis in that [RSA] must be included in all models. Table 4.7 presents fits for all the models that include [RSA] and correspond to ANOVA type logit models.

TABLE 4.7. Log-Linear Models for the Abortion Opinion Data

Model	*df*	G^2	$A-q$
[RSA][RSO][ROA][SOA]	10	6.12	−13.88
[RSA][RSO][ROA]	20	7.55	−32.45
[RSA][RSO][SOA]	20	13.29	−26.71
[RSA][ROA][SOA]	12	16.62	−7.38
[RSA][RSO][OA]	30	14.43	−45.57
[RSA][ROA][SO]	22	17.79	−26.21
[RSA][SOA][RO]	22	23.09	−20.91
[RSA][RO][SO][OA]	32	24.39	−39.61
[RSA][RO][SO]	42	87.54	3.54
[RSA][RO][OA]	34	34.41	−33.59
[RSA][SO][OA]	34	39.63	−28.37
[RSA][RO]	44	97.06	9.06
[RSA][SO]	44	101.9	13.9
[RSA][OA]	36	49.37	−22.63
[RSA][O]	46	111.1	19.1

The best fitting model is clearly [RSA][RSO][OA]. This model can be used directly to fit either the models in (1)

$$\begin{aligned}\log(m_{hi1k}/m_{hi2k}) &= w^1_{RS(hi)} + w^1_{A(R)}\,,\\ \log(m_{hi2k}/m_{hi3k}) &= w^2_{RS(hi)} + w^2_{A(k)}\,,\end{aligned}$$

the models in (2)

$$\begin{aligned}\log(m_{hi1k}/m_{hi3k}) &= v^1_{RS(hi)} + v^1_{A(k)}\\ \log(m_{hi2k}/m_{hi3k}) &= v^2_{RS(hi)} + v^2_{A(k)}\,,\end{aligned}$$

or some variation of these. As discussed earlier, the first pair of models looks at the odds of supporting legalized abortion to opposing legalized abortion and the odds of opposing legalized abortion to being undecided. The second pair of models examines the odds of supporting legalized abortion to undecided and the odds of opposing to undecided. Of these, the only odds that seem particularly interesting to the author are the odds of supporting to opposing. In the second pair of models, choosing the category "undecided" as the standard level to which other levels are compared is particularly unintuitive. The fact that undecided is the last category is no reason for it to be chosen as the standard of comparison. Either of the

other categories would be a better standard, so one of these should be used. Neither is obviously better than the other.

Neither of these pairs of models are particularly appealing, so we will only continue the analysis long enough to illustrate salient points and to allow comparisons with other models to be discussed later. The fitted values for [RSA][RSO][OA] are given in Table 4.8.

TABLE 4.8. Fitted Values for [RSA][RSO][OA]

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	65+
White	Male	Support	100.1	137.2	117.5	75.62	70.58	80.10
		Oppose	39.73	64.23	56.17	47.33	50.99	62.55
		Undec.	1.21	2.59	5.36	5.05	5.43	8.35
	Female	Support	138.4	172.0	152.4	101.8	101.7	110.7
		Oppose	43.49	63.77	57.68	50.44	58.19	68.43
		Undec.	2.16	4.18	8.96	8.76	10.08	14.86
Nonwhite	Male	Support	21.19	16.57	15.20	11.20	8.04	7.80
		Oppose	8.54	7.88	7.38	7.11	5.90	6.18
		Undec.	1.27	1.54	3.42	3.69	3.06	4.02
	Female	Support	21.40	26.20	19.98	16.38	13.64	12.40
		Oppose	4.24	6.12	4.77	5.12	4.92	4.83
		Undec.	0.36	0.68	1.25	1.50	1.44	1.77

We consider only the odds of support relative to opposed. The odds can be obtained from the fitted values. For example, the odds for young white males are $100.1/39.73 = 2.52$. The full table of odds is given in Table 4.9.

TABLE 4.9. Estimated Odds of Support versus Oppose

Legalized Abortion
(Based on the log-linear model [RSA][RSO][OA])

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	2.52	2.14	2.09	1.60	1.38	1.28
	Female	3.18	2.70	2.64	2.02	1.75	1.62
Nonwhite	Male	2.48	2.10	2.06	1.57	1.36	1.26
	Female	5.05	4.28	4.19	3.20	2.77	2.57

Note that the values from age to age vary by a constant multiple depending on the ages involved. The odds of support decrease steadily with age. The model has no inherent structure among the four race-sex categories;

however, the odds for white males and nonwhite males are surprisingly similar. Nonwhite females are most likely to support legalized abortion, white females are next, and males are least likely to support legalized abortion. Confidence intervals for log odds or log odds ratios can be found using the methods of Section 10.2 or, alternatively, the methods of Section 11.1.

If we pool categories, we can look at the set of three models generated by (3) or the set of two models generated by (4). The set of three models consists of the odds of supporting, the odds of opposing, and the odds of undecided (in each case, the odds are defined relative to the union of the other categories). The two models from (4) are essentially continuation ratio models. The most interesting definition of these models is obtained by taking the odds of supporting to opposing and the odds of undecided to having an opinion. Fitting the models involves fitting log-linear models to two sets of data.

Eliminating all undecideds from the data, we fit [RSA][RSO][OA] to the $2\times2\times2\times6$ table with only the opinion categories "support" and "oppose." The estimated expected cell counts are given in Table 4.10. Note that the estimated cell counts are very similar to those obtained when undecideds were included in the data. The odds of supporting relative to opposing are given below.

Odds of Support versus Opposed

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	2.52	2.14	2.09	1.60	1.38	1.28
	Female	3.18	2.70	2.64	2.01	1.75	1.62
Nonwhite	Male	2.48	2.11	2.06	1.57	1.36	1.26
	Female	5.08	4.31	4.22	3.22	2.79	2.58

Except for nonwhite females, the odds of support are essentially identical to those obtained with undecideds included. The G^2 for the fit without undecideds is 9.104 with 15 *df*. The G^2 for fitting [RSA][RO][SO][OA] is 11.77 on 16 *df*. The difference in G^2's is not large, so a logit model $\log(m_{hi1k}/m_{hi2k}) = R_{(h)} + S_{(i)} + A_{(k)}$ may fit adequately.

We now pool the support and oppose categories to get a $2\times2\times2\times6$ table in which the opinions are "support or oppose" and "undecided." Again, the model [RSA][RSO][OA] is fitted to the data. For this model, we report only the estimated odds.

Odds of Being Decided on Abortion

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	116.79	78.52	32.67	24.34	22.26	16.95
	Female	83.43	56.08	23.34	17.38	15.90	12.11
Nonwhite	Male	23.76	15.97	6.65	4.95	4.53	3.45
	Female	68.82	46.26	19.25	14.34	13.12	9.99

TABLE 4.10. Estimated Expected Cell Counts with Undecideds Eliminated

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	65+
White	Male	Support	100.2	137.7	117.0	75.62	70.22	80.27
		Oppose	39.78	64.35	55.98	47.38	50.78	62.73
	Female	Support	139.2	172.2	152.3	101.6	101.7	109.9
		Oppose	43.78	63.77	57.71	50.41	58.28	68.05
Nonwhite	Male	Support	20.67	16.96	15.48	11.00	8.07	7.81
		Oppose	8.33	8.04	7.52	7.00	5.93	6.19
	Female	Support	20.84	25.17	20.21	16.78	13.98	12.97
		Oppose	4.11	5.84	4.79	5.22	5.02	5.03

Again, the estimated odds vary from age to age by a constant multiple. The odds decrease with age, so older people are less likely to take a position. White males are most likely to state a position. Nonwhite males are least likely to state a position. White and nonwhite females have odds of being decided that are somewhat similar.

The G^2 for [RSA][RSO][OA] is 5.176 on 15 *df*. The G^2 for the smaller model [RSA][RO][SO][OA] is 12.71 on 16 *df*. The difference is very large. Although a main-effects-only logit model fits the support-opposition table quite well, to deal with the undecided category requires a race-sex interaction.

We have pretty much exhausted what can be done easily by fitting ANOVA type log-linear models using iterative proportional fitting. However, computer programs are readily available for direct fitting of logit models. We illustrate some results for modeling the odds of support relative to opposition with undecideds eliminated from the data.

The model that we considered in detail was [RSA][RSO][OA]. This is equivalent to

$$\log(m_{hi1k}/m_{hi2k}) = (RS)_{hi} + A_k \tag{5}$$

which models the odds of supporting legalized abortion. To fit this model directly, we need to provide a computer program with the counts for all cells indicating support, i.e.,

$$n_{hi1k}, \qquad h = 1, 2,\ i = 1, 2,\ k = 1, \ldots, 6\,,$$

and the total of support and opposition for all cells

$$N_{hik} = n_{hi\cdot k} = n_{hi1k} + n_{hi2k}\,.$$

For example, $n_{1111} = 96$, $n_{1211} = 140$, $n_{2111} = 24$, $n_{11\cdot 1} = 96 + 44 = 140$, $n_{12\cdot 1} = 140 + 43 = 183$, and $n_{21\cdot 1} = 24 + 5 = 29$. In addition, for each count

and total, we need to provide the program with the corresponding indices h, i, and k. Fitting model (5) directly gives $G^2 = 9.104$ on 15 df, exactly the results from fitting the equivalent model [RSA][RSO][OA].

The table of odds has suggested two things: (1) odds decrease as age increases and (2) the odds for males are about the same. We want to fit models that incorporate these suggestions. Of course, because the data are suggesting the models, formal tests of significance will be even less appropriate than usual, but G^2's still give a reasonable measure of the quality of model fit.

We model the fact that odds are decreasing with age by incorporating a linear trend in ages. We do not have specific ages to associate with the age categories, so we simply use the codes $k = 1, 2, \ldots, 6$ to indicate ages. These scores lead to fitting the model

$$\log(m_{hi1k}/m_{hi2k}) = (RS)_{hi} + \gamma k\,. \tag{6}$$

The G^2 is 10.18 on 19 df, so the linear trend in coded ages fits very well. [Recall that model (5) has $G^2 = 9.104$ on 15 df, so a test of model (6) versus model (5) has $G^2 = 10.18 - 9.104 = 1.08$ on $19 - 15 = 4$ df.]

To incorporate the idea that males have the same odds of support, we recode the indices of the data. Recall that to fit model (5), we had to specify three index variables along with the numbers supporting and the totals. The indices for the $(RS)_{hi}$ terms are $(h, i) = (1, 1), (1, 2), (2, 1), (2, 2)$. We could recode the problem with an index, say $g = 1, 2, 3, 4$, and fit the model

$$\log(m_{g1k}/m_{g2k}) = (RS)_g + A_k$$

and get exactly the same fit. We can choose the recoding as

(h, i)	(1,1)	(1,2)	(2,1)	(2,2)
g	1	2	3	4

Note that, together, the subscripts g and k still distinguish all of the cases for which data are provided.

This recoding can now be modified, so models that treat males the same can be specified. If we want to treat males the same, then the codes for white males $g = 1$ and nonwhite males $g = 3$ must be made the same. On the other hand, we still have distinct data for white males and nonwhite males, so the fact that there are two replications on males must be accounted for. To treat males the same, recode g as (f, e) with

	wm	wf	nm	nf
g	1	2	3	4
f	1	2	1	3
e	1	1	2	1

where e is an index for replications and the codes wm, wf, nm, nf indicate white males, white females, nonwhite males, and nonwhite females,

respectively. Now fit the model

$$\log(m_{fe1k}/m_{fe2k}) = (RS)_f + A_k\,. \tag{7}$$

The two male groups are only distinguished by the subscript e, and e does not appear on the right-hand side of the model, so the two male groups will be modeled identically. In fact, to use a logistic regression program, you typically do not even need to define the index e. But whether you define it or not, it exists implicitly in the model.

Model (7) is, of course, a reduced model relative to model (5). Model (7) has $G^2 = 9.110$ on 16 *df*, so the comparison between models has $G^2 = 9.110 - 9.104 = .006$ on $16 - 15 = 1$ *df*. We have lost almost nothing by going from model (5) to model (7).

Finally, we can write a model that incorporates both the trend in ages and the equality for males

$$\log(m_{fe1k}/m_{fe2k}) = (RS)_f + \gamma k\,. \tag{8}$$

This has $G^2 = 10.19$ on 20 *df*. Thus, relative to model (5), we have dropped 5 *df* from the model, yet only increased the G^2 by $10.19 - 9.10 = 1.09$.

For the alternative parametrization,

$$\log(m_{fe1k}/m_{fe2k}) = \mu + (RS)_f + \gamma k\,,$$

the estimates and standard errors using the side condition $(RS)_1 = 0$ are

Parameter	Estimate	SE	Est./SE
μ	1.071	.1126	9.51
$(RS)_1$	0	—	—
$(RS)_2$	.2344	.09265	2.53
$(RS)_3$	.6998	.2166	3.23
γ	−.1410	.02674	−5.27

All of the terms seem important. With this side condition, $(\widehat{RS})_2$ is actually an estimate of $(RS)_2 - (RS)_1$, so the z score 2.53 is an indication that white females have an effect on the odds of support that is different from males. Similarly, $(\widehat{RS})_3$ is an estimate of the difference in effect of nonwhite females and males. The estimated odds of support are

	Age					
Race-Sex	18-25	26-35	36-45	46-55	56-65	65+
Male	2.535	2.201	1.912	1.661	1.442	1.253
White female	3.204	2.783	2.417	2.099	1.823	1.583
Nonwhite female	5.103	4.432	3.850	3.343	2.904	2.522

These show the general characteristics discussed earlier. Also, they can be transformed into (conditional) probabilities of support. Probabilities are

generally easier to interpret than odds. The estimated probability that a white female between 46 and 55 years of age supports legalized abortion is $2.099/(1+2.099) = .677$. The odds are about 2, so the probability is about twice as great that such a person will support legalized abortion rather than oppose it.

Similar ideas of modeling can be applied to the odds of having made a decision on legalized abortion.

Finally, a word about computing. The computations for models (6), (7), and (8) were executed using a computer program specifically designed for logit models. This was done because computer programs based on iterative proportional fitting cannot handle the corresponding log-linear models. Iterative proportional fitting only works for ANOVA type models. However, programs for fitting general log-linear models (e.g., GLIM) can handle the log-linear models that correspond to (6), (7), and (8). The models are found in the usual way. Model (6) corresponds to

$$\log(m_{hijk}) = (RSA)_{hik} + (RSO)_{hij} + \gamma_j k$$

where we have added the highest-order interaction term not involving O and made the (RS) and γ terms depend on the opinion level j. Similarly, models (7) and (8) correspond to

$$\log(m_{fejk}) = (RSA)_{fek} + (RSO)_{fj} + (OA)_{jk}$$

and

$$\log(m_{fejk}) = (RSA)_{fek} + (RSO)_{fj} + \gamma_j k\,,$$

respectively.

In a somewhat different approach to treating response factors, Asmussen and Edwards (1983) allow the fitting of models that do not always include a term for the interactions among the explanatory factors. Instead, they argue that *log-linear models are appropriate for response factors as long as the model allows for collapsing over the response factors onto the explanatory factors*, cf. Section 5.3. These issues will also be discussed at the end of Section 6.8.

4.7 Logistic Discrimination and Allocation

How can you tell Swedes and Italians apart? How can you tell different species of irises apart? How can you identify people who are likely to have a heart attack or commit a crime? One approach is to collect data on individuals who are known to be in each of the populations of interest. The data can then be used to discriminate between the populations. To identify Swedes and Italians, one might collect data on height, hair color,

eye color, and skin complexion. To identify irises, one might measure petal length and width and sepal length and width. Typically, data collected on several different variables are combined to identify the likelihood that someone belongs to a particular population. In a standard discrimination–allocation problem, independent samples are taken from each population. The use of these samples to characterize the populations is referred to as discrimination. Allocation involves identifying the population of an individual for whom only the variable values are known. The factor of interest in these problems is the population, but it is not a response factor in the sense used elsewhere in this chapter. In particular, *discrimination data arises from conducting a retrospective study.* The reader may want to review the subsection of the chapter introduction that discusses retrospective and prospective studies.

There has been extensive work done on the problems of discrimination and allocation. Introductions to the subject are contained in Anderson (1984), Christensen (1990), Hand (1981), Lachenbruch (1975), McLachlan (1992), Press (1984), and Rao (1973). The review article by Cheng and Titterington (1994) relates discriminant analysis to neural networks. Recent work on logistic discrimination includes Cox and Ferry (1991) and O'Neill (1994). Probably, the two most commonly used methods of discrimination are Fisher's linear discriminant function and logistic regression. Fisher's method is based on the idea that each case corresponds to a fixed population and that the variables for each case are observations from a multivariate normal distribution. The normal distributions for the populations are assumed to have different means but the same variances and covariances. The logistic regression approach (or as presented here, the log-linear model approach) treats the distribution for each population as a multinomial. Much of the theoretical work on discriminant analysis is done in a Bayesian setting and both methods lend themselves to the easy computation of posterior probabilities for a case to be in a particular population.

The weakness of Fisher's method is that the assumption of normality with equal covariances is often patently false. The case variables are often percentages, rates, or categorical variables. Even when the case variables are continuous on the entire real line, they are often obviously skewed. Frequently the variance-covariance matrices in the various populations are not even similar, much less identical. Fortunately, Fisher's method is somewhat insensitive (robust) to many of these difficulties, cf. Lachenbruch, Sneeringeer, and Revo (1973) and Press and Wilson (1978). Fisher's method is also easily generalized to handle unequal covariance matrices. The strength of Fisher's method is that for normal data it is more efficient than logistic discrimination, cf. Efron (1975).

EXAMPLE 4.7.1. Aitchison and Dunsmore (1975, p. 212) consider 21 individuals with 1 of 3 types of Cushing's syndrome. Cushing's syndrome

is a medical problem associated with overproduction of cortisol by the adrenal cortex. The three types considered are related to specific problems with the adrenal gland, namely

A−adenoma
B−bilateral hyperplasia
C−carcinoma

The case variables considered are the rates at which two steroid metabolites are excreted in the urine. (These are measured in milligrams per day.) The two steroids are

TETRA – Tetrahydrocortisone

and

PREG – Pregnanetriol.

The data are listed in Table 4.11.

TABLE 4.11. Cushing's Syndrome Data

Case	Type	TETRA	PREG	Case	Type	TETRA	PREG
1	A	3.1	11.70	12	B	15.4	3.60
2	A	3.0	1.30	13	B	7.7	1.60
3	A	1.9	0.10	14	B	6.5	0.40
4	A	3.8	0.04	15	B	5.7	0.40
5	A	4.1	1.10	16	B	13.6	1.60
6	A	1.9	0.40	17	C	10.2	6.40
7	B	8.3	1.00	18	C	9.2	7.90
8	B	3.8	0.20	19	C	9.6	3.10
9	B	3.9	0.60	20	C	53.8	2.50
10	B	7.8	1.20	21	C	15.8	7.60
11	B	9.1	0.60				

The data determine the 3×21 table

	Case														
Type	1	2	3	4	5	6	7	8	$\cdots$	16	17	18	19	20	21
A	1	1	1	1	1	1	0	0	$\cdots$	0	0	0	0	0	0
B	0	0	0	0	0	0	1	1	$\cdots$	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	$\cdots$	0	1	1	1	1	1

The case variables TETRA and PREG are used to model the interaction in this table. The case variables are highly skewed, so, following Aitchison and Dunsmore, we analyze the transformed variables $TL \equiv \log(\text{TETRA})$ and $PL \equiv \log(\text{PREG})$. The transformed data are plotted in Figure 4.2.

Now consider the sampling scheme. For studies of this type, it is best modeled as involving independent samples from the three populations: A, B, and C. The sampling can be viewed as product-multinomial because

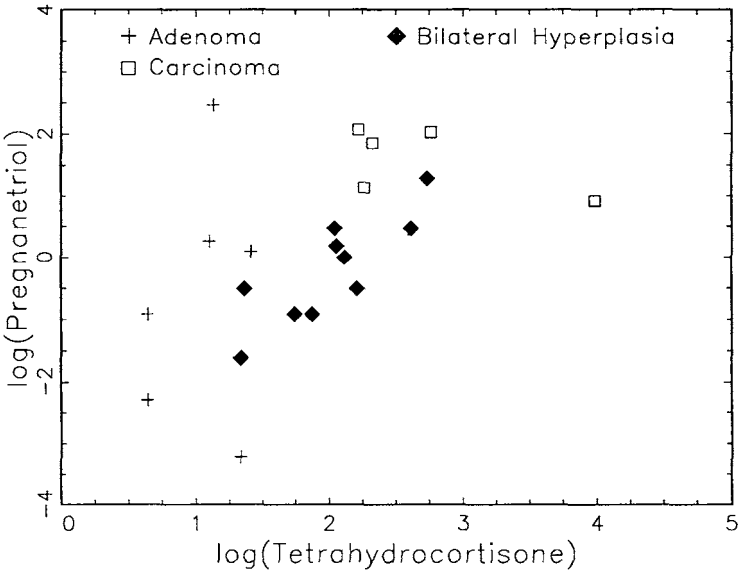


FIGURE 4.2. Cushing's Syndrome Data

all observations are intrinsically discrete. The categories for the product-multinomials consist of all the *observable* combinations of TL and PL. Although TL and PL are apparently continuous variables, the observations taken on TETRA and PREG are only known to be within $\pm.05$ mg and $\pm.005$ mg of their respective nominal values. Thus, the observations are discrete and product-multinomial sampling is appropriate. (Note that the PREG value for case 4 may be a typographical error.) The catch is that there are a huge number of possible categories. Most of these categories have no observations associated with them. If there are S observable combinations of the explanatory factors, we would like to perform a product-multinomial likelihood analysis of the $3 \times S$ table.

Unfortunately, the standard log-linear models for multinomial responses do not have maximum likelihood estimates because most of the column totals in the $3 \times S$ table are zero. To see that MLEs do not exist, observe that if they do exist, the fitted column totals must equal the observed column totals. Most of the S columns in the table will be unobserved, so most of the column totals will be zero. MLEs for log-linear models must be positive so that their logs can be taken; hence, the fitted column totals must all be positive. The fitted column totals cannot be both positive and zero.

In practice, the analysis is conducted as if the observed 3×21 table is obtained via product-multinomial sampling of the three populations. This works well in spite of the fact that the sampling scheme is palpably false. The 21 cases are included in the table precisely because they were observed.

Thus, each column total *must* be at least one. If the sampling scheme were truly product-multinomial, there would be a positive probability of getting column totals equal to zero. Section 11.4 contains a more detailed discussion of these issues and a justification for treating the 3×21 table as product-multinomial. In the current section, we simply present the standard methodology.

One of the tricky things about this is that it *looks like* logistic regression, except that we have more than two possibilities for the response. But treating this as a logistic regression is wrong. In a logistic regression, there are cases with predictor variables associated with them, and each case randomly and independently falls into a response category; e.g., have a coronary incident or don't. In a logistic regression, when the responses are 0s and 1s, every case is a sample from a different population. But in this logistic discrimination, there are only three populations being sampled. The sample sizes are larger, and the values of the predictor variables are actually the results of the sampling.

Because of the sampling scheme, when the samples from the various populations are of different sizes, the values m_{ij} are not directly useful in evaluating the relationship between populations and the predictor variables. For example, if we choose to sample 20,000 people from population A and only 10 from population B, the m_{1j}'s are not comparable to the m_{2j}'s. We must adjust for sample size before relating syndrome type to TL and PL. The evaluation of the relationship is based on the relative likelihoods of the three syndrome types. Thus, for any case j, our interest is in the relative sizes of p_{1j}, p_{2j}, and p_{3j}. Estimates of these quantities are easily obtained from the $\hat{m}_{ij}$'s. Simply take

$$\hat{p}_{ij} = \hat{m}_{ij}/n_{i\cdot}\,. \tag{1}$$

For a new patient of unknown syndrome type but whose values of TL and PL place him in category j, the most likely type of Cushing's syndrome is that which has the largest value among p_{1j}, p_{2j}, and p_{3j}. Clearly, we can estimate the most likely syndrome type. In practice, new patients are unlikely to fall into one of the 21 previously observed categories but the modeling procedure is flexible enough to allow allocation of individuals having any values of TL and PL. This will be discussed in detail in the subsection on allocation.

Discrimination

For each individual j, the variables $(TL)_j$ and $(PL)_j$ have been observed. We seek a model that can be used to classify observations into syndrome type. The main effects model is

$$\log(m_{ij}) = \alpha_i + \beta_j\,, \qquad i = 1, 2, 3\,, \qquad j = 1, \ldots, 21\,.$$

We want to use TL and PL to help model the interaction, so fit

$$\log(m_{ij}) = \alpha_i + \beta_j + \gamma_{1i}(TL)_j + \gamma_{2i}(PL)_j\,, \tag{2}$$

$i = 1, 2, 3$, $j = 1, \ldots, 21$.

This model is very similar to a log-linear version of the logit and logistic models discussed earlier. In particular, it has a separate term β_j for every combination of the explanatory variables. Taking differences gives, for example,

$$\log(m_{1j}/m_{2j}) = (\alpha_1 - \alpha_2) + (\gamma_{11} - \gamma_{12})(TL)_j + (\gamma_{21} - \gamma_{22})(PL)_j$$

which can be written as

$$\log(m_{1j}/m_{2j}) = \alpha + \delta_1(TL)_j + \delta_2(PL)_j\,.$$

Although this looks like a logistic regression model, it has a fundamentally different interpretation. Unlike logistic regression models, it is typically the case that

$$\log\left(\frac{m_{1j}}{m_{2j}}\right) \neq \log\left(\frac{p_{1j}}{p_{2j}}\right)\,.$$

Moreover, the ratio p_{1j}/p_{2j} is not even an odds of type A relative to type B. Both numbers are probabilities, but they are probabilities from different populations. The correct interpretation of p_{1j}/p_{2j} is as a likelihood ratio, specifically the likelihood of type A relative to type B. A value p_{ij} is the likelihood within population i of observing category j. Having fitted model (2), the estimate of the log of the likelihood ratio is

$$\log\left(\frac{\hat{p}_{1j}}{\hat{p}_{2j}}\right) = \log\left(\frac{\hat{m}_{1j}/n_{1\cdot}}{\hat{m}_{2j}/n_{2\cdot}}\right) = \log\left(\frac{\hat{m}_{1j}}{\hat{m}_{2j}}\right) - \log\left(\frac{n_{1\cdot}}{n_{2\cdot}}\right)\,.$$

It will be seen in Chapter 11 that, because interest is directed at comparing probabilities in *different* multinomials, asymptotic variances of estimates will be more complicated than for logistic regression.

Finally, it should be noted that *although odds depend on the sampling scheme, odds ratios do not.* Odds ratios are handled in exactly the same way regardless of whether the sampling scheme is prospective or retrospective.

The G^2 for model (2) is 12.30 on 36 degrees of freedom. As in Section 2.6, although G^2 is a valid measure of goodness of fit, G^2 cannot legitimately be compared to a χ^2 distribution. However, we can test reduced models. The model

$$\log(m_{ij}) = \alpha_i + \beta_j + \gamma_{1i}(TL)_j$$

has $G^2 = 21.34$ on 38 degrees of freedom and

$$\log(m_{ij}) = \alpha_i + \beta_j + \gamma_{2i}(PL)_j$$

has $G^2 = 37.23$ on 38 degrees of freedom. Neither of the reduced models provides an adequate fit. (Recall that χ^2 tests of model comparisons like these were valid.)

Table 4.12 contains estimated probabilities for the three populations. The probabilities are computed using equation (1) and model (2).

TABLE 4.12. Estimated Probabilities

	Group				Group		
Case	*A*	*B*	*C*	Case	*A*	*B*	*C*
1	.1485	.0012	.0195	12	.0000	.0295	.1411
2	.1644	.0014	.0000	13	.0000	.0966	.0068
3	.1667	.0000	.0000	14	.0001	.0999	.0000
4	.0842	.0495	.0000	15	.0009	.0995	.0000
5	.0722	.0565	.0003	16	.0000	.0907	.0185
6	.1667	.0000	.0000	17	.0000	.0102	.1797
7	.0000	.0993	.0015	18	.0000	.0060	.1879
8	.1003	.0398	.0000	19	.0000	.0634	.0733
9	.0960	.0424	.0000	20	.0000	.0131	.1738
10	.0000	.0987	.0025	21	.0000	.0026	.1948
11	.0000	.0999	.0003				

Table 4.13 illustrates a Bayesian analysis. For each case j, it gives the estimated posterior probability that the case belongs to each of the three syndrome types. The data consist of the observed TL and PL values in category j. Given that the syndrome type is i, the estimated probability of observing data in category j is $\hat{p}_{ij}$. Let $\pi(i)$ be the prior probability that the case is of syndrome type i. Bayes theorem gives

$$\hat{\pi}(i|Data) = \frac{\hat{p}_{ij}\pi(i)}{\sum_{i=1}^{3} \hat{p}_{ij}\pi(i)}\,.$$

Two choices of prior probabilities are used in Table 4.13: probabilities proportional to sample sizes, i.e., $\pi(i) = n_{i\cdot}/n_{\cdot\cdot}$ and equal probabilities $\pi(i) = \frac{1}{3}$. Prior probabilities proportional to sample sizes are *rarely appropriate*, but they relate in simple ways to standard output, so we give them more prominence than they probably deserve. Both of the sets of posterior probabilities are easily obtained. The table of proportional probabilities is just the table of $\hat{m}_{ij}$ values. This follows from two facts: first, $\hat{m}_{ij} = n_{i\cdot}\hat{p}_{ij}$ and second, the model fixes the column totals, so $\hat{m}_{\cdot j} = 1 = n_{\cdot j}$. To obtain the equal probabilities values, simply divide the entries in Table 4.12 by the sum of the three probabilities for each case. Cases that are misclassified by either procedure are indicated with a double asterisk in Table 4.13.

Table 4.14 summarizes the classifications. With proportional prior probabilities, 16 of 21 cases are correctly allocated. With equal prior probabilities, 18 of 21 cases are correctly allocated. While Table 4.14 is useful, it ignores

TABLE 4.13. Probabilities of Classification

			Proportional Prior Probabilities			Equal Prior Probabilities		
Case		Group	*A*	*B*	*C*	*A*	*B*	*C*
1		*A*	.89	.01	.10	.88	.01	.12
2		*A*	.99	.01	.00	.99	.01	.00
3		*A*	1.00	.00	.00	1.00	.00	.00
4		*A*	.50	.50	.00	.63	.37	.00
5	**	*A*	.43	.57	.00	.56	.44	.00
6		*A*	1.00	.00	.00	1.00	.00	.00
7		*B*	.00	.99	.01	.00	.99	.01
8	**	*B*	.60	.40	.00	.72	.28	.00
9	**	*B*	.58	.42	.00	.69	.31	.00
10		*B*	.00	.99	.01	.00	.97	.03
11		*B*	.00	1.00	.00	.00	1.00	.00
12	**	*B*	.00	.29	.71	.00	.17	.83
13		*B*	.00	.97	.03	.00	.93	.07
14		*B*	.00	1.00	.00	.00	1.00	.00
15		*B*	.01	.99	.00	.01	.99	.00
16		*B*	.00	.91	.09	.00	.83	.17
17		*C*	.00	.10	.90	.00	.05	.95
18		*C*	.00	.06	.94	.00	.03	.97
19	**	*C*	.00	.63	.37	.00	.46	.54
20		*C*	.00	.13	.87	.00	.07	.93
21		*C*	.00	.03	.97	.00	.01	.99

the clarity of the allocations. For example, case 4 with proportional probabilities is essentially a toss-up between types A and B. That information is lost in Table 4.14. (The probability of type A is slightly greater than one-half.) Another problem with Table 4.14 is that it tends to overestimate how well the discrimination would work on other data. The data were used to form a discrimination procedure and Table 4.14 evaluates how well it works by allocating the same data. This double dipping tends to make the discrimination procedure look better than it really is. Cross-validation can be used to reduce the bias introduced; for related work, see Geisser (1977) and Gong (1986). Finally, it is of interest to note that the difference in Table 4.14 between proportional probabilities and equal probabilities is that under proportional probabilities, one additional case in each of A and C is misclassified into B. That occurs because the prior probability for B is about twice as great as the values for A and C.

TABLE 4.14. Summary of Classifications

	Proportional Prior Probabilities True Group			Equal Prior Probabilities True Group		
Allocated to Group	A	B	C	A	B	C
A	5	2	0	6	2	0
B	1	7	1	0	7	0
C	0	1	4	0	1	5

Readers who are familiar with normal theory discrimination may be interested in the analysis of these data contained in Christensen (1990). Taking the logs of tetrahydrocortisone and pregnanetriol is important in using Fisher's linear discrimination because the original data are clearly nonnormal. Logistic discrimination imposes no such normality requirement. Without the log transform, Fisher's method misclassifies seven observations including five of the six in type A. Logistic discrimination on the untransformed data with proportional priors only misclassifies four observations and gets five of six correct in type A.

Allocation

If you stop and think about it, discrimination seems like a remarkably silly thing to do. Why take cases from known populations and reclassify them when the process of reclassification introduces errors? The reason discrimination is interesting is because one can use a model that discriminates between cases from known populations to predict the population of an unknown case. In our example, a new patient can be measured for TL and PL, and then diagnosed as to type of Cushing's syndrome without direct examination of the adrenal cortex. (I have no idea if this is an accurate

description of medical practice, but it illustrates the kind of thing that can be done.) We now consider the problem of allocating new cases to the populations.

Model (2) includes a separate term β_j for each case, so it is not clear how model (2) can be used to allocate future cases. We will begin with logit models and then work back to an allocation model. Model (2) has 30 parameters, only 9 of which are really of interest. Of these nine, only six are estimable. From (2), we can model the probability ratio of type A relative to type B

$$\begin{aligned} &\log(p_{1j}/p_{2j}) \\ &\quad = \log(m_{1j}/m_{2j}) - \log(n_{1.}/n_{2.}) \qquad (3) \\ &\quad = (\alpha_1 - \alpha_2) + (\gamma_{11} - \gamma_{12})(TL)_j + (\gamma_{21} - \gamma_{22})(PL)_j - \log(n_{1.}/n_{2.})\,. \end{aligned}$$

The log-likelihoods of A relative to C are

$$\begin{aligned} &\log(p_{1j}/p_{3j}) \\ &\quad = \log(m_{1j}/m_{3j}) - \log(n_{1.}/n_{3.}) \qquad (4) \\ &\quad = (\alpha_1 - \alpha_3) + (\gamma_{11} - \gamma_{13})(TL)_j + (\gamma_{21} - \gamma_{23})(PL)_j - \log(n_{1.}/n_{3.})\,. \end{aligned}$$

Fitting model (2) gives the estimated parameters.

Par.	Est.	Par.	Est.	Par.	Est.
α_1	0.0	γ_{11}	-16.29	γ_{21}	-3.359
α_2	-20.06	γ_{12}	-1.865	γ_{22}	-3.604
α_3	-28.91	γ_{13}	0.0	γ_{23}	0.0

where the estimates with values of 0 are really side conditions imposed on the collection of estimates to make it unique.

For a new case with values TL and PL, we plug estimates into equations (3) and (4) to get

$$\log(\hat{p}_1/\hat{p}_2) = 20.06 + (-16.29 + 1.865)TL + (-3.359 + 3.604)PL - \log(6/10)$$

and

$$\log(\hat{p}_1/\hat{p}_3) = 28.91 - 16.29(TL) - 3.359(PL) - \log(6/5)\,.$$

For example, if the new case has a tetrahydrocortisone reading of 4.1 and a pregnanetriol reading of 1.10, then $\log(\hat{p}_1/\hat{p}_2) = .24069$ and $\log(\hat{p}_1/\hat{p}_3) = 5.4226$. The likelihood ratios are

$$\begin{aligned} \hat{p}_1/\hat{p}_2 &= 1.2721 \\ \hat{p}_1/\hat{p}_3 &= 226.45 \end{aligned}$$

and by division,

$$\hat{p}_2/\hat{p}_3 = 226.45/1.2721 = 178.01\,.$$

It follows that type A is a little more likely than type B and that both are much more likely than type C

One can also obtain estimated posterior probabilities for a new case. The posterior odds are

$$\frac{\hat{\pi}(1|Data)}{\hat{\pi}(2|Data)} = \frac{\hat{p}_1}{\hat{p}_2}\frac{\pi(1)}{\pi(2)} \equiv \hat{O}_2$$

and

$$\frac{\hat{\pi}(1|Data)}{\hat{\pi}(3|Data)} = \frac{\hat{p}_1}{\hat{p}_3}\frac{\pi(1)}{\pi(3)} \equiv \hat{O}_3\,.$$

Using the fact that $\hat{\pi}(1|Data) + \hat{\pi}(2|Data) + \hat{\pi}(3|Data) = 1$, we can solve for $\hat{\pi}(i|Data)$, $i = 1, 2, 3$. In particular,

$$\begin{aligned}
\hat{\pi}(1|Data) &= \left[1 + \frac{1}{\hat{O}_2} + \frac{1}{\hat{O}_3}\right]^{-1} = \frac{\hat{O}_2\hat{O}_3}{\hat{O}_2\hat{O}_3 + \hat{O}_3 + \hat{O}_2}\,,\\
\hat{\pi}(2|Data) &= \frac{1}{\hat{O}_2}\left[1 + \frac{1}{\hat{O}_2} + \frac{1}{\hat{O}_3}\right]^{-1} = \frac{\hat{O}_3}{\hat{O}_2\hat{O}_3 + \hat{O}_3 + \hat{O}_2}\,,\\
\hat{\pi}(3|Data) &= \frac{1}{\hat{O}_3}\left[1 + \frac{1}{\hat{O}_2} + \frac{1}{\hat{O}_3}\right]^{-1} = \frac{\hat{O}_2}{\hat{O}_2\hat{O}_3 + \hat{O}_3 + \hat{O}_2}\,.
\end{aligned}$$

Using TETRA = 4.10 and PREG = 1.10, the assumption $\pi(i) = n_{i\cdot}/n_{\cdot\cdot}$ and more numerical accuracy in the parameter estimates than was reported earlier,

$$\begin{aligned}
\hat{\pi}(1|Data) &= .433\\
\hat{\pi}(2|Data) &= .565\\
\hat{\pi}(3|Data) &= .002\,.
\end{aligned}$$

Assuming $\pi(i) = 1/3$ gives

$$\begin{aligned}
\hat{\pi}(1|Data) &= .560\\
\hat{\pi}(2|Data) &= .438\\
\hat{\pi}(3|Data) &= .002\,.
\end{aligned}$$

Note that the values of tetrahydrocortisone and pregnanetriol used are identical to those for case 5; thus, the $\hat{\pi}(i|Data)$'s are identical to those listed in Table 4.13 for case 5.

To use the log-linear model approach illustrated here, one needs to fit a 3×21 table. Typically, a data file of 63 entries is needed. Three rows of the data file are associated with each of the 21 cases. Each data entry has to be identified by case and by type. In addition, the case variables should be included in the file in such a way that all three rows for a case include the corresponding case variables, TL and PL. Model (2) is easily fitted using GLIM.

It is easy to just fit log-linear or logistic models to data such as that in Table 4.11 and get $\hat{m}_{ij}$'s or $\hat{p}_{ij}$'s. If you treat these values as estimated

probabilities for being in the various populations, you are doing a Bayesian analysis with prior probabilities proportional to sample sizes. This is rarely an appropriate methodology.

4.8 Exercises

EXERCISE 4.8.1. The auto accident data of Example 3.2.4 was actually a subset of a four-dimensional table. The complete data are given in Table 4.15. Analyze the data treating severity of injury as a response variable. What conclusions can you reach from examining the $\hat{m}_{hijk}$'s, the odds, and odds ratios?

TABLE 4.15. Automobile Accident Data

		Small Cars ($h = 1$)			
		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver	No	350	150	60	112
Ejected (i)	Yes	26	23	19	80

		Standard Cars ($h = 2$)			
		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver	No	1878	1022	148	404
Ejected (i)	Yes	111	161	22	265

EXERCISE 4.8.2. Breslow and Day (1980) present data on the occurrence of esophageal cancer in Frenchmen. Explanatory factors are age and alcohol consumption. High consumption was taken to be anything over the equivalent of one liter of wine per day. The data are given in Table 4.16. Analyze the data as a logit model. In your analysis, consider the information on ordered age categories.

EXERCISE 4.8.3. The data in the previous experiment is a series of 2×2 tables collected under five different age conditions. This is the same situation as the Mantel-Haenszel setup of Exercise 3.8.9. The Mantel-Haenszel test is one of conditional independence given age. It assumes that the model of no three-factor interaction holds. Respecify the test in terms of logit models.

EXERCISE 4.8.4. Haberman (1978) reports data from the National Opinion Research Center on attitudes toward abortion (cf. Table 4.17). The data

TABLE 4.16. Occurrence of Esophageal Cancer

Age	Alcohol Consumption	Cancer Yes	Cancer No
25-34	High	1	9
	Low	0	106
35-44	High	4	26
	Low	5	164
45-54	High	25	29
	Low	21	138
55-64	High	42	27
	Low	34	139
65-74	High	19	18
	Low	36	88
75+	High	5	0
	Low	8	31

were collected over 3 years. Analyze the abortion attitude data treating attitude as a response variable.

Respondents were identified by their years of education and their religious group. The groups used were Catholics, Southern Protestants, and other Protestants. Southern Protestants were taken as Protestants who live in or south of Texas, Oklahoma, Arkansas, Kentucky, West Virginia, Maryland, and Delaware. Attitudes toward abortion were determined by whether the respondent thought that legal abortions should be available under three sets of circumstances. The three circumstances are (a) a strong chance exists of a serious birth defect, (b) the woman's health is threatened, and (c) the pregnancy was the result of rape. A negative response in the table consists of negative responses to all circumstances. A positive response is three positives. A mixed response is any other pattern. Find an appropriate model for the data. Interpret the model and draw conclusions from the estimates. (Haberman also presents similar data based on three different circumstances: the child is not wanted, the family is poor, and the mother unmarried.)

EXERCISE 4.8.5. Feigl and Zelen (1965), Cook and Weisberg (1982), and Johnson (1985) give data on survival of 33 leukemia patients as a function of their white blood cell count and the existence of a certain morphological characteristic in the cells. The characteristic is referred to as either AG positive or AG negative. The binary response is survival of at least 52 weeks beyond the time of diagnosis. The data are given in Table 4.18. Fit a logistic regression model with separate slopes and intercepts for AG positives and negatives. Examine the data for influential observations. Consider whether a log transformation of the white blood cell count is useful. Evaluate models with (a) the same slope for both AG groups, (b) the same intercept for both

TABLE 4.17. Abortion Attitudes among Caucasian Christians

Year	Religion	Years of Education	Attitude		
			Negative	Mixed	Positive
1974	Prot.	0-8	7	16	49
	Prot.	9-12	10	26	219
	Prot.	12+	4	10	131
1974	Prot. S.	0-8	1	19	30
	Prot. S.	9-12	5	21	106
	Prot. S.	12+	2	11	87
1974	Cath.	0-8	3	9	29
	Cath.	9-12	15	30	149
	Cath.	12+	11	18	69
1973	Prot.	0-8	4	16	59
	Prot.	9-12	6	24	197
	Prot.	12+	4	11	124
1973	Prot. S.	0-8	4	16	34
	Prot. S.	9-12	6	29	118
	Prot. S.	12+	1	4	82
1973	Cath.	0-8	2	14	32
	Cath.	9-12	16	45	141
	Cath.	12+	7	20	72
1972	Prot.	0-8	9	12	48
	Prot.	9-12	13	43	197
	Prot.	12+	4	9	139
1972	Prot. S.	0-8	9	17	30
	Prot. S.	9-12	6	10	97
	Prot. S.	12+	1	8	68
1972	Cath.	0-8	14	12	32
	Cath.	9-12	18	50	131
	Cath.	12+	8	13	64

AG groups, and (c) the same slope and intercept. Examine each model for influential observations.

TABLE 4.18. Data on Leukemia Survival

Survival	Cell Count	AG	Survival	Cell Count	AG
1	2,300	+	1	4,400	−
1	750	+	1	3,000	−
1	4,300	+	0	4,000	−
1	2,600	+	0	1,500	−
0	6,000	+	0	9,000	−
1	10,500	+	0	5,300	−
1	10,000	+	0	10,000	−
0	17,000	+	0	19,000	−
0	5,400	+	0	27,000	−
1	7,000	+	0	28,000	−
1	9,400	+	0	31,000	−
0	32,000	+	0	26,000	−
0	35,000	+	0	21,000	−
0	52,000	+	0	79,000	−
0	100,000	+	0	100,000	−
0	100,000	+	0	100,000	−
1	100,000	+			

EXERCISE 4.8.6. Finney (1941) and Pregibon (1981) present data on the occurrence of vasoconstriction in the skin of the fingers as a function of the rate and volume of air breathed. The data are reproduced in Table 4.19. A constriction value of 1 indicates that constriction occurred. Analyze the data.

EXERCISE 4.8.7. Mosteller and Tukey (1977) reported data on verbal test scores for sixth graders. They used a sample of 20 Mid-Atlantic and New England schools taken from *The Coleman Report*. The dependent variable y was the mean verbal test score for each school. The predictor variables were x_1 — staff salaries per pupil, x_2 — percent of sixth grade fathers employed in white collar jobs, x_3 — a composite score measuring socioeconomic status, x_4 — the mean score on a verbal test administered to teachers, and x_5 — one-half of the sixth grade mothers' mean number of years of schooling. Schools meet a performance standard set for them (by me) if their average verbal test score is above 37. The data are given in Table 4.20.

(a) Using logistic regression on the 0-1 scores, find a good model for predicting whether schools meet the standard.

(b) Using standard regression on y, find several of the best predictive models. Compare these to your logistic regression model.

TABLE 4.19. Data on Vasoconstriction

Constriction	Volume	Rate	Constriction	Volume	Rate
1	0.825	3.7	0	2.0	0.4
1	1.09	3.5	0	1.36	0.95
1	2.5	1.25	0	1.35	1.35
1	1.5	0.75	0	1.36	1.5
1	3.2	0.8	1	1.78	1.6
1	3.5	0.7	0	1.5	0.6
0	0.75	0.6	1	1.5	1.8
0	1.7	1.1	0	1.9	0.95
0	0.75	0.9	1	0.95	1.9
0	0.45	0.9	0	0.4	1.6
0	0.57	0.8	1	0.75	2.7
0	2.75	0.55	0	0.03	2.35
0	3.0	0.6	0	1.83	1.1
1	2.33	1.4	1	2.2	1.1
1	3.75	0.75	1	2.0	1.2
1	1.64	2.3	1	3.33	0.8
1	1.6	3.2	0	1.9	0.95
1	1.415	0.85	0	1.9	0.75
0	1.06	1.7	1	1.625	1.3
1	1.8	1.8			

TABLE 4.20. Verbal Test Scores

Obs.	x_1	x_2	x_3	x_4	x_5	Score	y
1	3.83	28.87	7.20	26.60	6.19	1	37.01
2	2.89	20.10	−11.71	24.40	5.17	0	26.51
3	2.86	69.05	12.32	25.70	7.04	0	36.51
4	2.92	65.40	14.28	25.70	7.10	1	40.70
5	3.06	29.59	6.31	25.40	6.15	1	37.10
6	2.07	44.82	6.16	21.60	6.41	0	33.90
7	2.52	77.37	12.70	24.90	6.86	1	41.80
8	2.45	24.67	−0.17	25.01	5.78	0	33.40
9	3.13	65.01	9.85	26.60	6.51	1	41.01
10	2.44	9.99	−0.05	28.01	5.57	1	37.20
11	2.09	12.20	−12.86	23.51	5.62	0	23.30
12	2.52	22.55	0.92	23.60	5.34	0	35.20
13	2.22	14.30	4.77	24.51	5.80	0	34.90
14	2.67	31.79	−0.96	25.80	6.19	0	33.10
15	2.71	11.60	−16.04	25.20	5.62	0	22.70
16	3.14	68.47	10.62	25.01	6.94	1	39.70
17	3.54	42.64	2.66	25.01	6.33	0	31.80
18	2.52	16.70	−10.99	24.80	6.01	0	31.70
19	2.68	86.27	15.03	25.51	7.51	1	43.10
20	2.37	76.73	12.77	24.51	6.96	1	41.01

EXERCISE 4.8.8. *The Logistic Distribution.*
Show that $F(x) = e^x/(1 + e^x)$ satisfies the properties of a cumulative distribution function (cdf). Any random variable with this cdf is said to have a *logistic distribution*.

EXERCISE 4.8.9. *Stimulus–Response Studies.*
The effects of a drug or other stimulus are often studied by choosing r doses of the drug (levels of the stimulus), say $x_1, \ldots, x_r$, and giving the dose x_j to each of N_j subjects. The data consist of the number y_j who exhibit some predetermined response. Often this response is the death of the subject, but it can be any measure of the effectiveness of the stimulus. Typically in *dose-response studies*, interest centers on the median effective dose, the $ED(50)$, or if the response is death, the median lethal dose, the $LD(50)$. The $LD(50)$ is that dose for which the probability is 0.5 that a subject will die. Frequently, a model of the form

$$\log\left[p_j/(1-p_j)\right] = \alpha + \beta \log(x_j)$$

is fitted to such data. Assume this model holds for any and all doses.

(a) Is the sampling scheme appropriate for a logistic regression?

(b) How could you estimate the $LD(50)$?

Exercises 11.8.2 and 11.8.3 give additional results on inference for the $LD(50)$.

Suppose that for each individual in the population there is a minimum dose x to which the individual will respond. (The individual is assumed to respond to all doses larger than x.) Let the random variable X be this minimum susceptibility for an individual chosen at random.

(c) Give an estimate of the median of X.

(d) Give an estimate of the 90th percentile of the distribution of X.

(e) What is the distribution of X?

EXERCISE 4.8.10. *Probit Analysis.*
An alternative to the logistic analysis of dose-response data is *probit analysis*. In probit analysis, the model is

$$\Phi^{-1}(p_j) = \alpha + \beta \log(x_j)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Graph $\Phi(\cdot)$ and the logistic cdf and compare the general shapes of the distributions. In light of the two previous exercises, give a brief summary of the similarities and differences of logit and probit analysis. For more information on probit analysis, the interested reader should consult Finney (1971).

EXERCISE 4.8.11. Woodward et al. (1941) report several data sets, one of which examines the relationship between exposure to chloracetic acid

and the death of mice. Ten mice were exposed at each dose level. The data are given in Table 4.21. Doses are measured in grams per kilogram of body weight. Fit the logistic regression model of Exercise 4.8.9 and estimate the $LD(50)$. Try to determine how well the model fits the data.

TABLE 4.21. Lethality of Chloracetic Acid

Dose	Fatalities	Dose	Fatalities
.0794	1	.1778	4
.1000	2	.1995	6
.1259	1	.2239	4
.1413	0	.2512	5
.1500	1	.2818	5
.1588	2	.3162	8

EXERCISE 4.8.12. Consider a sample of $j = 1, \ldots, r$ independent binomials $y_j \sim \text{Bin}(N_j, p_j)$, each with a covariate x_j. Suppose that for some cumulative distribution function $F(\cdot)$,

$$p_j = F(x_j).$$

Show that for some transformation $z_j = g(x_j)$ and parameters α and β, this logistic regression model holds:

$$\log\left(\frac{p_j}{1-p_j}\right) = \alpha + \beta z_j.$$

EXERCISE 4.8.13. The data of Exercise 4.8.2 are actually a retrospective study. A sample of cancer patients was compared to a sample of men drawn from the electoral lists of the department of Ille-et-Vilaine in Brittany. Reanalyze the data in light of this knowledge.

EXERCISE 4.8.14. Any multinomial response model can be viewed as the model for an $I \times J$ table. Assume product-multinomial sampling from J independent multinomials each with I categories. Define

$$\pi_{ij} = p_{ij} \Big/ \sum_{h=i}^{I} p_{hj}$$

so that the continuation ratios introduced in Section 4.2 are

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \frac{p_{ij}}{\sum_{h=i+1}^{I} p_{hj}}.$$

Let $r_{ij} = n_{\cdot j} - \sum_{h=1}^{i-1} n_{hj}$; show that the product-multinomial likelihood

$$\prod_{j=1}^{J} \left\{ \frac{n_{\cdot j}!}{\prod_{i=1}^{I} n_{ij}!} \prod_{i=1}^{I} p_{ij}^{n_{ij}} \right\}$$

can be written as the product of binomial likelihoods, i.e.,

$$\prod_{j=1}^{J} \prod_{i=1}^{I-1} \binom{r_{ij}}{n_{ij}} \pi_{ij}^{n_{ij}} (1 - \pi_{ij})^{r_{ij} - n_{ij}}.$$

Using this result and maximum likelihood estimation, show that a set of continuation ratio models can be fitted simultaneously to the entire table by fitting each continuation ratio model separately. Note that the chi-square statistics for fitting each continuation ratio model can be added to get a chi-square statistic for the entire table.

EXERCISE 4.8.15. Give the log-linear model corresponding to (4.1.1).

EXERCISE 4.8.16. Analyze the trauma data that are described in Example 13.2.2.

5

Independence Relationships and Graphical Models

As mentioned in Section 3.7, all of the general principles of testing and estimation presented for three-factor tables also apply when there are additional classification factors. The main difference in working with higher-dimensional tables is that things become more complicated. First, there are many more ANOVA type models to consider. For example, in a four-factor table, there are 113 ANOVA models that include all of the main effects. In five-factor tables, there are several thousand models to consider. Second, a great many of the models require iterative methods for obtaining maximum likelihood estimates. Finally, interpretation of higher-dimensional models is more difficult.

In this chapter, we examine interpretations of models for four and higher-dimensional tables, graphical models, conditions that allow tables to be collapsed, and a variety of graphical models known as recursive causal models.

5.1 Model Interpretations

This section provides tools for interpreting log-linear models for higher-dimensional tables. The interpretations are based on independence and conditional independence. The tools are based on viewing higher-dimensional tables as three-dimensional tables. In Section 2, we present alternative methods based on exploiting the relationships between graph theory and conditional independence.

An example of a model with four factors is

$$\begin{aligned}\log(m_{hijk}) &= u + u_{1(h)} + u_{2(i)} + u_{3(j)} + u_{4(k)} \\ &\quad + u_{12(hi)} + u_{13(hj)} + u_{23(ij)} + u_{123(hij)} \\ &\quad + u_{14(hk)} + u_{24(ik)} \ .\end{aligned}$$

Eliminating redundant parameters gives

$$\log(m_{hijk}) = u_{123(hij)} + u_{14(hk)} + u_{24(ik)} \ .$$

The shorthand notation for this model is [123][14][24]. Our discussion of model interpretations will be based exclusively on the shorthand notation for models.

Consider the model [123][124]. If we think of all combinations of factors 1 and 2 as a single factor, then we get a three-factor table with factors (12), 3, and 4. The model [(12)3][(12)4] becomes a three-dimensional model of conditional independence. Given the levels of factors 1 and 2, factor 3 is independent of factor 4.

This trick of combining factors to reduce a four-factor model into a three-factor model is very useful. The model [123][14] can be considered as a three-factor model in which all combinations of factors 2 and 3 are a single factor. The model [123][14] can then be interpreted as saying that given factor 1, factor 4 is independent of factors 2 and 3. Note that the model puts no constraints on the relationship between factors 2 and 3, and that conditional probabilities involving factors 2, 3, and 4 can change with the level of factor 1, cf. Example 1.1.5.

Using the principle of combining factors, it is easy to see that [123][4] indicates that factor 4 is independent of factors 1, 2, and 3, but that the relationship between factors 1, 2, and 3 is unspecified. Also, [12][34] indicates that factors 1 and 2 may be related, factors 3 and 4 may be related, but 1 and 2 are independent of 3 and 4.

A second useful trick in interpreting models is looking at larger models. If a particular model is true, then any larger model is also true. If the larger model has an interpretation in terms of independence, then the smaller model admits the same interpretation.

For example, consider the three-factor models [12][3] and [12][23]. The smaller model [12][3] indicates that factors 1 and 2 are independent of factor 3. In particular, factors 1 and 3 are independent given the level of factor 2. This is the interpretation of the larger model [12][23]. The interpretation of the larger model is also valid for the smaller model. Note, however, that if two models are both valid and both are interpretable, then the smaller model gives the more powerful interpretation. Often, *we want to identify the smallest interpretable model that fits.*

Now consider the four-factor model [12][13][14]. One larger model is [123][14], so factors 2 and 3 are independent of factor 4 given factor 1.

Similarly, [12][134] and [13][124] are also larger models, so we find that given factor 1, the other three factors are all independent. Note that the structure of the model [12][13][14] makes this interpretation almost self-evident. Factor 1 is included in all three terms, so it is the variable that is fixed in the conditional probabilities. Factors 2, 3, and 4 are in separate terms, so they are independent given factor 1.

EXERCISE 5.1. By examining the probabilities p_{hijk}, show that the three larger models imply conditional independence for factors 2, 3, and 4 in [12][13][14].

A more complicated example is [12][13][24]. One larger model is [13][124]. Thus, given factor 1, factor 3 is independent of factors 2 and 4. Another larger model is [24][123]; thus, given factor 2, factor 4 is independent of factors 1 and 3.

Finally, consider the model [12][13][14][23]. The larger model [123][14] implies that 2 and 3 are independent of 4 given 1. In fact, this is the only simple interpretation associated with [12][13][14][23]. To see this, ignore factor 4. The three-factor model has the terms [12][13][23], which has no simple interpretation as a three-dimensional model. The next largest model is to replace [12][13][23] with [123]. If we do this in the four-factor model, we replace [12][13][14][23] with [123][14]. Any other model with a simple interpretation would have to be larger than [123][14]. However, because [123][14] already has a simple interpretation, we have the best explanation available. (Recall that the smaller the model, the more powerful the interpretation in terms of independence.)

Table 5.1 summarizes the discussion above and also includes some additional models. Note that it is the pattern of the models that determines interpretability. Just as [12][13][14] indicates that 2, 3, and 4 are independent given 1, the model [12][23][24] indicates that factors 1, 3, and 4 are independent given factor 2. Any relabeling of the factors in Table 5.1 gives another interpretable model. It is important to remember that while models imply certain interpretations, more than one model generates the same interpretation. For example, the model [123][14] gives the interpretation that given 1, factors 2 and 3 are independent of factor 4. Conversely, the condition that given 1, factors 2 and 3 are independent of factor 4 implies that [123][14] must hold. However, the independence condition is also consistent with the smaller model [12][13][23][14]. This smaller model is equivalent to the independence condition along with the additional condition that there is no u_{123} interaction.

Goodman (1970, 1971) and Haberman (1974a) introduced the concept of *decomposable* log-linear models. These models are also called *multiplicative*. *The class of decomposable models consists of all models that have closed form maximum likelihood estimates.* They also have simple interpretations in terms of independence or conditional independence. For example, all models for three factors other than [12][13][23] are decomposable. In Table

TABLE 5.1. Some Models and Their Conditional Independence Interpretations

Model	Interpretation
[123][124]	Given 1 and 2, factors 3 and 4 are independent.
[123][14][24]*	Given 1 and 2, factors 3 and 4 are independent.
[123][14]	Given 1, factor 4 is independent of factors 2 and 3.
[12][13][14][23]*	Given 1, factor 4 is independent of factors 2 and 3.
[123][4]	Factor 4 is independent of factors 1, 2, and 3.
[12][23][34][41]	Given 2 and 4, factors 1 and 3 are independent. Given 1 and 3, factors 2 and 4 are independent.
[12][13][14]	Given 1, factors 2, 3, and 4 are all independent.
[12][13][24]	Given 1, factor 3 is independent of factors 2 and 4. Given 2, factor 4 is independent of factors 1 and 3.
[12][34]	Factors 1 and 2 are independent of factors 3 and 4.
[12][13][4]	Factor 4 is independent of factors 1, 2, and 3. Given 1, factor 2 is independent of factor 3.
[12][3][4]	Factor 3 is independent of factors 1, 2, and 4. Factor 4 is independent of factors 1, 2, and 3.
[1][2][3][4]	All factors are independent of all other factors.

*These models imply their interpretations; however, the interpretations do not imply the models.

5.1, all models except the two with asterisks and [12][23][34][41] are decomposable. Note that [12][23][34][41] is not decomposable but is still characterized by its conditional independence relations. It is particularly easy to work with decomposable models because they have very simple structure. Often, results that are difficult or impossible to prove for arbitrary log-linear models can be shown for decomposable models, e.g., Bedrick (1983) and Koehler (1986). An exact characterization of decomposable models is given in the next section.

5.2 Graphical and Decomposable Models

Models that have interpretations in terms of conditional independence are known as *graphical* models. The terminology stems from the relationship of these models to graph theory. Berge (1973) gives a discussion of graph theory that is particularly germane but does not give statistical applications. Edwards and Kreiner (1983) give an overview of the use of graphical log-linear models. More recently, Edwards (1995) provides an introduction to the uses of graphical models in statistics, including applications other than log-linear models. More advanced recent books include Whittaker (1990) and Lauritzen (1996).

Graphical models are determined by their two-factor interactions. The basic idea is that any graphical model containing all of the terms u_{12}, u_{13}, and u_{23} must also include u_{123}. To extend this, consider a graphical model that includes u_{12}, u_{13}, u_{23}, u_{24}, and u_{34}. The terms u_{12}, u_{13}, and u_{23} imply that u_{123} must be in the graphical model and u_{23}, u_{24}, and u_{34} imply that u_{234} must be in the model. A graphical model must contain u_{1234} if it includes all six of the two-factor terms that can be formed from the four factors, i.e., if it includes u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, and u_{34}.

Definition 5.2.1. A model is *graphical* if, whenever the model contains all two-factor terms generated by a higher-order interaction, the model also contains the higher-order interaction. In graph theory, the corresponding idea is that of a *conformal* graph.

Obviously, we need to discuss both models and the two-factor effects generated by higher-order terms. The model [1234][345] is determined by the four-factor term u_{1234} and the three-factor term u_{345}. The four-factor term u_{1234} generates the two-factor terms u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, and u_{34}. The three-factor u_{345} term subsumes the two-factor terms u_{34}, u_{35}, and u_{45}. We will refer to the four-factor term [1234] as generating [12], [13], [14], [23], [24], and [34]. Similarly, the three-factor term [345] generates the two-factor terms [34], [35], and [45]. Conversely, any graphical model that contains the two-factor effects [12], [13], [14], [23], [24], and [34] must include

[1234] (or a larger term that subsumes [1234]) and a graphical model that includes [34], [35], and [45] also includes either [345] or a larger term.

EXAMPLE 5.2.2. *Three-Factor Models.*
The two-factor terms that are possible with three-factors are [12], [13], and [23]. The two-factor effects generate only one higher-order interaction, [123]. The only three-factor models that contain all of the two-factor terms are [123] and [12][13][23]. The model [123] contains all the two-factor effects and the higher-order term, so [123] is graphical. The model [12][13][23] contains all the two-factor effects generated by [123] but does not contain the higher-order term, so it is not graphical. None of the other three-factor models contain all of the two-factor interactions, so, by default, they are all graphical models.

EXAMPLE 5.2.3. *Four-Factor Models*
The model [123][24] is graphical because it includes the three-factor term [123] and *it does not contain all of the two-factor terms generated by any other higher-order terms.* This follows because all higher-order terms other than [123] involve factor 4, so each generates at least 2 two-factor terms that involve factor 4. The model [123][24] includes only one such term, [24], so, by default, the model does not include all the two-factor terms for any higher-order interaction other than [123]. Similarly, the model [123][124] is graphical because the two-factor terms that are present only generate the three-factor interactions in the model.

A model that is not graphical is [12][13][14][23]. It includes all of [12], [13], and [23], but it does not include [123]. The model [123][124][234] is not graphical because it contains all six of the possible two-factor effects but does not contain [1234]. Except for the models indicated by asterisks, all of the models in Table 5.1 are graphical models.

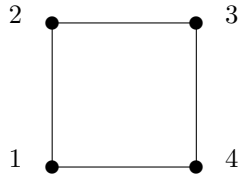
Any log-linear model can be embedded in a graphical model. This follows immediately from the fact that the saturated model is the graphical model having all possible two-factor effects. *To interpret a specific log-linear model, one seeks the smallest graphical models that contain the specific model.*

EXAMPLE 5.2.4. The nongraphical model [12][13][14][23] is a submodel of the graphical model [123][14]. The nongraphical model retains the conditional independence interpretation, 2 and 3 independent of 4 given 1 which is appropriate for the larger graphical model; however, the nongraphical model involves additional constraints.

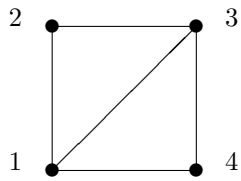
Amazingly, graphical models can be displayed graphically. (Wonders never cease!) In this context, graphs are not directly related to Cartesian coordinates but rather to graph theory. A graph consists of *vertices* (nodes)

and *edges. Vertices correspond to factors in log-linear models. Edges correspond to two-factor effects.* Note that graphs based on two-factor effects would be useless without a convention that dictates how two-factor effects determine a log-linear model. Thus, pictures of graphical models are worthless until after graphical models have been defined. A key feature of this subject is the one-to-one correspondence between graphical log-linear models and graphs. Every model determines a graph and every graph determines a model.

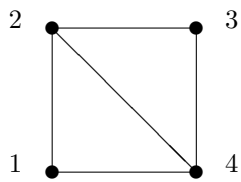
EXAMPLE 5.2.5. Consider the model [12][23][34][41]. Factors are points on a graph (*vertices*) and two-factor interactions are allowable paths (*edges*) between points. The graph is given below.



Now consider the model [123][134]. The two-factor terms generated by [123] are [12], [23], [13] and the terms [13], [34], [14] are generated by [134]. Note that [13] is common to both sets of two-factor terms. The corresponding graph is given below.

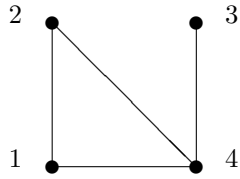


We can also read log-linear models directly from the corresponding graph. For example, the graph



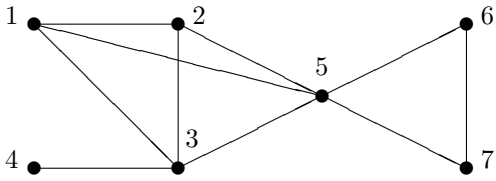
has the two-factor effects [12], [24], [14]; these generate the three-factor term [124]. The graph also contains the edges [23], [34], [24] that generate [234]. We have accounted for all of the edges in the graph, so the model is [124][234].

As another example, consider the following graph.



Again the edges [12], [24], [14] generate the three-factor term [124]. However, this graph does not have all of the edges that generate [234] because the graph does not contain [23]. The term [34] is not included in any larger term, so it must be included separately. The model is [124][34].

Finally, consider a seven-factor graph.

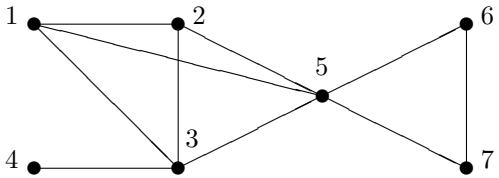


The graph contains all possible edges between the vertices 1, 2, 3, and 5; thus, the graphical log-linear model includes the term [1235]. Similarly, the graph contains all possible edges between the vertices 5, 6, and 7; therefore the log-linear model includes the term [567]. Finally, the graph contains the isolated edge [34], so this term must be in the model. These three terms account for all of the edges in the graph, so the graphical log-linear model is [1235][34][567].

Note that the graph also contains all possible edges between other sets of vertices; for example, all of the edges between 1, 2, and 5 are in the graph so the model includes [125]. However, [125] has already been forced into the model by the inclusion of [1235]. The graphical log-linear model is determined by the largest sets of vertices that include all possible edges between them. The set $\{1, 2, 5\}$ is unimportant because it is contained in the set $\{1, 2, 3, 5\}$. A set of vertices for which all the vertices are connected by edges is *complete*. Both the sets $\{1, 2, 5\}$ and $\{1, 2, 3, 5\}$ are complete. A maximal complete set (i.e., a complete set that is not contained in any

other complete set), is called a *clique*. *The cliques of a graph determine the graphical log-linear model*. The set $\{1, 2, 3, 5\}$ is a clique, but the set $\{1, 2, 5\}$ is not maximal, so it is not a clique. In the graph of the model [1235][34][567], the cliques are $\{1, 2, 3, 5\}$, $\{3, 4\}$, and $\{5, 6, 7\}$. There is an obvious correspondence between the cliques and the $[\cdot]$ notation defining the model. In the future, we will simply indicate the cliques as [1235], [134], and [567]. Because the cliques determine the model, the concept of a clique is of fundamental importance. Surprisingly, that importance need not be made explicit in the remainder of this discussion. However, in Wermuth's method of model selection, the role of cliques cannot be ignored, cf. Section 6.5.

Perhaps the most important reason for graphing log-linear models is that independence relations can be read directly from the graph. To do this, we need to introduce the concept of a chain. A *chain* is simply a sequence of edges that lead from one factor (vertex) to another factor. In the graph



there are a huge number of chains. For example, there is a chain from 1 to 2 to 5, a chain from 1 to 2 to 5 to 6, a chain from 1 to 2 to 5 to 7 to 6, a chain from 1 to 2 to 5 to 3 to 4, and many others. Note that a chain involves not only the end points but also all the intermediate points. In other words, there is a chain from 1 to 3 to 5 to 7, but the graph contains no chain from 1 to 3 to 7 because the graph does not include the edge [37]. We allow chains to begin and end at the same point, e.g., 3 to 1 to 5 to 3; in other words, *round-trips are allowed*. However, we do not allow the path to include a factor more than once. For example, we do not allow 3 to 5 to 6 to 7 to 5 to 1. Even though this path never uses the same edge twice, it does go through factor 5 twice. In a sense, there is no real loss in excluding such paths because we can still get from factor 3 to factor 1 by taking the path 3 to 5 to 1. We are just *not allowing ourselves to drive around in circles*. A formal definition of chains is given below.

Definition 5.2.6. Let h and j be factors and let $\{i_1, \ldots, i_k\}$ be a sequence of factors that are distinct from each other and from h and j. The sequence of edges $C_{hj} = \{[hi_1], [i_1 i_2], \ldots, [i_k j]\}$ is a *chain* between h and j. A graph contains the chain C_{hj} if the graph contains all of the edges included in the chain.

We get a degenerate chain if we start at h, go to another vertex i, and back to h. This is the only situation in which an edge could appear twice within the sequence of edges defining a chain. Nonetheless, this chain only contains one edge.

The key result on independence follows.

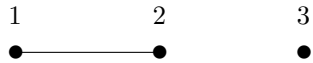
Theorem 5.2.7. Let the sets A, B, and C denote disjoint subsets of the factors in a graphical model. The factors in A are independent of the factors in B given C if and only if every chain between a factor in A and a factor in B involves at least one factor in C.

Proof. See Darroch, Lauritzen, and Speed (1980). □

EXAMPLE 5.2.8. The four-factor model [12][13][24] is illustrated below. It is graphical, so Theorem 5.2.7 applies. Rewrite the model as [31][12][24]. By the theorem, factor 3 is independent of 2 and 4 given 1, factors 3 and 1 are independent of 4 given 2, and factors 3 and 4 are independent given 1 and 2. There are three independence conditions here and all are necessary. For example, the model that only specifies 3 independent of 2 and 4 given 1 is [31][124], not [12][13][24].

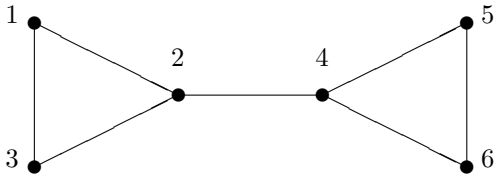
Theorem 5.2.7 also implies certain marginal independence relations. For example, factor 3 is independent of factor 2 given 1. This is a statement about the marginal distribution of factors 1, 2, and 3.

Consider the model [12][3]. There are no chains connecting factors 1 and 2 with 3, so every chain that connects them involves at least one member of the empty set. Thus, 1 and 2 are independent of 3 given the factors in the empty set; i.e., 1 and 2 are independent of 3.



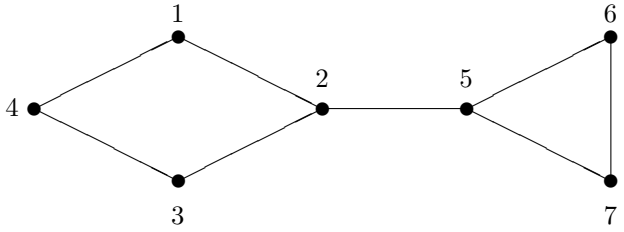
Marginally, we can conclude that 3 is independent of 2 and that 3 is independent of 1.

In the model [123][24][456], 1 and 3 are independent of 5 and 6 given 2 and 4. Similarly, 1 and 3 are independent of 4, 5, and 6 given 2. Also, 5 and 6 are independent of 1, 2, and 3 given 4.



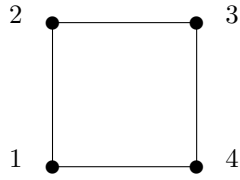
Many marginal independence relationships hold for this model. For example, 1 and 3 are independent of 4 and 5 given 2.

Exercise 5.2. (a) Graph the 10 graphical models in Table 5.1.
(b) List 10 of the independence relationships in [12][23][34][41][25][567].



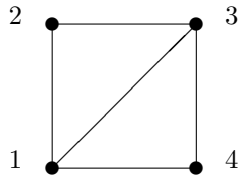
The decomposable models discussed at the end of the previous section form a subset of the graphical models. They are the graphical models that have the additional condition of being *chordal*. The terms given in parentheses in the example below are graph theory terms.

Example 5.2.9. Suppose that a model contains the interactions [12], [23], [34], [41]. We can start at *any* of the points and travel in a cycle (*closed chain*) back to that point. For example, we can travel from 1 to 2, from 2 to 3, from 3 to 4, and from 4 back to 1.



The model (graph) is *chordal* if every such cycle among four or more vertices has a shortcut. A shortcut is called a *chord*. The cycle given above has two

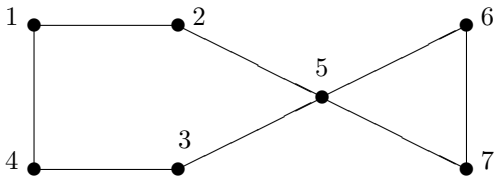
possible shortcuts: [31] and [24]; adding either or both of these would make the model chordal. For example, if the model also includes [31], a cycle from 1 back to 1 can be shortened by traveling [12], [23], [31].



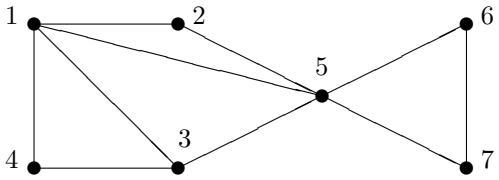
Similarly, a trip from 2 back to 2 can be shortened by traveling [23], [31], [12]. The graphical model generated by the terms [12], [23], [34], [41], and [31] is [123][134]. This decomposable model has the interpretation that given factors 1 and 3, factors 2 and 4 are independent. The maximum likelihood estimates are $\hat{m}_{hijk} = n_{hij\cdot} n_{h\cdot jk} / n_{h\cdot j\cdot}$.

The *length* of a chain is the number of edges in it. The closed chain [12], [23], [34], [41] has length four. The closed chain [12], [23], [31] has length three. A closed chain among four or more vertices is a closed chain of length four or more.

Consider the model [12][25][53][34][41][567] given below in graphical form.



This is not decomposable because the closed chain [12], [25], [53], [34], [41] involves five vertices and has no chords. The model requires the addition of at least two additional two-factor effects to convert it into a decomposable model. Adding one edge to the offending chain still leaves a cycle of length four without a chord. For example, adding [15] leaves the cycle [15], [53], [34], [41] without a chord. The following graph adds both [15] and [13].



This is now the graph of a decomposable log-linear model. All cycles of length four or more have a chord. The corresponding log-linear model is [143][135][125][567].

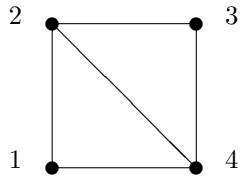
These ideas are formalized in the following definition.

Definition 5.2.10. Consider the chain $C_{hj} = \{[hi_1], [i_1 i_2], \ldots, [i_k j]\}$. The *length* of the chain is the number of edges (distinct elements) in C_{hj}. A chain C_0 is a *reduced chain* relative to C_{hj} if C_0 is a chain between h and j with length less than that of C_{hj} and if every factor involved in C_0 is also involved in C_{hj}. Note that it is the factors (vertices) that define reduced, not the effects (edges). A *closed chain* is a chain between a factor h and itself with length greater than 1. In particular, the chain from h to h, $C_{hj} = \{[hi], [ih]\}$, contains only one (distinct) edge, so it has length 1 and does not form a closed chain. Any two-factor term $[i_r i_s]$ is a *chord* of the closed chain $C_{hh} = \{[hi_1], \ldots, [i_{r-1} i_r], [i_r i_{r+1}], \ldots, [i_{s-1} i_s], [i_s i_{s+1}], \ldots, [i_k h]\}$ if the sequence $\{[hi_1], \ldots, [i_{r-1} i_r], [i_r i_s], [i_s i_{s+1}], \ldots, [i_k h]\}$ is a closed reduced chain relative to C_{hh}. It is allowable for either factor in $[i_r i_s]$ to be h. A model is *chordal* if every closed chain of length $k \geq 4$ generated by the model has a chord that is in the model. A model is *decomposable* if it is both graphical and chordal.

By definition, a closed chain must have a length of at least 2; it follows immediately that a closed chain must have a length of at least 3. Clearly, a closed chain of length three cannot have a chord, so it is natural that the definition of chordal models involves closed chains of length 4 or more. It is possible for a model to be chordal without being graphical. Chordal models have restrictions on the two-factor terms; they place no requirements on higher-order terms. In graph theory, decomposable models correspond to *acyclic hypergraphs.*

Example 5.2.11. The effects [12], [23], [34] define a chain from 1 to 4. The effects [12], [24] define a reduced chain from 1 to 4. The effects in the model [12][23][34][41] define a closed chain from 1 back to 1 but also from 2 back to 2, from 3 back to 3, and from 4 back to 4. To see the last of these,

observe that the model contains the closed chain [41], [12], [23], [34]. The possible chords for these closed chains are [31] and [24]. To see that [24] is a chord, observe that [12], [24], [41] defines a closed reduced chain of [12], [23], [34], [41].



The nongraphical model [12][23][34][41][24] is chordal because any closed chain of length four has a chord that is in the model. This model has no closed chains of length greater than four because it involves only four factors. The model [12][23][34][41][24] is not decomposable because it is not graphical. It contains all of [12], [41], and [24] but not [124]. The corresponding decomposable model is [124][234]. This model generates precisely the two-factor terms [12], [23], [34], [41], and [24], so it is both graphical and chordal.

With four factors, the only graphical model that is not decomposable is [12][23][34][41].

Decomposable models have closed form estimates. We illustrate one simple case.

EXAMPLE 5.2.12. Consider the graph

The corresponding model is [12][13][24]. The probability structure can be read from the model,

$$p_{hijk} = \frac{p_{hi\cdot\cdot}p_{h\cdot j\cdot}p_{\cdot i\cdot k}}{p_{h\cdot\cdot\cdot}p_{\cdot i\cdot\cdot}},$$

where the terms in the numerator are determined by the terms in the model (the marginal probabilities in the numerator correspond to the margins fitted by the model) and the terms in the denominator correspond to the factors that appear in more than one term in the model. Marginal probabilities are estimated from marginal tables, e.g., $\hat{p}_{hi\cdot\cdot} = n_{hi\cdot\cdot}/n_{\cdot\cdot\cdot\cdot}$. The estimated expected counts are

$$\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot}\hat{p}_{hijk} = \frac{n_{hi\cdot\cdot}n_{h\cdot j\cdot}n_{\cdot i\cdot k}}{n_{h\cdot\cdot\cdot}n_{\cdot i\cdot\cdot}}.$$

Decomposable models are closely related to *recursive causal models*. Recursive causal models use ideas from directed graphs to indicate causation. As mentioned in the introduction to Chapter 4, causation is not something that can be inferred from data. Any causation must be inferred from other sources. Recursive causal models are introduced in Section 4. The interested reader can also consult the relevant literature, e.g., Wermuth and Lauritzen (1983), Kiiveri, Speed, and Carlin (1984), and the fine expository paper by Kiiveri and Speed (1982).

The interplay between graph theory and statistics is a fascinating subject with implications for log-linear models, *covariance selection, factor analysis, structural equation models, artificial intelligence*, and *database management*. Reviews are given by Kiiveri and Speed (1982), Edwards (1995), Whittaker (1990), and Lauritzen (1996). For applications to log-linear models, see the references listed in the previous paragraph along with Darroch, Lauritzen, and Speed (1980). These articles cite a wealth of related work including the important contributions of Leo Goodman and Shelby Haberman.

5.3 Collapsing Tables

One important function of statistics is to summarize large batches of numbers. This is such a fundamental aspect of statistics that it is easily overlooked. For example, formal theories of statistics are generally based on the use of sufficient statistics, cf. Cox and Hinkley (1974). Intuitively, a sufficient statistic is simply a summary of the data that is sufficient for drawing valid conclusions about the data. The use of sufficient statistics is an enormous advantage both theoretically and practically.

An early step in analyzing many sets of data is the construction of a table. Tables organize data in a way that makes the data more understandable. Clearly, small tables are easier to understand than large tables. For example, a 3×5 table is typically easier to understand than a $3 \times 5 \times 4$ table. In this section, we establish conditions that, if satisfied, allow us to collapse a $3 \times 5 \times 4$ table of counts into a 3×5 table and still draw valid conclusions. Recall that collapsing is not always possible. Simpson's paradox is precisely the result of collapsing a table that cannot be validly collapsed. First, we discuss collapsing in three-factor tables and then extend the discussion to higher-order tables.

Collapsed tables are also used in analysis of variance. The three-factor ANOVA model

$$\begin{aligned} y_{ijk\ell} &= \mu + \alpha_i + \beta_j + \gamma_k \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + e_{ijk\ell} \,, \end{aligned}$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $\ell = 1, \ldots, N$, is a model for analyzing the $I \times J \times K$ table of means $\bar{y}_{ijk\cdot}$. It is well known that with no three-factor $(\alpha\beta\gamma)$ interaction, each of the two-factor interactions can be examined by looking at the corresponding two-factor marginal table. For example, the two-factor interaction $(\alpha\beta)$ can be investigated using the $I \times J$ table of $\bar{y}_{ij\cdot\cdot}$'s.

The situation with tables of counts is more complex. In general, *if a log-linear model has no three-factor interaction and if all two-factor interactions exist, it is not valid to draw conclusions about two-factor interactions from the two-factor marginal tables.*

As discussed earlier, two-factor interactions are closely related to odds ratios. A three-factor table is said to be collapsible over factor 1 if the odds ratios in the marginal table $p_{\cdot jk}$ are identical to the odds ratios for each row of the three-way table. In other words, we can collapse on rows (factor 1) if for all i, j, j', k, and k',

$$\frac{p_{\cdot jk}p_{\cdot j'k'}}{p_{\cdot j'k}p_{\cdot jk'}} = \frac{p_{ijk}p_{ij'k'}}{p_{ij'k}p_{ijk'}} \,. \tag{1}$$

If this is true, we can draw valid inferences about the relationship between factors 2 and 3 by looking only at the marginal table. (This is clearly an easier task than examining a separate two-way table for each level of factor 1.)

As shown in equation (3.2.3) in the subsection Odds Ratios and Independence Models, equation (1) holds if rows and columns are independent given layers. So, under this model, the relationship between columns and layers does not depend on rows. Similarly, the row-layer relationship can be investigated in the row-layer marginal table if rows and columns are independent given layers.

Note that in order to have equation (1) hold, it is not necessary that rows and columns be independent given layers. It is easily seen that if rows and layers are independent given columns, then (1) still holds. Moreover, if both models [13][23] and [12][23] hold, then we must have [1][23], so rows are independent of both columns and layers. Obviously, in this case, collapsing over rows is allowable.

The validity of collapsing over a factor is a property of the parameters p_{ijk}. Data analysis is based on the observed values n_{ijk}. It is important to realize that the MLE of the odds ratio $p_{\cdot jk}p_{\cdot j'k'}/p_{\cdot j'k}p_{\cdot jk'}$ under either model [13][23] or [12][23] is $n_{\cdot jk}n_{\cdot j'k'}/n_{\cdot j'k}n_{\cdot jk'}$ and that this is also the MLE from the marginal table of $n_{\cdot jk}$'s. Thus, data analysis is identical whether working with the three-dimensional table or with the collapsed table. Collapsed tables are such a useful and intuitive tool that we have already used them in data analysis. The reader should note that collapsed tables were an integral part of both Examples 3.2.2 and 3.2.3.

Our results on collapsing three-factor tables are summarized in the next theorem.

Theorem 5.3.1.

(a) If the model [13][23] holds, then the relationship between factors 2 and 3 can be examined in the marginal table $n_{\cdot jk}$ and the relationship between factors 1 and 3 can be examined in the marginal table $n_{i\cdot k}$.

(b) If either model [13][23] or [12][23] holds, then the relationship between factors 2 and 3 can be examined in the marginal table $n_{\cdot jk}$.

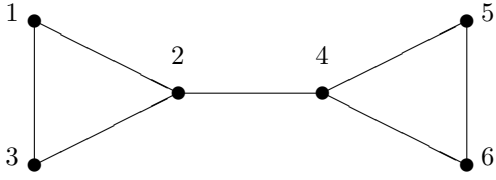
(c) If model [1][23] holds, then the relationship between factors 2 and 3 can be examined in the marginal table $n_{\cdot jk}$.

To extend collapsibility conditions to higher-order tables, use the same tricks as were used in Section 1 for interpreting higher-order models: reindexing and using larger models. Consider a table with five factors: 1, 2, 3, 4, and 5. Suppose our model is [1234][45][35]. This is contained in the model [1234][345]. By considering this as a three-factor table with factors 1-2, 3-4, and 5, we have that factor 1-2 and factor 5 are independent given factor 3-4. Thus, collapsing over factor 5 to examine the marginal table of factors 1, 2, 3, and 4 is valid. Also, collapsing over factors 1 and 2 gives a valid marginal table for examining factors 3, 4, and 5.

It is particularly easy to read off collapsibility from a graphical model.

Corollary 5.3.2. Let the sets A, B, and C denote a partition of the factors in a graphical model such that every chain between a factor in A and a factor in B involves at least one factor in C; then the relationships among the factors in A and C can be examined in the marginal table obtained by summing over the factors in B.

EXAMPLE 5.3.3. The model [123][24][456] can be graphed as below.



It follows that accurate conclusions can be drawn from the marginal tables $n_{123\cdots}$, $n_{1234\cdot\cdot}$, $n_{\cdots 456}$, and $n_{\cdot 2\cdot 456}$.

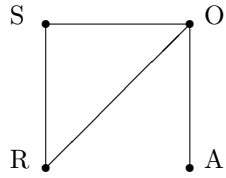
5.4 Recursive Causal Models

This section examines a class of graphical models that are useful for analyzing causal relationships. Causation is not something that can be established by data analysis. Establishing causation requires logical arguments that go beyond the realm of numerical manipulation. For example, a well-designed randomized experiment can be the basis for conclusions of causality, but the analysis of an observational study yields information only on correlations. When observational studies are used as a basis for causal inference, the jump from correlation to causation must be made on nonstatistical grounds. In this section, we consider a class of graphical models that have causation built into them. The discussion focuses on appropriate graphs and their interpretations. Not all of the graphical models in this class correspond to log-linear models; thus, the new class is distinct from the graphical models considered in Section 2. For the models considered here, the numerical process of estimation is exceedingly simple.

The graphical models considered in this section are *recursive causal models*. Unlike most of the methods considered in Chapter 4, recursive causal models allow for multiple response factors. With multiple response factors, a given factor can serve as both a response, relative to some causal factors, and as a cause for other response factors. The term "recursive" indicates that response factors are not allowed to serve, even indirectly, as causes of themselves.

We begin with a discussion of models that involve only one response factor. In particular, we consider the abortion opinion data discussed in Sections 3.7 and 4.6.

EXAMPLE 5.4.1. *Abortion Opinion Data.*
The factors involved in the abortion opinion data are race R, sex S, opinion O, and age A. In Chapter 6, one of the better models found for these data is [RSO][OA]. This is a graphical model.



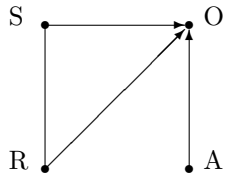
The model [RSO][OA] indicates that Race and Sex are independent of Age given Opinion. While this may explain the data well, it is difficult to imagine a social process that could cause independence between such tangible characteristics as Race, Sex, and Age given something as ephemeral as

Opinion. In particular, it violates Asmussen and Edwards' (1983) criteria for response models (see the ends of Sections 4.6 and 6.8).

In Chapter 4, we have argued that when analyzing a response, one should condition on all explanatory factors. With Opinion taken as a response, any log-linear model should include the interaction term [RSA] for the explanatory factors. In Section 4.6, we found that [RSA][RSO][OA] was a reasonable model. This model is not graphical. For example, the three-factor terms in the model, [RSA] and [RSO], imply the existence of [SA] and [SO] interactions. Taken together with the [OA] term, the model includes all of the two-factor terms included in [SAO]. By definition, if all these two-factor effects are included, a graphical model must also include [SAO]. Thus, [RSA][RSO][OA] is not graphical.

Note that based on the model [RSA][RSO][OA], any logit model for Opinion has an effect (RS) for Race-Sex interaction and a main effect A for age. In the discussion below, we present a recursive causal model that incorporates the same effects. The difference is that the recursive causal model is not a log-linear model but a *conjunction* of log-linear models.

While [RSO][OA] is a graphical model, it is not a recursive causal model for Opinion. Given below is the graph of a recursive causal model in which Opinion is the response, Race, Sex, and Age are direct causes of Opinion, and there is a joint effect for Race and Sex.



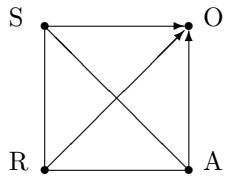
This is very similar to the graph for [RSO][OA]; however, some of the edges have been replaced by arrows. Arrows are called *directed edges*. Edges without arrowheads are *undirected edges*. Directed edges point to response factors; O is the only response factor. In this graph, the directed edges originate at the explanatory factors. The factors R, S, and A are each called a *direct cause* of O because there is a directed edge from each of R, S, and A to O. The undirected edge between R and S is unchanged from the graph of [RSO][OA]; the edge represents an interaction between R and S. Age involves no undirected edges.

With no loss of generality, the probability model corresponding to these four factors can be written as

$$\begin{aligned}\Pr(R=h,S=i,O=j,A=k)&\\ =\Pr(O=j|R=h,S=i,A=k)&\Pr(R=h,S=i,A=k).\end{aligned}$$

Here, the probability is written as the product of a conditional probability of the response factor given its direct causes, $\Pr(O = j|R = h, S = i, A = k)$ and another term, $\Pr(R = h, S = i, A = k)$, that involves only the explanatory factors. Each of these terms is to be modeled with a log-linear model.

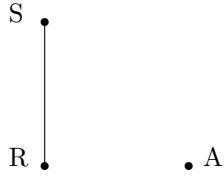
The term $\Pr(O = j|R = h, S = i, A = k)$ is a conditional probability; so, as discussed earlier, the corresponding log-linear model should include an [RSA] term. The log-linear model used is the graphical model that incorporates the explanatory factor edges for [RSA] and changes directed edges involving the four factors to undirected edges. In the following graph, the directed edges are retained to emphasize that there are two steps involved.



Changing directed edges to undirected edges, the graphical log-linear model is clearly the saturated model, cf. Section 2. The maximum likelihood estimate of $\Pr(R = h, S = i, O = j, A = k)$ is $n_{hijk}/n_{....}$ and, thus, the maximum likelihood estimate of $\Pr(O = j|R = h, S = i, A = k) \equiv p_{hijk}/p_{hi\cdot k}$ is

$$\frac{\hat{p}_{hijk}}{\hat{p}_{hi\cdot k}} = \frac{n_{hijk}}{n_{hi\cdot k}}.$$

The probability model for the explanatory factor term $\Pr(R = h, S = i, A = k)$ is also a log-linear model determined by the graph of the recursive causal model. The log-linear model is the graphical model obtained by dropping the response factor and the directed edges. It is given below.



The log-linear model for this graph is [RS][A]. It determines a marginal distribution for the explanatory factors. The model is that

$$\Pr(R = h, S = i, A = k) = \Pr(R = h, S = i)\Pr(A = k).$$

or, equivalently,

$$p_{hi\cdot k} = p_{hi\cdot\cdot} p_{\cdot\cdot\cdot k}.$$

The maximum likelihood estimates are

$$\hat{p}_{hi\cdot k} = \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}} \frac{n_{\cdot\cdot\cdot k}}{n_{\cdot\cdot\cdot\cdot}}.$$

Combining the two sets of results gives

$$\begin{aligned}&\Pr(R = h, S = i, O = j, A = k)\\ &\quad = \Pr(O = j | R = h, S = i, A = k)\Pr(R = h, S = i)\Pr(A = k)\end{aligned}$$

or, equivalently,

$$p_{hijk} = \frac{p_{hijk}}{p_{hi\cdot k}} \left(p_{hi\cdot\cdot}\right) p_{\cdot\cdot\cdot k}\,.$$

This probability model appears to be a saturated log-linear model but is not. Taking logs gives $\log(p_{hijk})$ as the sum of four additive terms, one of which involves all four indices. This would seem to be a saturated log-linear model. However, a two-stage modeling procedure was used, so the simple-minded approach is not appropriate. Using the maximum likelihood estimates from each stage gives

$$\hat{p}_{hijk} = \frac{n_{hijk}}{n_{hi\cdot k}} \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}} \frac{n_{\cdot\cdot\cdot k}}{n_{\cdot\cdot\cdot\cdot}}$$

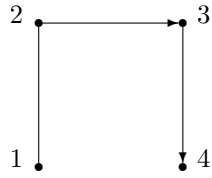
and estimated expected cell counts

$$\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot} \frac{n_{hijk}}{n_{hi\cdot k}} \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}} \frac{n_{\cdot\cdot\cdot k}}{n_{\cdot\cdot\cdot\cdot}}.$$

It is interesting to note that there is no data reduction involved in the estimation. An important, if often underemphasized, element of statistical modeling is that large amounts of data are reduced to manageable form by the use of sufficient statistics. This is certainly true for ANOVA type log-linear models where maximum likelihood estimates are completely determined by various *marginal* tables. The $\hat{m}_{hijk}$'s given above require knowledge of the n_{hijk}'s, so no data reduction has occurred. This is not always true; data reduction does occur for some recursive causal models.

We now consider recursive causal graphs with four factors, of which two are responses.

EXAMPLE 5.4.2. *Two Response Factors.*
Assume multinomial sampling for a table with four factors. Consider the recursive causal graph given below.



Factors 1 and 2 are purely explanatory and interact. Factor 3 has one direct cause, which is factor 2. Factor 4 has one direct cause, factor 3. In general, the probability model for four factors can be written in three terms:

$$\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k) = \Pr(F_1 = h, F_2 = i) \times \Pr(F_3 = j|F_1 = h, F_2 = i)\Pr(F_4 = k|F_1 = h, F_2 = i, F_3 = j).$$

Based on the graph, we write the recursive causal probability model as

$$\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k) = \Pr(F_1 = h, F_2 = i)\Pr(F_3 = j|F_2 = i)\Pr(F_4 = k|F_3 = j).$$

The first term, $\Pr(F_1 = h, F_2 = i)$, involves only the purely explanatory factors. The second term, $\Pr(F_3 = j|F_2 = i)$, involves the distribution for factor 3 given its direct cause, the purely explanatory factor 2. Note that factor 1 is an indirect cause of 3 because of its relationship with factor 2; however, only the direct cause F_2 is involved in the probability model. The third term, $\Pr(F_4 = k|F_3 = j)$, involves the distribution of F_4 given its direct cause F_3.

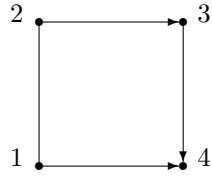
Again, estimation of probabilities is simple. The probability $\Pr(F_1 = h, F_2 = i)$ is estimated using the graphical log-linear model obtained by dropping the response factors 3 and 4 and all directed edges. In other words, it is estimated from the saturated model [12] for the marginal table involving only the first two factors. The term $\Pr(F_3 = j|F_2 = i)$ is estimated using the saturated model for the 2, 3 marginal table. The last term, $\Pr(F_4 = k|F_3 = j)$, is estimated from the saturated model for the 3, 4 marginal table. Thus,

$$\begin{aligned}\hat{p}_{hijk} &= \hat{p}_{hi\cdot\cdot}\frac{\hat{p}_{\cdot ij\cdot}}{\hat{p}_{\cdot i\cdot\cdot}}\frac{\hat{p}_{\cdot\cdot jk}}{\hat{p}_{\cdot\cdot j\cdot}} \\ &= \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}}\frac{n_{\cdot ij\cdot}}{n_{\cdot i\cdot\cdot}}\frac{n_{\cdot\cdot jk}}{n_{\cdot\cdot j\cdot}}.\end{aligned}$$

Of course, the estimated expected cell counts are simply

$$\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot}\hat{p}_{hijk}\,.$$

The graphical model can be made more interesting by inserting a direct cause between 1 and 4.



The recursive causal probability model is now

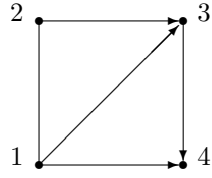
$$\begin{aligned}&\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k)\\ &\quad = \Pr(F_1 = h, F_2 = i)\Pr(F_3 = j|F_2 = i)\Pr(F_4 = k|F_1 = h, F_3 = j).\end{aligned}$$

Again, the first term $\Pr(F_1 = h, F_2 = i)$, involves only the purely explanatory factors. The second term, $\Pr(F_3 = j|F_2 = i)$, involves the distribution for factor 3 given its direct cause. The difference between this model and the previous one is that the third term, $\Pr(F_4 = k|F_1 = h, F_3 = j)$, now involves the distribution of F_4 given both of its direct causes F_1 and F_3. Estimation is based on saturated models for appropriate marginal tables and yields

$$\begin{aligned}\hat{p}_{hijk} &= \hat{p}_{hi\cdot\cdot}\frac{\hat{p}_{\cdot ij\cdot}}{\hat{p}_{\cdot i\cdot\cdot}}\frac{\hat{p}_{h\cdot jk}}{\hat{p}_{h\cdot j\cdot}}\\ &= \frac{n_{hi\cdot\cdot}}{n_{\cdots\cdot}}\frac{n_{\cdot ij\cdot}}{n_{\cdot i\cdot\cdot}}\frac{n_{h\cdot jk}}{n_{h\cdot j\cdot}}.\end{aligned}$$

Note that a saturated model is used for the explanatory factors only because the graph indicates use of a saturated model. Unlike response factors, the model for explanatory factors is not required to be a saturated model for the appropriate marginal table.

Consider one final graph.



The probability model is

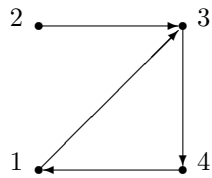
$$\begin{aligned}\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k) &= \Pr(F_1 = h, F_2 = i)\\ &\quad \times \Pr(F_3 = j|F_1 = h, F_2 = i)\Pr(F_4 = k|F_1 = h, F_3 = j).\end{aligned}$$

Again, the first term, $\Pr(F_1 = h, F_2 = i)$, involves only the purely explanatory factors. The second term, $\Pr(F_3 = j|F_1 = h, F_2 = i)$, involves the distribution for factor 3 given both of its direct causes. The third term, $\Pr(F_4 = k|F_1 = h, F_3 = j)$ also conditions only on direct causes. Estimation is based on appropriate marginal tables,

$$\hat{p}_{hijk} = \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}} \frac{n_{hij\cdot}}{n_{hi\cdot\cdot}} \frac{n_{h\cdot jk}}{n_{h\cdot j\cdot}}$$

with estimated expected cell counts $\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot}\hat{p}_{hijk}$.

In general, a *causal graph* for a set of factors C includes a set M of purely explanatory factors, also called external or *exogenous* factors and a set $C - M$ of response factors, also called internal or *endogenous* factors. The exogenous factors have an undirected graph associated with them. Each endogenous factor is the end point for one or more directed edges. The directed edges can originate at either exogenous or endogenous factors. A causal graph is *recursive* if no endogenous factor is a cause of itself; in other words, if there are no directed pathways that lead from an endogenous factor back to itself. For example, the causal graph



is not recursive. Factor 2 is exogenous. The other three factors are endogenous. There is a directed path from 3 to 4 to 1 and back to 3, so 3 is a cause of itself and the graph is not recursive. In this example, all of the endogenous factors are causes of themselves.

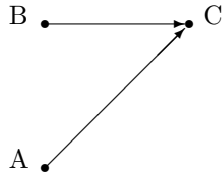
Generally, for a factor $F_i \in C - M$, its *direct causes* are the factors from which a directed edge goes to F_i. Let the set of all such factors be D_i. The probability for a recursive causal model is

$$\Pr(F_i = f_i \,:\, F_i \in C) = \Pr(F_i = f_i \,:\, F_i \in M) \prod_{F_i \in C-M} \Pr(F_i = f_i|D_i).$$

The term $\Pr(F_i = f_i \,:\, F_i \in M)$ depends on the marginal graph for M, i.e., the graph that drops all endogenous factors and directed edges. Estimates of $\Pr(F_i = f_i \,:\, F_i \in M)$ are maximum likelihood estimates from the corresponding graphical log-linear model. Estimation of a term of the form $\Pr(F_i = f_i|D_i)$ is based on the saturated model for the marginal table with factors in $\{F_i\} \cup D_i$.

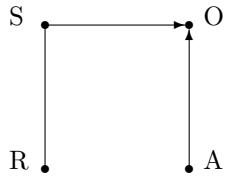
Putting arrowheads on edges and defining a new method for stating probability models has nothing fundamental to do with causation. These probability models can be used even when the causation suggested by the graph exists only in the head of the data analyst. Moreover, if enough models with nonsensical causations are fit, one that fits well may be found. Obviously, a well-fitting model does not establish that the causal patterns in the model are true. However, if the graph is a reasonable statement of a causal process, a well-fitting probability model adds credence to the hypothesized causal process. Evaluating how well recursive causal models fit is discussed later in this section.

A useful and interesting concept in recursive causal models is that of *configuration* $>$. It allows one to relate recursive causal models to decomposable log-linear models, cf. Section 2. Three factors A, B, and C are in configuration $>$ if C is caused by both A and B, but there is neither a directed nor an undirected edge between A and B. This is illustrated below.



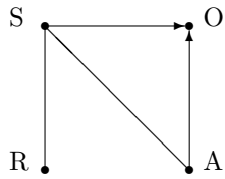
Recursive causal models are typically not log-linear, but there is a substantial intersection between the classes of models. The key result is that *a recursive causal graph contains no factors that are in configuration* $>$ *and the graph restricted to the exogenous variables is decomposable if and only if the recursive causal probability model is identical to the probability model determined by a decomposable log-linear model*, cf. Wermuth and Lauritzen (1983).

Example 5.4.3. Consider the recursive causal graph given below. This is similar to one used in Example 5.4.1; however, the factor R has been eliminated as a direct cause for O.



The factors S, A, O are in configuration >; thus, the recursive causal graph is not equivalent to a decomposable log-linear model.

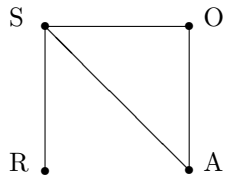
By connecting the nodes for S and A, the configuration > can be eliminated.



The probabilities for this recursive causal graph are the probabilities for the model [RS][SA] in the exogenous marginal table, obtained by collapsing over the response factor O, times the conditional probabilities for $\Pr(O = j|S = i, A = k)$ from the saturated model collapsing over R. This gives

$$p_{hijk} = \left(\frac{p_{hi\cdot\cdot}p_{\cdot i\cdot k}}{p_{\cdot i\cdot\cdot}}\right)\left(\frac{p_{\cdot ijk}}{p_{\cdot i\cdot k}}\right) = \left(\frac{p_{hi\cdot\cdot}p_{\cdot ijk}}{p_{\cdot i\cdot\cdot}}\right).$$

Note that these are exactly the same *probabilities* as determined by the decomposable log-linear model [RS][SOA] defined by the *underlying undirected graph*



In the underlying undirected graph, directed edges are changed to undirected edges. The probability models are the same, so independence relationships are the same. For the decomposable model, the independence relationship is that R is independent of O and A given S. This holds for both the log-linear model and the recursive causal model.

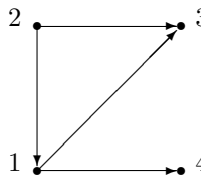
The probability models associated with decomposable log-linear models are identical to probability models for recursive causal models that (1) have no configurations > and (2) have a decomposable exogenous factor graph. The exogenous factor graph is the graph with all response factors and directed edges eliminated. It follows that decomposable models can be thought of as a subset both of graphical log-linear models and of recursive

causal models. A recursive causal model graph with a decomposable exogenous factor graph and no configurations $>$ can be transformed into a decomposable log-linear model graph simply by changing directed edges to undirected edges. See Wermuth and Lauritzen (1983) and Kiiveri, Speed, and Carlin (1984) for the validity of these statements.

EXERCISE 5.3. In Example 5.4.1, it was stated that the model [RSO][OA] is not a recursive causal model for Opinion. However, [RSO][OA] is decomposable, so it corresponds to some recursive causal model. Explain why [RSO][OA] is not a recursive causal model for Opinion, and by changing the endogenous factor, give a recursive causal model that does correspond to [RSO][OA].

We now present a conditional independence result that holds for general recursive causal models. *Any endogenous (response) factor F_i is independent of the factors F_j for which it is not a direct or indirect cause given D_i, the direct causes of F_i.* Independence among exogenous factors is determined by the exogenous factor graph.

EXAMPLE 5.4.4. In the model associated with the following graph, factor 4 is independent of the factors for which it is not a cause, i.e., factors 2 and 3, given factor 1 which is the direct cause of 4.

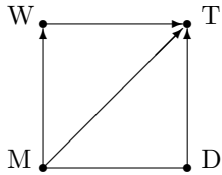


Another independence relation that can be read from the graph is that 3 is independent of 4 given 1 and 2, the direct causes of 3. Note that the graph contains no configurations $>$, so the independence relations are the same as in the underlying undirected graph. The decomposable model is [41][123] which has 4 independent of 2 and 3 given 1. The decomposable model also implies that 3 is independent of 4 given 1 and 2.

Wermuth and Lauritzen (1983) and Kiiveri, Speed, and Carlin (1984) examine the validity of conditional independence statements for recursive causal models. Another natural application of recursive causal graphs is in the analysis of structural equation models. Interpretations and conditional independence results hold as for the analysis of discrete data; see Kiiveri and Speed (1982).

Birch (1963), Goodman (1973), and Fienberg (1980) have examined methods of model selection that apply to recursive causal models. The methods are illustrated through an example.

EXAMPLE 5.4.5. Suppose muscle tension data similar to that of Examples 3.7.1 and 4.5.1 has been collected. The factors involved are T, the change in muscle tension, W, the weight of the muscle, M, the muscle type and D, the drug administered. Each factor is at two levels. For ease of exposition, we will treat the sampling as multinomial. The graph below indicates a possible recursive causal scheme.



Change in muscle tension T is hypothesized to have all of W, M, and D as direct causes. Muscle weight W has muscle type M as a direct cause. The purely explanatory factors are M and D; they are allowed to interact with each other. Write $\Pr(T = h, W = i, M = j, D = k) = p_{hijk}$. Note that the indexing has changed from previous examples. The first factor is now at the upper right of the graph rather than the lower left and the order is counterclockwise rather than clockwise. *These changes are important for verifying the maximum likelihood estimates presented.* The expected cell counts for the recursive causal model are

$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{..jk}}{n_{....}} \right) \left(\frac{n_{.ij.}}{n_{..j.}} \right) \left(\frac{n_{hijk}}{n_{.ijk}} \right).$$

The graph given above has one configuration > involving W, D, and T, so the probability model is not a decomposable log-linear model.

The lack of fit of the model can be tested in the usual way. Some algebra shows that

$$\begin{aligned} G^2 &= 2 \sum_{hijk} n_{hijk} \log \left(\frac{n_{hijk}}{\hat{m}_{hijk}} \right) \\ &= 2 \sum_{ijk} n_{.ijk} \log \left(\frac{n_{.ijk}}{n_{..jk} n_{.ij.} / n_{..j.}} \right). \end{aligned}$$

This is precisely the lack of fit statistic for testing the log-linear model [WM][MD] against the saturated model in the marginal table for W, M, and D. With each factor at two levels, it follows that the statistic has 2

degrees of freedom. In fact, the log-linear model [WM][MD] is consistent with the only conditional independence result available from the graph; the response factor W is independent of D given M, the direct cause of W.

The relationship of the lack of fit test with the log-linear model [WM][MD] can be seen through the probability modeling procedure. Recall that the basis of the modeling procedure is that one can always write

$$\begin{aligned}\Pr(T=h,W=i,M=j,D=k) &= \Pr(M=j,D=k)\\ &\quad\times\Pr(W=i|M=j,D=k)\Pr(T=h|W=i,M=j,D=k).\end{aligned}$$

and, based on the graph, the recursive causal probability model is

$$\begin{aligned}\Pr(T=h,W=i,M=j,D=k) &= \Pr(M=j,D=k)\\ &\quad\times\Pr(W=i|M=j)\Pr(T=h|W=i,M=j,D=k).\end{aligned}$$

The only real modeling being done is replacing $\Pr(W=i|M=j,D=k)$ with $\Pr(W=i|M=j)$. The factor T is extraneous. The perfectly general statement

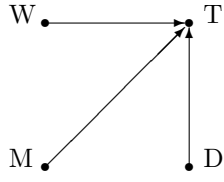
$$\Pr(W=i,M=j,D=k)=\Pr(M=j,D=k)\Pr(W=i|M=j,D=k)$$

has been replaced with

$$\begin{aligned}&\Pr(W=i,M=j,D=k)\\ &\quad=\quad\Pr(M=j,D=k)\Pr(W=i|M=j)\\ &\quad=\quad\Pr(M=j)\Pr(D=k|M=j)\Pr(W=i|M=j).\end{aligned}$$

This is just the model for conditional independence of W and D given M. It is not surprising that the test statistic involves only the aspect of the model that is not always true.

There is no particular reason to believe in a relationship between the type of drug used and the muscle type. It is also questionable whether muscle type really has an effect on weight. The graph that incorporates these ideas involves dropping one directed edge and the undirected edge. It is given below.



Note that by dropping M as a direct cause of W, W has been transformed into an exogenous factor. The estimated expected cell counts are

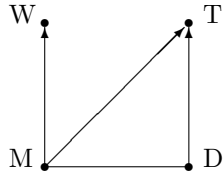
$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{\cdot i\cdot\cdot}}{n_{....}} \frac{n_{\cdot\cdot j\cdot}}{n_{....}} \frac{n_{\cdot\cdot\cdot k}}{n_{....}} \right) \left(\frac{n_{hijk}}{n_{\cdot ijk}} \right).$$

This model can be checked for general lack of fit as above. It is not difficult to see that the test is identical to that for complete independence [W][M][D] in the marginal table collapsing over T. The likelihood ratio chi-squared has 4 degrees of freedom.

This new model was obtained from the previous one by dropping edges in the previous graph; thus, this model is a reduced model relative to the previous one.

It follows that a likelihood ratio test can be performed for comparing the two models. Not surprisingly, the test simplifies to that of [W][M][D] versus [WM][MD].

Another possible model eliminates the effect of muscle weight W on tension.

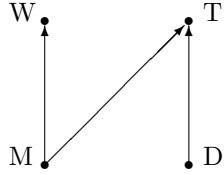


The maximum likelihood estimates are

$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{\cdot\cdot jk}}{n_{....}} \right) \left(\frac{n_{\cdot ij\cdot}}{n_{\cdot\cdot j\cdot}} \right) \left(\frac{n_{h\cdot jk}}{n_{\cdot\cdot jk}} \right).$$

The graph contains no configurations $>$, so the probability model and thus the maximum likelihood estimates are the same as for the decomposable log-linear model [WM][MTD] determined by the underlying undirected graph.

Consider one final model.



This incorporates independence of M and D from the exogenous factor graph and it has one conditional independence relation involving responses: W independent of T and D given M. These two relationships imply that the pair W and M are independent of D. The maximum likelihood estimates are

$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{..j.}}{n_{....}}\right) \left(\frac{n_{...k}}{n_{....}}\right) \left(\frac{n_{.ij.}}{n_{..j.}}\right) \left(\frac{n_{h.jk}}{n_{..jk}}\right).$$

The likelihood ratio test statistic for this model can be separated into the sum of three terms. The first term is the statistic for testing [M][D] versus [MD]. The second term is the statistic for testing [WM][MD] versus [WMD]. The last term is for testing [TMD][WMD] versus [TWMD]. The following series of equalities establishes the result.

$$\begin{aligned}
G^2 &= 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{hijk}}{\hat{m}_{hijk}}\right) \\
&= 2\sum_{hijk} n_{hijk} \left[\log(n_{hijk}) - \log(\hat{m}_{hijk})\right] \\
&= 2\sum_{hijk} n_{hijk} \log(n_{hijk}) - 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{..j.} n_{...k}}{n_{....}}\right) \\
&\quad - 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{.ij.}}{n_{..j.}}\right) - 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{h.jk}}{n_{..jk}}\right) \\
&= 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{hijk}\, n_{.ijk}}{n_{.ijk}\, n_{..jk}} n_{..jk}\right) \\
&\quad - 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{..j.} n_{...k}}{n_{....}}\right) \\
&\quad - 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{.ij.}}{n_{..j.}}\right) - 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{h.jk}}{n_{..jk}}\right) \\
&= 2\sum_{hijk} \left[n_{hijk} \log(n_{..jk}) - n_{hijk} \log\left(\frac{n_{..j.} n_{...k}}{n_{....}}\right)\right] \\
&\quad + 2\sum_{hijk} \left[n_{hijk} \log\left(\frac{n_{.ijk}}{n_{..jk}}\right) - n_{hijk} \log\left(\frac{n_{.ij.}}{n_{..j.}}\right)\right] \\
&\quad + 2\sum_{hijk} \left[n_{hijk} \log\left(\frac{n_{hijk}}{n_{.ijk}}\right) - n_{hijk} \log\left(\frac{n_{h.jk}}{n_{..jk}}\right)\right] \\
&= 2\sum_{hijk} n_{hijk} \left[\log(n_{..jk}) - \log\left(\frac{n_{..j.} n_{...k}}{n_{....}}\right)\right] \\
&\quad + 2\sum_{hijk} n_{hijk} \left[\log\left(n_{.ijk}\right) - \log\left(\frac{n_{.ij.} n_{..jk}}{n_{..j.}}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ 2\sum_{hijk} n_{hijk}\left[\log(n_{hijk}) - \log\left(\frac{n_{h\cdot jk}n_{\cdot ijk}}{n_{\cdot\cdot jk}}\right)\right] \\
= \quad & 2\sum_{jk} n_{\cdot\cdot jk}\left[\log\left(n_{\cdot\cdot jk}\right) - \log\left(\frac{n_{\cdot\cdot j\cdot}n_{\cdot\cdot\cdot k}}{n_{\cdot\cdot\cdot\cdot}}\right)\right] \\
&+ 2\sum_{ijk} n_{\cdot ijk}\left[\log\left(n_{\cdot ijk}\right) - \log\left(\frac{n_{\cdot\cdot jk}n_{\cdot ij\cdot}}{n_{\cdot\cdot j\cdot}}\right)\right] \\
&+ 2\sum_{hijk} n_{hijk}\left[\log(n_{hijk}) - \log\left(\frac{n_{\cdot ijk}n_{h\cdot jk}}{n_{\cdot\cdot jk}}\right)\right].
\end{aligned}
$$

The three terms in the last equality are precisely the three test statistics that were claimed. The existence of such breakdowns for G^2 is quite general.

5.5 Exercises

EXERCISE 5.5.1. Using the methods of Section 5.1, discuss the independence relationships for all of the models given below.

(a) [123][24][456]

(b) [12][13][23][24][456]

(c) [123][124][456]

(d) [123][24][456][15]

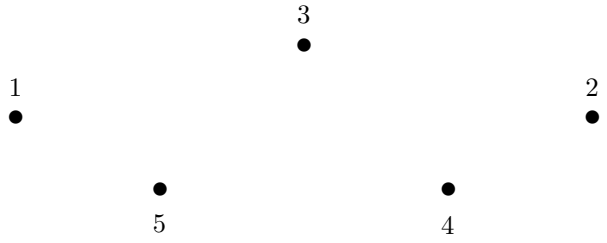
(e) [123][24][456][15][36]

EXERCISE 5.5.2. In the saturated log-linear model for a four-dimensional table, let $u_{34} = 0$ and let all of the corresponding higher-order terms also be zero, e.g., $u_{134} = 0$.

(a) Based on this model, find a formula for $\hat{m}_{hijk}$ without using graphical methods.

(b) Use graphical methods to find $\hat{m}_{hijk}$.

EXERCISE 5.5.3. The vertices for a five-factor model are given below. Connect the dots to give a graphical representation of the model [123][135][34][24]. Use the illustration to show that [123][135][34][24][25] is not a graphical model.



Exercise 5.5.4. Which of the models given below are graphical? Graph them. Which of these are decomposable? Discuss the independence relationships for all of the models. For each model, what marginal tables will provide valid inferences?

(a) [123][24][456]

(b) [12][13][23][24][456]

(c) [123][124][456]

(d) [123][24][456][15]

(e) [123][24][456][15][36]

Exercise 5.5.5. Consider all of the graphs in Example 5.4.2. Classify each as equivalent or not equivalent to a decomposable log-linear model. For those that are equivalent, prove the equivalence of the probability models.

6

Model Selection Methods and Model Evaluation

This chapter examines methods of selecting models for high-dimensional tables. The model selection methods considered are the stepwise methods, e.g., forward selection and backward elimination, a modified backward elimination method from Aitkin (1978, 1979) that controls the experimentwise error rate, and a backward elimination method from Wermuth (1976) that is restricted to decomposable models. In addition, we discuss the use of the model selection criteria presented in Section 3.6. Of course, it would be foolish to choose a model simply because some model selection procedure presents it to you as a good model. Other considerations such as model interpretability and the consistency of the data with model assumptions may dictate choosing some other model. It is always wise to *use model selection methods to produce several apparently good models* that can be investigated further. In line with this approach, the analysis of residuals and influential observations is also discussed in this chapter.

Modeling is a useful process both for prediction of future observables and for describing the relationships between factors. Large models always reproduce the data on which they were fitted better than smaller models. The saturated model always provides a perfect fit of the data. However, smaller models have more powerful interpretations and are often better predictive tools than large models. *Often, our goal is to find the smallest model that fits the data.*

In model fitting, there are two approaches to specifying models: the descriptive approach and the causative approach. The descriptive approach simply describes the relationships that are observed. For example, in a three-way table, one might find that given a young child's educational sta-

tus, her father's educational status and her mother's educational status are independent. This can be used to describe the data, but it would be foolish to suggest that the child's status in any way determines her parents' status. In analyzing these factors, it makes sense to consider the parents' status as fixed and to determine its effect on the child's status. However, a statistical relationship between parents' status and child's status does not imply causation. Causation cannot be inferred on statistical grounds; it must be inferred from the subject matter. For these reasons, we will concentrate on descriptive modeling. Nonetheless, causation has important implications for statistical modeling. Causation is closely related to the existence of response factors; the analysis of response factors is treated in Chapter 4, Section 5.4, and is also discussed in Chapter 11.

6.1 Stepwise Procedures for Model Selection

Stepwise procedures assume an initial model and then use rules for adding or deleting terms to arrive at a final model. Stepwise procedures are categorized in three ways: *forward selection*, in which terms are added to an initial small model; *backward elimination*, in which terms are removed from an initial large model; and *composite methods*, in which terms can either be added to or removed from the initial model. Methods for choosing initial models and examples will be considered in the subsequent two sections.

Because of the huge number of terms available in high-dimensional models, more effort is expended in selecting an initial model than is commonly used in regression analysis. Moreover, because ANOVA type models use parameters that are not uniquely defined, stepwise procedures must be adjusted so that nonsensical models are not considered.

Often, at any given point, stepwise procedures are applied only to examine terms that involve the same number of factors. For example, sometimes in examining three-factor interactions, two-factor interactions are ignored and any three-factor interactions that are implied by the existence of higher-order interactions are forced into the model. (In this case, higher-order interactions are those that involve four or more factors.) We begin by considering this particular procedure. Later, an improved method (implemented in BMDP) will be discussed.

Stepwise procedures are sequential in that they assume a current model and look to add or delete terms one at a time to that model. When considering s-factor terms, the basic forward selection rule is

FS: (a) Add the s-factor term not already in the model that has the most significant test statistic.

(b) Continue adding terms until no term achieves a predetermined minimum level of significance.

The basic backward elimination rule is

BE: (a) Delete the s-factor term with the least significant test statistic among s-factor terms that are not forced into the model.

(b) Continue until all terms maintain a predetermined minimum level of significance.

The backward elimination procedure is based on comparing models and does not consider whether the reduced models fit relative to the saturated model. It is possible that a model may fit globally and that dropping an s-factor term may be acceptable but that the new smaller model may not fit globally. One might want to modify the procedure so that it stops before eliminating any effect that will cause the saturated model test to be rejected.

Note that a term can be forced into the model either by the sampling scheme or by the presence in the model of a higher-order term that implies the existence of the term in question. For example, in the model [1235][234], when considering three-factor terms for elimination, all of [123], [125], [135], and [235] are forced into the model by having [1235] in the model. The only three-factor term eligible for elimination is [234].

The composite method alternates between applying the forward selection rule and the backward elimination rule. For forward selection, the test statistics are the statistics for testing the current model against the larger models in which one additional term has been added. For backward elimination, the test statistics are the statistics for testing the current model against the reduced models in which one term has been eliminated. "Significance" of test statistics is measured by their P *values*. A test statistic fails to achieve a predetermined minimum level of significance, say α, if $P > \alpha$ and maintains that level of significance if $P < \alpha$. The level α is often taken as .10, .05, or .01.

As an alternative to considering only s-factor terms, one can consider adding or deleting either simple or multiple effects. *Adding a simple effect consists of adding an effect that does not imply the simultaneous addition of any other effects.* For example, if we have four factors, say R, S, O, and A, and the model is [RSO][SA], the only simple effects that could be added are [RA] and [OA]. To see this, note two things. First, all other two-factor terms are already in the model, so these are the only two-factor terms that can be added. Second, to add any three-factor terms, e.g., [RSA], also implies the addition of a new two-factor term. Therefore, adding any three-factor term implies the addition of more than one effect and thus is not the addition of a simple effect.

If we consider the deletion of simple effects from [RSO][SA], the only possible deletions are the [RSO] and [SA] terms. Deleting the [RSO] term leaves the model [RS][RO][SO][SA]. Deleting the SA effect leaves [RSO][A].

Addition of a multiple effect involves incorporating a new factor into some effect that already exists in the model. For the model [RSO][SA], possible multiple effects are constructed by adding A to [RSO] (giving [RSOA][SA]), by adding R to [SA] (giving [RSO][RSA]), and by adding O to [SA] (giving [RSO][OSA]). In addition, the term [A] is implicitly in the model, so the terms [RA] and [OA] can be added, giving [RSO][SA][RA] and [RSO][SA][OA], respectively. Also, the term [RO] is implicitly in the model, so the term [ROA] can be added, yielding the model [RSO][SA][ROA].

In forward selection, considering either addition of simple effects or addition of multiple effects is appropriate. Because addition of multiple effects involves consideration of a wider variety of additional effects, addition of multiple effects is generally preferred. In backward elimination, deletion of simple effects is the appropriate procedure.

As defined here, deleting multiple effects is the same procedure as deleting simple effects. For the model [RSO][SA], deletion of any factor from [RSO] leaves all of the implicit terms [RS], [SO], and [RO] unaffected. Thus, deletion of multiple effects allows consideration of only the models [RS][SO][RO][OA] and [RSO][A]. These are the same models considered in the deletion of simple effects. There is a key difference between adding and deleting multiple effects. Adding multiple effects to [RSO][SA] allows addition of interesting nonsimple effects like [ROA] because [RO] is implicitly in the model. Deleting a factor from an implicit term such as [RO] has absolutely no effect on the model [RSO][SA]. Other definitions of what it means to delete a multiple effect are possible, but the ones I am acquainted with can give stupid results. To be safe, when using computer software, one should always specify deletion of simple effects.

One reasonable approach to backward elimination in, say, a five-factor model is to eliminate first the five-factor effect if possible, then any unnecessary four-factor effects. When eliminating three-factor effects, restrict attention only to those three-factor effects not forced into the model by the included four-factor effects. Similarly, only consider for elimination two-factor effects that are not forced in by the included three- and four-factor effects. Also, *any effects forced into the model by the sampling scheme should never be considered for elimination.*

As is well known from regression analysis, the main virtue of stepwise methods is that they are fast and cheap. Their virtue is directly related to their fault. They are fast and cheap because the procedures put severe limits on the number of models that are considered. Because only a limited number of models are examined, the procedures can easily miss the best models. Stepwise methods do not give the best model based on any overall criteria of model fit (cf. Section 6); in fact, they can give models that contain none of the terms that are in the best models. Forward selection

is a notoriously bad method of variable selection because it starts from an inadequate model and there is no guarantee that it will ever arrive at an adequate model. Backward elimination should give an adequate model if the initial model is adequate, but the only way to ensure an adequate initial model is to use the saturated model. Combined methods improve on forward selection simply because they allow consideration of more models. However, combined methods do not ensure finding the best models either. A nontechnical problem with stepwise procedures is that they give a unique "best" answer. Typically, no uniquely correct model exists. *If stepwise methods are to be used, it is wise to use several variations and therefore arrive at several candidate models. These models should be evaluated on their interpretability and their consistency with model assumptions to arrive at one or more final models.*

6.2 Initial Models for Selection Methods

In this section, we discuss a variety of methods for arriving at an appropriate initial model from which to begin the search for a well-fitting model. *The examples in this section deal only with initial model selection.* An example incorporating various stepwise procedures is given in Section 3. This section examines three approaches to picking an initial model: all s-factor effects models, models based on tests of marginal and partial association, and models based on testing each term in the saturated model last.

6.2.1 All s-Factor Effects

The simplest way to choose an initial model is to take one that consists of all effects of a particular level s. For example, the initial model can be the model of all main effects, or all two-factor effects, or all three-factor effects, and so on. The initial model can be chosen as either the smallest of these models that fits the data or the largest of these models that does not fit the data. In particular, for a four-factor table, one can test the models

(a) [1][2][3][4]

(b) [12][13][14][23][24][34]

(c) [123][124][234][134]

(d) [1234]

against each other to determine the smallest model that fits the data. Suppose it is model (b). We can then consider eliminating terms from model (b). Another approach is to look at the largest model that does not fit the data. That would be model (a). We can then consider selecting terms

to add to model (a). Elimination and selection can be performed either by ad hoc methods or by using the formal rules for backward elimination and forward selection.

Note that if we begin with model (b) [model (a)], it is very tempting to restrict attention to deleting (adding) only two-factor terms. Considering only two-factor terms is a substantial reduction in work as compared to investigating all levels of terms, but this simplification runs the risk of both missing some important terms and leaving in some unimportant terms.

Finally, it should be noted that if a combined stepwise procedure is to be used, either of the initial models is appropriate. However, different initial models may give different results.

EXAMPLE 6.2.1. Reconsider the data of Examples 3.7.1 and 4.5.1 on the relationship between two drugs and muscle tension. For each mouse, a muscle was identified and its tension was measured. A randomly chosen drug was given to the mouse and the muscle tension was measured again. The muscle was then tested to identify which type of muscle it was. The weight of the muscle was also measured. Factors and levels are tabulated below.

Factor	Abbreviation	Levels
Change in Muscle Tension	T	High, Low
Weight of Muscle	W	High, Low
Muscle	M	Type 1, Type 2
Drug	D	Drug 1, Drug 2

The sampling is product multinomial with the total count for each muscle type fixed. The data are

			Drug	
Tension	Weight	Muscle	Drug 1	Drug 2
	High	Type 1	3	21
		Type 2	23	11
High				
	Low	Type 1	22	32
		Type 2	4	12
	High	Type 1	3	10
		Type 2	41	21
Low				
	Low	Type 1	45	23
		Type 2	6	22

The test statistics are given below. Clearly, the only model that fits the data is the model of all three-factor interactions.

Model	*df*	G^2	*P*
[TWM][TWD][TMD][WMD]	1	0.11	.74
[TW][TM][WM][TD][WD][MD]	5	47.67	.00
[T][W][M][D]	11	127.4	.00

6.2.2 Examining Each Term Individually

An intuitively appealing method of selecting an initial model is to examine each term in the saturated model and include only those terms that are important. The question arises as to how one decides which terms are important. One reasonable approach is testing whether the terms are nonzero. One can then include the terms that are significantly different from zero and drop the rest. Unfortunately, the saturated model is overparametrized to the point that any terms except the u_{1234}'s can be dropped without affecting the model. The problem is in determining how to test whether the terms are zero. There are many possibilities. For example, to test whether $u_{123(hij)}$ is important in a four-dimensional table, we can test [12][13][234][134] versus [123][234][134] or we can test [234] versus [123][234] or any of a very large number of other model comparisons. The problem is to decide on which tests to examine.

Two methods have been proposed: testing each term last and the method suggested by Brown (1976) in which tests of marginal and partial association are performed for each term. These methods are examined in the next two subsections.

6.2.3 Tests of Marginal and Partial Association

Brown (1976) proposed looking at two tests for each term in the saturated model: a test of marginal association and a test of partial association. These tests can be used in a variety of ways to choose an initial model.

To test a particular term for *marginal association*, collapse over any factors not included in the term. The test of marginal association is based on the marginal table and consists of testing the model that involves only the term in question against the largest submodel that does not include the term. (Main effects are not tested for marginal association.)

EXAMPLE 6.2.2. The test of marginal association for the $u_{1234(hijk)}$'s is the test of [123][124][134][234] versus [1234]. The test of marginal association for the $u_{123(hij)}$'s is the test of [12][23][13] versus [123]. The test of marginal association for the $u_{24(ik)}$'s is the test of [2][4] versus [24].

The test for *partial association* depends on the number of factors involved in the term. If the term involves s factors, the test of partial association is a test of the model with all s-factor (interaction) terms against the reduced

model in which the term in question is dropped out. Thus, in a test of partial association, all other effects are fixed at a certain level of interaction.

EXAMPLE 6.2.3. In a four-dimensional table, the test of partial association for the $u_{123(hij)}$'s is the test of [124][134][234] versus [123][124][134][234]. The test of partial association for the $u_{24(ik)}$'s is the test of [12][13][14][23][34] versus [12][13][14][23][24][34]. In a four-dimensional table, the test of partial association for u_{1234} is identical to the test of marginal association.

Note that the degrees of freedom for the tests are the degrees of freedom for dropping the term in question; they are typically the same in the two tests.

There are a number of ways of choosing an initial model using Brown's tests: (a) include all terms with significant marginal tests, (b) include all terms with significant partial tests, (c) include all terms for which either the marginal or partial test is significant, (d) include all terms for which both the marginal and partial tests are significant.

Method (d) always gives the smallest model. Method (c) always gives the largest model. Method (d) can be used to determine an initial model for forward selection. Method (c) determines a model that might be used with backward elimination. Any of the four methods would give an appropriate initial model for combined stepwise selection.

An obvious ad hoc model selection approach is to restrict attention to models that are between the small model of method (d) and the large model of method (c). Perhaps the main fault with this method is that important terms could have been missed in model (c).

EXAMPLE 6.2.4. Brown's tests for the muscle tension data are presented in Table 6.1. The WMD term is clearly significant as is the WM term. In addition, several terms involving the change in muscle tension appear to be important, e.g., T, TM, TD, and possibly TMD.

Using significance levels of $\alpha = .01$ and $\alpha = .10$, the four initial models suggested by Brown's tests are

	$\alpha = .01$	$\alpha = .10$
Method (a):	[WMD][TMD]	[WMD][TMD][TWM]
Method (b):	[WMD][TD][TM]	[WMD][TMD]
Method (c):	[WMD][TMD]	[WMD][TMD][TWM]
Method (d):	[WMD][TD]	[WMD][TMD].

6.2.4 Testing Each Term Last

The basis of this method is testing whether each term can be dropped from the saturated model without a significant loss of explanatory power.

TABLE 6.1. Brown's Tests for the Muscle Tension Data

Effect	Partial Association G^2	P	Marginal Association G^2	P
T	6.04	.01	—	—
W	3.55	.06	—	—
M	1.18	.28	—	—
D	0.08	.78	—	—
TW	2.35	.13	0.06	.80
TM	6.81	.01	5.27	.02
WM	63.66	.00	62.25	.00
TD	6.02	.01	6.37	.01
WD	0.65	.42	1.12	.29
MD	0.17	.68	1.40	.24
TWM	1.00	.32	2.63	.10
TWD	0.01	.93	0.04	.85
TMD	2.86	.09	6.01	.01
WMD	35.65	.00	40.49	.00
TWMD	0.14	.70	0.14	.70

The problem with this method is that it requires a reparametrization of the model. For example, the model $\log(m_{hijk}) = u_{24(ik)} + u_{1234(hijk)}$ is a saturated model, but if we drop the $u_{24(ik)}$'s, we get $\log(m_{hijk}) = u_{1234(hijk)}$ which is still a saturated model. Dropping the $u_{24(ik)}$'s does not change the model. To test every term against the saturated model requires a regression parametrization in which dropping any term really reduces the model.

We begin with a simple example that assumes familiarity with estimation for analysis of variance under the "usual" constraints. After the example, we deal with the question of reparametrization. The discussion of reparametrization involves a more sophisticated use of linear model ideas than has been used thus far in the book.

EXAMPLE 6.2.5. Consider again the muscle-tension data of Example 6.2.1. This involves four factors each at two levels. We begin by examining a similar normal theory ANOVA model

$$\begin{aligned} y_{hijk} \quad = \quad & \mu + \alpha_h + \beta_i + \gamma_j + \eta_k \\ & + (\alpha\beta)_{hi} + (\alpha\gamma)_{hj} + (\alpha\eta)_{hk} \\ & + (\beta\gamma)_{ij} + (\beta\eta)_{ik} + (\gamma\eta)_{jk} \\ & + (\alpha\beta\gamma)_{hij} + (\alpha\beta\eta)_{hik} + (\alpha\gamma\eta)_{hjk} \\ & + (\beta\gamma\eta)_{ijk} + (\alpha\beta\gamma\eta)_{hijk} + e_{hijk} \, . \end{aligned}$$

With two levels in each factor, every interaction has one degree of freedom and corresponds to a contrast. For example, under the "usual" side conditions, the $(\alpha\beta)$ interaction contrast is

$$(\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22} \, .$$

The estimate of the contrast is

$$\bar{y}_{11\cdot\cdot} - \bar{y}_{12\cdot\cdot} - \bar{y}_{21\cdot\cdot} + \bar{y}_{22\cdot\cdot} \ .$$

Log-linear model estimation is analogous to the ANOVA procedure.

We are dealing with a saturated model

$$\log(m_{hijk}) = u + u_{1(h)} + \cdots + u_{234(ijk)} + u_{1234(hijk)},$$

so $\hat{m}_{hijk} = n_{hijk}$ for all h, i, j, and k. Define new parameters λ corresponding to each interaction contrast. For example, the u_{12} interaction corresponds to a contrast

$$4\lambda_{12} = u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)}$$

or, equivalently,

$$16\lambda_{12} = 4[u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)}].$$

This particular definition of the λ's relates to two things that will be examined later. One is ease of computation of the estimates; the other is a useful reparametrization of the saturated model. Let $w_{hijk} = \log(n_{hijk})$. The estimated contrast is

$$\begin{aligned} 16\hat{\lambda}_{12} &= 4[\bar{w}_{11\cdot\cdot} - \bar{w}_{12\cdot\cdot} - \bar{w}_{21\cdot\cdot} + \bar{w}_{22\cdot\cdot}] \\ &= [w_{11\cdot\cdot} - w_{12\cdot\cdot} - w_{21\cdot\cdot} + w_{22\cdot\cdot}] \ . \end{aligned}$$

Applying the usual side conditions, $u_{12(h\cdot)} = u_{12(\cdot i)} = 0$, leads to the parameter estimates

$$\frac{\hat{\lambda}_{12}}{4} = \hat{u}_{12(11)} = -\hat{u}_{12(12)} = -\hat{u}_{12(21)} = \hat{u}_{12(22)}.$$

Obviously, if you know one of the $\hat{u}_{12(hi)}$'s, you know them all. It is simpler to focus on $\hat{\lambda}_{12}$.

The estimates of all the $\hat{\lambda}$'s are 1/16th of the sums of 8 w_{hijk}'s minus the sum of the remaining 8 w_{hijk}'s, so all $\hat{\lambda}$'s have the same asymptotic standard error

$$\mathrm{SE}(\hat{\lambda}) = \left[\frac{1}{16}\sum_{hijk}\frac{1}{n_{hijk}}\right]^{1/2} \ .$$

The standard error depends on having a saturated model and is a generalization of the result for log odds ratios given earlier. Details are given in Section 10.2.

If we do an analysis of variance on the w_{hijk}'s, the sums of squares for various terms equal $16\hat{\lambda}^2$. Table 6.2 shows an analysis of variance. Table 6.3

TABLE 6.2. Analysis of Variance on $\log(n_{hijk})$ for the Muscle Tension Data

Source	*df*	SS
T	1	.2208
W	1	.3652
M	1	.0000
D	1	.9238
TW	1	.0522
TM	1	.4202
WM	1	5.631
TD	1	.1441
WD	1	.0080
MD	1	.2167
TWM	1	.1123
TWD	1	.0018
TMD	1	.2645
WMD	1	3.286
TWMD	1	.01188

gives values of $|16\hat{\lambda}|$ and $|z| = |\hat{\lambda}/\mathrm{SE}(\hat{\lambda})| = |16\hat{\lambda}/\mathrm{SE}(16\hat{\lambda})|$. The z values can be used to test $\lambda = 0$. The estimates in Table 6.3 were obtained from Table 6.2. For example, the source T has a sum of squares of .2208, so $|16\hat{\lambda}_T| = \sqrt{(16).2208}$. The standard error is $\mathrm{SE}(16\hat{\lambda}_T) = \sqrt{\sum(1/n_{hijk})}$. The main reason for using estimates of 16λ rather than λ is that the 16λ estimates are more comparable to another reparametrization of the saturated model that will be used later.

The important terms in Table 6.3 are λ_{WMD}, λ_{WM}, and perhaps λ_D. In other words, the main effect for D, the WM interaction, and the WMD interaction are the important terms in the model. By our rule for including lower-order terms, the inclusion of λ_{WMD} implies the model [WMD] which automatically includes both [WM] and [D]. It is interesting to note that the factor T does not appear in any important terms.

As mentioned at the beginning of the subsection, testing each term last requires that the saturated model be reparametrized into a regression model. A method is needed for relating the reparametrized results back to the original parametrization. This is most easily done when each factor is at only two levels. If each factor is at only two levels, there is one degree of freedom for each u term. Still, there are an infinite number of possible parametrizations. It is necessary to (arbitrarily) choose one.

EXAMPLE 6.2.6. Consider a $2 \times 2 \times 2 \times 2$ table. The model

$$\begin{aligned}\log(m_{hijk}) &= u + u_{1(h)} + u_{2(i)} + u_{3(j)} + u_{4(k)} \\ &\quad + u_{12(hi)} + u_{13(hj)} + u_{14(hk)} + u_{23(ij)} + u_{24(ik)} + u_{34(jk)}\end{aligned} \tag{1}$$

TABLE 6.3. Muscle Tension Data: Estimates and Test Statistics for Model (2)

λ	$\lvert 16\hat{\lambda}\rvert$	$\lvert z\rvert$
T	1.880	1.44
W	2.417	1.85
M	0.002	0.00
D	3.846	2.94
TW	0.914	0.70
TM	2.593	1.98
WM	9.492	7.26
TD	1.518	1.16
WD	0.358	0.27
MD	1.862	1.42
TWM	1.340	1.03
TWD	0.172	0.13
TMD	2.057	1.57
WMD	7.251	5.55
TWMD	0.436	0.33

$\mathrm{SE}(16\hat{\lambda}) = 1.307$.

$$+ u_{123(hij)} + u_{124(hik)} + u_{134(hjk)} + u_{234(ijk)} + u_{1234(hijk)}$$

can be reparametrized as

$$\begin{aligned}\log(m_{hijk}) &= \lambda + (-1)^{h-1}\lambda_1 + (-1)^{i-1}\lambda_2 + (-1)^{j-1}\lambda_3 + (-1)^{k-1}\lambda_4 \\ &\quad + (-1)^{h+i-2}\lambda_{12} + (-1)^{h+j-2}\lambda_{13} + (-1)^{h+k-2}\lambda_{14} \\ &\quad + (-1)^{i+j-2}\lambda_{23} + (-1)^{i+k-2}\lambda_{24} + (-1)^{j+k-2}\lambda_{34} \\ &\quad + (-1)^{h+i+j-3}\lambda_{123} + (-1)^{h+i+k-3}\lambda_{124} \\ &\quad + (-1)^{h+j+k-3}\lambda_{134} + (-1)^{i+j+k-3}\lambda_{234} \\ &\quad + (-1)^{h+i+j+k-4}\lambda_{1234}\,.\end{aligned} \tag{2}$$

This parametrization gives the same estimates as using the "usual" side conditions, i.e., $0 = u_{1(\cdot)} = u_{2(\cdot)} = u_{3(\cdot)} = u_{4(\cdot)} = u_{12(\cdot i)} = u_{12(h\cdot)} = \cdots = u_{1234(\cdot ijk)} = u_{1234(h\cdot jk)} = u_{1234(hi\cdot k)} = u_{1234(hij\cdot)}$. Model (2) is given in matrix form in Example 10.4.1.

Other sets of side conditions correspond to other reparametrizations. For example, another frequently used set of side conditions are $0 = u_{1(1)} = u_{2(1)} = u_{3(1)} = u_{4(1)} = u_{12(1i)} = u_{12(h1)} = \cdots = u_{1234(1ijk)} = u_{1234(h1jk)} = u_{1234(hi1k)} = u_{1234(hij1)}$. Here, all u terms are set equal to zero for which any of h, i, j, or k is 1. If we let $\delta_{ab} = 1$ when $a = b$ and 0 otherwise where a and b are any symbols, these side conditions correspond to the

reparametrized model

$$\begin{aligned}\log(m_{hijk}) &= \gamma + \delta_{h2}\gamma_1 + \delta_{i2}\gamma_2 + \delta_{j2}\gamma_3 + \delta_{k2}\gamma_4 \\ &\quad + \delta_{(h,i)(2,2)}\gamma_{12} + \delta_{(h,j)(2,2)}\gamma_{13} + \delta_{(h,k)(2,2)}\gamma_{14} \\ &\quad + \delta_{(i,j)(2,2)}\gamma_{23} + \delta_{(i,k)(2,2)}\gamma_{24} + \delta_{(j,k)(2,2)}\gamma_{34} \\ &\quad + \delta_{(h,i,j)(2,2,2)}\gamma_{123} + \delta_{(h,i,k)(2,2,2)}\gamma_{124} \\ &\quad + \delta_{(h,j,k)(2,2,2)}\gamma_{134} + \delta_{(i,j,k)(2,2,2)}\gamma_{234} \\ &\quad + \delta_{(h,i,j,k)(2,2,2,2)}\gamma_{1234}\,. \end{aligned} \tag{3}$$

Except for λ_{1234} and γ_{1234}, these parametrizations are *not* equivalent and can lead to different conclusions about which terms should be in a model.

As mentioned earlier, a primary difficulty in testing each term last is in relating the tests for the reparametrized model to tests for ANOVA type models. In the special case where each factor has two levels (categories), the relationship is simple, because each term in the ANOVA type models has one degree of freedom, just as each test in the reparametrized model has one degree of freedom. If a particular term has a large test statistic, the corresponding main effect or interaction is included in the model. For example, if we reject $H_0 : \lambda_{12} = 0$ (or $H_0 : \gamma_{12} = 0$), then our ANOVA model includes $u_{12(hi)}$. This implies that the ANOVA model will include (at least implicitly) $u_{1(h)}$ and $u_{2(i)}$ regardless of whether λ_1 (γ_1) and λ_2 (γ_2) are significantly different from zero. Note that because λ_{12} and γ_{12} are not equivalent, the results of this procedure depend on the parametrization chosen.

In fact, identifying important effects by testing all effects last does not provide a good end model. It provides an initial model from which some method of exploration (e.g., forward selection or combined stepwise) can be used to determine a final model.

Example 6.2.7. Consider again the muscle tension data of Example 6.2.1. The λ values defined in Example 6.2.5 using the usual side conditions are exactly the same as the λ values defined in model (2). For example,

$$4\lambda_{12} = u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)}.$$

We have already obtained estimates of the λ's, standard errors, and $|z|$ scores.

Similarly, if we use the parametrization of model (3) and the related side conditions, we find, for example, that

$$\gamma_{12} = u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)};$$

however, $4\lambda_{12} \neq \gamma_{12}$. Some computer programs, e.g., GLIM, routinely provide estimates and standard errors for the parametrization of model (3). The results along with $|z|$ values are reported in Table 6.4. There are now

at least six interesting terms: γ_{WMD}, γ_{MD}, γ_{WM}, γ_D, γ_M, and γ_W. The $|z|$ value for γ_{WD} is also quite large. Again using the rule of including lower-order terms, the inclusion of γ_{WMD} implies the model [WMD], which, in turn, implies the inclusion of all of the other interesting terms. Once again, factor T does not appear.

The results from these two parametrizations are reasonably consistent for this data set, but it is not difficult to see how the analysis could go awry. Consider the terms for TWM and TMD. Using model (2), we get $|z(\hat{\lambda}_{TWM})| = 1.03$ and $|z(\hat{\lambda}_{TMD})| = 1.57$, so, although neither is significant, the TMD term seems considerably more important than the TWM term. However, in model (3), $|z(\hat{\gamma}_{TWM})| = 0.80$ and $|z(\hat{\gamma}_{TMD})| = 0.80$. In model (3), the TMD and TWM terms seem to be of equal importance. The problem is that the parameters are model dependent. Because they are from different models, $8\lambda_{TWM} \neq \gamma_{TWM}$. As a result, the test statistics are testing different things. The only exception to this is that $16\lambda_{TWMD} = \gamma_{TWMD}$.

The relationships between parameters in models (1), (2), and (3) are complex and, in the our view, often not worth pursuing. The simplest way to avoid the complexities of alternative parametrizations is to deal directly with model (1) and its submodels. In our view, the method of testing each term last can give some rough ideas about the analysis, but usually should not be considered to give anything more than *rough* ideas.

Again, we note a rather curious phenomenon in this example. The experiment was conducted to investigate changes in muscle tension. Neither parametrization shows the significance of any effect involving T, the change in tension. It is theoretically possible that none of the other factors relate to change in muscle tension, but in most studies of this type, the investigator conducts the experiment because he or she knows that there are relationships between the other factors and change in tension. As we saw from the tests of partial and marginal association, such relationships exist. The method employed has simply failed to find them.

Two final comments on the choice of an initial model. *Any effects that are forced into the model to deal with the sampling scheme should be included in any initial model and never deleted in any model selection method.* Also, one is rarely interested in models that are smaller than the model of complete independence. It is common practice to include at least the main effect for every factor in an initial model.

6.3 Example of Stepwise Methods

We now give detailed examples of forward selection and backward elimination. Reconsider the data of Example 3.7.2 in which there are four factors defining a $2 \times 2 \times 3 \times 6$ table. Recall that the factors are

TABLE 6.4. Muscle Tension Data — Model (3): Estimates, Standard Errors, and Test Statistics

γ	$\hat{\gamma}$	SE	$\lvert z\rvert$
T	−0.000	.8165	0.00
W	1.992	.6154	3.24
M	2.037	.6138	3.32
D	1.946	.6172	3.15
TW	0.716	.8569	0.84
TM	0.578	.8570	0.67
WM	−3.742	.8199	4.56
TD	−0.742	.9024	0.82
WD	−1.571	.6765	2.32
MD	−2.684	.7179	3.74
TWM	−0.888	1.104	0.80
TWD	−0.304	.9781	0.31
TMD	0.810	1.010	0.80
WMD	3.407	.9620	3.54
TWMD	0.436	1.307	0.33

Factor	Abbreviation	Levels
Race	R	White, Nonwhite
Sex	S	Male, Female
Opinion	O	Yes = Supports Legalized Abortion
		No = Opposed to Legalized Abortion
		Und = Undecided
Age	A	18-25, 26-35, 36-45, 46-55, 56-65, 66+ years

The data are repeated in Table 6.5.

We begin by fitting the all three-factor model, the all two-factor model, and the complete independence (all one-factor) model.

Model	*df*	G^2
[RSO][RSA][ROA][SOA]	10	6.12
[RS][RO][RA][SO][SA][OA]	37	26.09
[R][S][O][A]	62	121.47

Clearly, both the all three-factor and the all two-factor models fit the data relative to the saturated model. Comparing the all three-factor and all two-factor models gives

$$\begin{aligned} G^2 &= 26.09 - 6.12 = 19.97, \\ df &= 37 - 10 = 27, \end{aligned}$$

so there is no reason to reject the all two-factor model. The model of complete independence does not fit.

TABLE 6.5. Abortion Opinion Data

Race	Sex	Opinion	Age					
			18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Und	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Und	1	4	9	9	10	16
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Und	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Und	1	2	1	1	1	1

6.3.1 Forward Selection

First, we consider forward selection using the model of complete independence as our initial model. As a criterion for adding terms, we add the most significant term as long as the significance level is below .10. For those of us who are not wild about significance tests, some comments based directly on the likelihood ratio test statistics are also included. It should be remembered that the formal method of forward selection is based on the significance levels.

Three methods of forward selection have been discussed. These involve adding two-factor terms, adding simple effects, and adding multiple effects. Starting with the model of complete independence, the first step in all three methods is the same. We require the following fits:

Model	Added Term	*df*	G^2
[RS][O][A]	RS	61	119.45
[RO][S][A]	RO	60	107.48
[RA][S][O]	RA	57	115.46
[SO][R][A]	SO	60	112.28
[SA][R][O]	SA	57	120.75
[OA][R][S]	OA	52	59.78
[R][S][O][A]		62	121.47

All of the models with two-factor terms are compared to the model of complete independence; e.g., to test the RS term, $G^2 = G^2([R][S][O][A]) - G^2([RS][O][A]) = 121.47 - 119.45 = 2.02$. The degrees of freedom are $62 - 61 = 1$. The results of the tests are summarized below.

Term	*df*	G^2	*P*
RS	1	2.02	.1557
RO	2	13.99	.0009
RA	5	6.01	.3055
SO	2	9.19	.0101
SA	5	0.72	.9820
OA	10	61.69	.0000

The term OA has the smallest *P value*, so [OA] is added to the model of complete independence.

With the new model [R][S][OA], the second step of adding either two-factor effects or simple effects again remains the same. At this point, adding any three-factor term would imply adding more than one effect, so simple effects are only two-factor effects. Consideration of the addition of multiple effects leads to a different second step.

To add either a two-factor effect or a simple effect requires the following fits:

Model	Added Term	*df*	G^2
[RS][OA]	RS	51	57.76
[RO][OA][S]	RO	50	45.79
[RA][OA][S]	RA	47	53.77
[SO][OA][R]	SO	50	50.59
[SA][OA][R]	SA	47	59.06
[R][S][OA]		52	59.78

Again, all the models with an additional two-factor term are compared to the model [R][S][OA].

Term	*df*	G^2	*P*
RS	1	2.02	.1557
RO	2	13.99	.0009
RA	5	6.01	.3055
SO	2	9.19	.0101
SA	5	0.72	.9820

The term RO is added to the model, giving a base model of [RO][OA][S].

If addition of multiple effects to [R][S][OA] is considered, two more models must be evaluated. Adding an additional factor to the OA term leads to the models [R][SOA] and [S][ROA]. These models have the fits

Model	Added Term	*df*	G^2
[R][SOA]	SOA	35	47.91
[S][ROA]	ROA	35	32.31
[R][S][OA]		52	59.78

They are tested against [R][S][OA].

Term	*df*	G^2	*P*
SOA	17	11.87	.8082
ROA	17	27.47	.0516

These *P values* are larger than the *P value* for RO, so the term RO is still added to [R][S][OA], giving the new model [RO][OA][S].

In this example, addition of two-factor effects and addition of simple effects turn out to be identical procedures. We now follow this procedure to its conclusion. After establishing the end model, we will indicate how these procedures would have differed if our model selection procedure had been modified slightly. Finally, we will follow the method of addition of multiple effects to its conclusion.

After the second step of forward selection, the simple effects and two-factor effects procedures had arrived at a base model of [RO][OA][S]. The only simple effects that can be added are two-factor effects. There are four possible effects that can be added. The necessary model fits are

Model	Added Term	*df*	G^2
[RS][RO][OA]	RS	49	43.77
[RA][RO][OA][S]	RA	45	38.82
[SO][RO][OA]	SO	48	36.60
[SA][RO][OA]	SA	45	45.07
[RO][OA][S]		50	45.79

This leads to the following differences:

Term	*df*	G^2	*P*
RS	1	2.02	.1557
RA	5	6.97	.2228
SO	2	9.19	.0101
SA	5	0.72	.9820

Thus, [SO] is added to the model. This turns out to be the last step at which anything is added to the model. The next step examines

Model	Added Term	*df*	G^2
[RS][SO][RO][OA]	RS	47	34.25
[RA][SO][RO][OA]	RA	43	29.63
[SA][SO][RO][OA]	SA	43	35.33
[SO][RO][OA]		48	36.60

and

Term	*df*	G^2	*P*
RS	1	2.35	.1252
RA	5	6.97	.2228
SA	5	1.27	.9382

At this stage, all of the *P values* are in excess of .10, so no new term is added and the final model is [SO][RO][OA].

To see how adding two-factor effects can differ from adding simple effects, suppose that our criterion for stepping is having *P values* less than .15. With this criterion, the term RS would be added to the model, giving a new model of [RS][SO][RO][OA]. If we restrict ourselves to adding two-factor effects, the next step involves adding RA or SA. If we allow addition of simple effects, the three-factor term RSO could also be added. This is the first time that a simple effect is a three-factor effect because [RS][SO][RO][OA] is the first model that contains all three of the two-factor effects that correspond to a three-factor effect. In particular, with [RS], [SO], and [RO] in the model, adding [RSO] is adding a simple effect.

Recall that the rationale for starting our search with the model of complete independence was based in part on the fact that the all two-factor model gave an adequate fit. This provides a rationale for considering only the addition of two-factor terms. Unfortunately, the test of the all two-factor model against the all three-factor model can have very little power for identifying individual three-factor terms that are important. It is not safe to ignore all three-factor terms based on this one test. Thus, it is dangerous to consider adding only two-factor effects. By considering addition of simple effects, we at least admit the possibility of examining important three-factor effects.

We now return to examining forward selection with the addition of multiple effects. Recall that after the second step, we had arrived at a base model of [RO][OA][S]. The multiple effects that can be added involve adding one new factor to a term already in the model, so these are the remaining two-factor effects (which are also simple effects) plus RSO, ROA, and SOA. Given below are statistics for fitting each model plus the differences in *df*'s and G^2's between the various models and [RO][OA][S]. Tests are based on these differences. We use a cutoff of $\alpha = .05$. (For this example, the first two steps do not change when $\alpha = .05$ is used instead of $\alpha = .10$.)

Model	Added Term	df	G^2	Differences df	Differences G^2	Differences P
[OR][OA][SA]	SA	45	45.07	5	0.72	.9820
[RA][OR][OA][S]	RA	45	38.82	5	6.97	.2228
[RO][OA][OS]	SO	48	36.60	2	9.19	.0101
[RO][OA][SR]	RS	49	43.77	1	2.02	.1557
[RO][SOA]	SOA	33	33.92	17	11.87	.8082
[ROA][S]	ROA	35	32.31	15	13.48	.5654
[RSO][OA]	RSO	45	24.77	5	21.02	.0008
[RO][OA][S]	—	50	45.79	—	—	—

For testing against [RO][OA][S], the model with the smallest P *value* is [RSO][OA], with $P = .0008$. The P *value* is less than .05, so we take [RSO][OA] as our working model. Multiple effects that can be added to this are SA, RA, SOA, ROA, RSA, and RSOA. The statistics are given below.

Model	Added Term	df	G^2	Differences df	Differences G^2	Differences P
[RSOA]	RSOA	0	0.00	45	24.77	.9938
[RSO][OA][SA]	SA	40	23.50	5	1.27	.9382
[RSO][OA][RA]	RA	40	17.79	5	6.97	.2228
[RSO][SOA]	SOA	30	22.09	15	2.68	.9998
[RSO][ROA]	ROA	30	11.29	15	13.48	.5655
[RSO][OA][RSA]	RSA	30	14.43	15	10.34	.7980
[RSO][OA]	—	45	24.77	—	—	—

For testing against the model [RSO][OA], every model has a P *value* greater than .05, so no new terms are added. The final model is [RSO][OA].

Because the addition of multiple effects leads to considering more models than the addition of simple effects, the author prefers the multiple effect option if you insist on doing forward selection.

6.3.2 Backward Elimination

We now consider applying backward elimination to the initial model containing all two-factor terms. We will use a cutoff value of $\alpha = .05$. Given below are statistics for the model of all two-factor terms and the six models in which one of the two-factor terms has been dropped. (At this stage, the simple effects are precisely the two-factor effects.) The df and G^2 for testing each model against the saturated model are given. With this information, each reduced model can be tested against the all two-factor model by taking differences in the df's and G^2's. For each of the reduced models, the differences are listed along with the P *value* for the test.

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AS][AR][AO][OS][OR][SR]	—	37	26.09	—	—	—
[AS][AR][OS][OR][SR]	AO	47	89.24	10	63.15	.0000
[AR][AO][OS][OR][SR]	AS	42	27.28	5	1.19	.9461
[AS][AO][OS][OR][SR]	AR	42	32.98	5	6.89	.2289
[AS][AR][AO][OR][SR]	OS	39	36.12	2	10.03	.0067
[AS][AR][AO][OS][SR]	OR	39	41.33	2	15.24	.0005
[AS][AR][AO][OS][OR]	SR	38	28.36	1	2.27	.1319

Deleting the AS term gives the largest *P value*, so we choose the reduced model [AR][AO][OS][OR][SR].

Once again, the simple effects are the two-factor effects. The necessary statistics are

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AR][AO][OS][OR][SR]	—	42	27.28	—	—	—
[AR][OS][OR][SR]	AO	52	89.93	10	62.65	.0000
[AO][OS][OR][SR]	AR	47	34.25	5	6.97	.2228
[AR][AO][OR][SR]	OS	44	36.80	2	9.52	.0086
[AR][AO][OS][SR]	OR	44	42.53	2	15.26	.0005
[AR][AO][OS][OR]	SR	43	29.63	1	2.35	.1252

Deleting the AR term gives the largest *P value*, so the new model is [AO][OS][OR][SR].

Simple effects are still two-factor effects, so the necessary statistics are

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AO][OS][OR][SR]	—	47	34.25	—	—	—
[A][OS][OR][SR]	AO	57	95.94	10	61.69	.0000
[AO][OR][SR]	OS	49	43.77	2	9.52	.0085
[AO][OS][SR]	OR	49	48.57	2	14.32	.0008
[AO][OS][OR]	SR	48	36.60	1	2.35	.1252

We now delete SR and use the base model [AO][OS][OR]. The statistics are

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AO][OS][OR]	—	48	36.60	—	—	—
[A][OS][OR]	AO	58	98.29	10	61.69	.0000
[AO][OR][S]	OS	50	45.79	2	9.19	.0101
[AO][OS][R]	OR	50	50.59	2	13.99	.0009

None of the *P values* is greater than .05, so we stop deleting terms and go with the model [AO][OS][OR]. Note that this model has the nice interpretation that given people's opinions; race, sex, and age are independent.

In this example, simple effects were always two-factor effects. If our cutoff level had been $\alpha = .01$, this would not have happened. With a cutoff of .01, the term OS can be dropped from [AO][OS][OR], yielding the model [AO][OR][S]. Deleting two-factor terms leads us to consider the reduced models [A][OR][S] (eliminating AO) and [AO][R][S] (eliminating OR). If we consider dropping simple effects, we would also consider the reduced model [AO][OR] (eliminating S). However, as mentioned earlier, it is typically not a good idea to drop main effects. If the initial model was the model of all three-factor effects, the difference between dropping three-factor effects and dropping simple effects could be substantial. Dropping simple effects is a more general procedure and seems more reasonable to the author.

6.3.3 Comparison of Stepwise Methods

Forward selection with $\alpha = .10$ and addition of simple effects lead to the model [RO][SO][OA]. It is easily seen that $\alpha = .05$ would lead to the same model. Forward selection with multiple effects and $\alpha = .05$ leads to [RSO][OA]. Backward elimination of simple effects from the all two-factor model with $\alpha = .05$ leads to [RO][SO][OA]. Frankly, we are lucky to have two methods give the same model. There is no reason that this needs to happen. Which is a better model? One is a special case of the other, so the test statistic is $36.60 - 24.77 = 11.83$ with $48 - 45 = 3$ degrees of freedom. This suggests quite strongly that [RSO][OA] is the better model. Note that it will not always be possible to test the results of different procedures because the models may not be comparable.

Stepwise methods are very sensitive to the cutoff values used. They are also very sensitive to the initial model. For backward elimination, we started with the model of all two-factor effects. The importance of the single three-factor effect RSO was washed out in testing the all two-factor model against the all three-factor model. Hence, it was decided to go with the two-factor model. If we had considered Brown's measures of partial and marginal association, we would have been better off (at least for these data). Brown's measures are given in Table 6.6. The term [RSO] stands out as a clearly important effect. In fact, even without using a stepwise procedure, Brown's tests clearly suggest the model [RSO][OA], but that is a function of these particular data.

In this section, we have used several variations on stepwise regression. This is not just a pedagogical device. *If stepwise methods are to be used in spite of their well-known weaknesses, it is important to use several variations.* This allows the data to indicate several candidate models. *These models should be compared to see how well they fit the model assumptions. They should also be compared for interpretability.*

TABLE 6.6. Brown's Measures of Association

Effect	*df*	Partial G^2	*P*	Marginal G^2	*P*
R	1	1552.90	.00	—	—
S	1	25.21	.00	—	—
O	2	1532.82	.00	—	—
A	5	55.14	.00	—	—
RS	1	2.27	.13	2.02	.16
RO	2	15.24	.00	13.99	.00
SO	2	10.03	.01	9.19	.01
RA	5	6.89	.23	6.01	.31
SA	5	1.19	.95	0.72	.98
OA	10	63.15	.00	61.69	.00
RSO	2	10.51	.01	9.48	.01
RSA	5	2.55	.77	1.70	.89
ROA	10	7.17	.71	6.51	.77
SOA	10	1.43	1.00	1.41	1.00
RSOA	10	6.12	.81	6.12	.81

6.3.4 Computer Commands

BMDP-4F performs these stepwise procedures and gives initial models, including measures of partial and marginal association. Commands for backward elimination of simple effects are

```
/ INPUT     FILE = 'C:\LOGLIN\ABORT.DAT'.
         FORMAT = FREE.
         VARIABLES = 5.
/ VARIABLE  NAMES = R, S, A, O, N.
/ TABLE     INDEX = R, S, A, O.
            COUNT = N.
/ STAT      ALL.
/ FIT       MODEL = AS, AR, AO, OS, OR, SR.
            ASSOCIATION = 4.
            DELETE = SIMPLE.
            STEP = 10.
            PROB = .05.
/ PRINT     LINE = 79.
/ END
```

The "step" command specifies how many steps are allowed in the procedure. The "prob" command specifies the probability for stopping the procedure. With this program, *both* the *P value* for the individual term and the *P value* for testing the model against the saturated model must be less than the specified probability for the procedure to stop.

6.4 Aitkin's Method of Backward Selection

Aitkin (1978, 1979) suggests a model selection method that is closely related to the *all s-factor effects* method described in Section 2. After using backward selection to pick an all s-factor model, Aitkin's method provides for testing every model intermediate between the all s-factor model and the all $s-1$-factor model. The procedure also incorporates ideas on *simultaneous testing* that control the overall error rate for all tests performed.

Aitkin begins by testing the all $s-1$-factor model against the all s-factor model at a level, say, γ_s. (The choice of γ_s will be discussed later.) This is actually a test of whether the s-factor effects are needed in the model. Except for the choice of γ_s, this is exactly what was done in the subsection of Section 2 on All s-Factor Effects.

To describe the procedure precisely, we need some additional notation. Let G_s^2 be the likelihood ratio test statistic and let d_s be the degrees of freedom for testing the all s-factor model against the saturated model. To test the need for s-factor effects, we reject the null hypothesis of no s-factor effects if

$$G_{s-1}^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s) \; .$$

This is a test for the adequacy of the all $s-1$-factor model. Aitkin then identifies the smallest value of s for which the all s-factor effects model adequately fits the data. This model has s as the largest value such that $G_{s-1}^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s)$.

EXAMPLE 6.4.1. Consider the muscle tension data of Example 6.2.1. This is a four-factor table. As given in Example 6.2.1, the all s-factor models are

s	Model	d_s	G_s^2
4	[TWMD]	0	0
3	[TWM][TWD][TMD][WMD]	1	0.11
2	[TW][TM][WM][TD][WD][MD]	5	47.67
1	[T][W][M][D]	11	127.4

For reasons to be considered later, suppose $\gamma_4 = .05$, $\gamma_3 = .185$, and $\gamma_2 = .265$, then Aitkin's tests are

$s-1$ versus s	$G_{s-1}^2 - G_s^2$	$\chi^2(1-\gamma_s, d_{s-1} - d_s)$
3 versus 4	$0.11 - 0 = .11$	$\chi^2(.95, 1) = 3.841$
2 versus 3	$47.67 - 0.11 = 47.56$	$\chi^2(.815, 4) = 6.178$
1 versus 2	$127.4 - 47.67 = 79.7$	$\chi^2(.735, 6) = 7.638$

The largest value of s for which $G_{s-1}^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s)$ is $s = 3$. The model [TWM][TWD][TMD][WMD] fits the data according to Aitkin's criteria.

Before discussing Aitkin's method in general, we examine of its application to the race, sex, opinion, age data.

EXAMPLE 6.4.2. For the race, sex, opinion, age data of Section 3, take $\gamma_s = .10$ for all s. Recall from the previous section that the model of all two-factor effects fits well. Because of this, the procedure will never consider three-factor effects. In particular, it will never consider the model [RSO][OA] which fits very well.

The model of all two-factor effects has 37 degrees of freedom and $G^2 = 26.09$ for testing against the saturated model. The model of complete independence has 62 degrees of freedom for testing against the saturated model. In Aitkin's method, a model, say X, with all main effects and some two-factor effects is deemed inadequate if

$$G_X^2 - 26.09 > \chi^2(.90, 62 - 37) = 34.39$$

or, equivalently, if

$$G_X^2 > 34.38 + 26.09 = 60.47 .$$

Any model that is not inadequate is adequate. A *minimal adequate model* is an adequate model that has no submodel that is deemed adequate. *Our primary interest is in identifying minimal adequate models.*

Among models with two-factor effects, the most informative models will be small models that fit and large models that do not fit. If a small model fits, any larger model also fits. If a large model does not fit, then any smaller model does not fit.

We begin by looking for small models that fit adequately. In particular, consider the first step of forward selection from the complete independence model.

Model	Added Effect	G^2
[R][S][OA]	AO	59.78
[R][SA][O]	SA	120.75
[RA][S][O]	RA	115.46
[R][SO][A]	SO	112.28
[RO][S][A]	RO	107.47
[RS][O][A]	RS	109.45

We are in luck! One of these models, [R][S][OA], has $G^2 < 60.47$, so it is deemed adequate. Thus, we have an extremely small model that is adequate and any larger model must also be adequate. The model [R][S][OA] is clearly a minimal adequate model. We have also established that any other minimal adequate models must have at least 2 two-factor effects (we have already checked all models with only 1 two-factor effect) and none of the two-factor effects can be OA (otherwise it will have [R][S][OA] as a submodel).

We now look for relatively large models that do not fit. Typically, a good approach is to look at the first step of backward elimination from the all two-factor effects model and hope to find some that do not fit adequately. For these data, however, we just established that any model larger than [R][S][OA] fits. In the first step of backward elimination, one two-factor effect is dropped out. Unless the two-factor effect dropped out is OA, we already know that the model will fit. The only model we need consider is [RS][RO][RA][SO][SA]; this model has a G^2 of 89.24. Once again, we are in luck. The value 89.24 is greater than the critical value 60.47, so [RS][RO][RA][SO][SA] is deemed inadequate. Thus, any model that does not include OA must be inadequate. Combining our two results, we have established that the only minimal adequate model is [R][S][OA].

We now consider what would occur if γ_2 was somewhat larger. Aitkin has suggested a method of choosing the γ_s's that will be discussed later. It begins with a level, say $\alpha = .05$, and for $t = 4$ factors with $s = 2$, Aitkin's choice is $\gamma_2 = 1 - (1-\alpha)^{\binom{4}{2}} = .265$, where $\binom{4}{2} = 6$ is the number of combinations of four things taken two at a time. Upon establishing that $\chi^2(.735, 25) = 28.97$ where $25 = 62 - 37$, a model X consisting of two-factor effects is deemed inadequate if

$$G_X^2 > 28.97 + 26.09 = 55.06\,.$$

The model [R][S][OA] has $G^2 = 59.78$, so it is no longer deemed adequate. However, it is very close to meeting the adequacy criterion. It makes sense to consider models that include [OA] and another two-factor effect. In particular, this is precisely the second step in forward selection starting with complete independence and adding simple effects. The models considered and their G^2's are

Model	G^2
[R][SA][OA]	59.06
[RA][S][OA]	53.77
[R][SO][OA]	50.59
[RO][S][OA]	45.79
[RS][OA]	57.76

The models [R][SA][OA] and [RS][OA] have $G^2 > 55.06$, so they are deemed inadequate. The models [RA][S][OA], [R][SO][OA], and [RO][S][OA] are adequate and no smaller models are adequate, so the models [RA][S][OA], [R][SO][OA], and [RO][S][OA] are minimally adequate models. Working from the all two-factor model down, the model of all two-factor effects except [OA] is inadequate ($G^2 = 89.24$), so all adequate models contain [OA].

The inadequate models, [R][SA][OA] and [RS][OA] need to be considered as to the additional terms needed to make them adequate. If we add RA,

SO, or RO, then we have made them larger than one of our minimally adequate models. The only two-factor effects that can generate additional minimally adequate models are SA and RS. If we add the appropriate effect to each model, we get [RS][SA][OA] in both cases. The G^2 for this model is 57.04. Because G^2 is greater than the critical value 55.06, [RS][SA][OA] is considered inadequate. Therefore, the only minimally adequate models are [RA][S][OA], [R][SO][OA], and [RO][S][OA]. All of these have simple interpretations in terms of conditional independence.

Aitkin's method applied to these data gives smaller models than any of the stepwise methods considered. Unfortunately, it missed the important [RSO] interaction.

General Discussion

We now present a general discussion of Aitkin's method. The method begins with a model of all s-factor effects that adequately fits the data, while the model with only the $s-1$-factor effects does not fit the data. The crux of the method is in identifying intermediate models that also give an adequate fit. If X is a *model that contains all $s-1$-factor effects and some but not all of the s-factor effects*, Aitkin tests the adequacy of X by rejecting adequacy of fit if

$$G_X^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s)$$

Here, G_X^2 is the likelihood ratio test statistic for testing model X against the saturated model. Note that the same criterion for rejection $\chi^2(1-\gamma_s, d_{s-1}-d_s)$ is used for any such model X. Also, if the all $s-1$-factor model is indeed an adequate fit, then because $G_X^2 - G_s^2 < G_{s-1}^2 - G_s^2$, the probability of a false rejection for any and all such models X is less than the probability of a false rejection of the all $s-1$-factor model.

The fact that the criterion of rejection does not depend on the particular model X leads to two important observations. If X is an adequate model, then any larger model, say W, must also be deemed adequate because $G_X^2 - G_s^2 \geq G_W^2 - G_s^2$. Similarly, if X is inadequate, then any smaller model W must also be deemed inadequate because $G_X^2 - G_s^2 \leq G_W^2 - G_s^2$. These facts together with Aitkin's restriction to only considering s-factor terms make it practical to examine all models that involve s-factor terms.

For the data of Example 6.2.1, Aitkin's method seeks to examine every model that is intermediate between the all three-factor model [TWM][TWD][TMD]
[WMD] and the all two-factor model [TW][TM][WM][TD][WD][MD]. For example, in testing the intermediate, model [TWM][TWD][TMD], the value of $G^2_{[TWM][TWD][TMD]} - G_3^2 = 35.65$ is larger than $\chi^2(.815, 4) = 6.178$, so the model [TWM][TWD][TMD] is not considered adequate. (The test statistic was obtained from Table 6.1 and $.815 = 1-\gamma_3$ from earlier in the section.) Note that the χ^2 value uses the 4 degrees of freedom as-

sociated with testing the all two-factor model against the all three-factor model. Similarly, any other model intermediate between the all two-factor and all three-factor models is tested against the all three-factor model using $\chi^2(.815, 4)$.

If the all two-factor model is adequate, the probability of a false rejection when testing the all two- and all three-factor models is $\gamma_3 = .185$. If the all two-factor model is adequate, then any intermediate model is adequate. Similarly, if an intermediate model is inadequate, then the all two-factor model must be inadequate. Because tests for intermediate models use the same χ^2 value but have smaller G^2 values than the all 2 versus all 3 test, a false rejection occurs for an intermediate model if and only if a false rejection occurs for the all two-factor model. (This is similar in spirit to Scheffé's method of multiple comparisons.)

Aitkin defines a subset of s-factor effects as a *minimal adequate subset* if no proper subset defines an adequate model. (The model is the model of all $s-1$-factor effects plus the subset of s-factor effects.) Typically, there will be several such models. Each of these models may be reduced further by testing any smaller-order (e.g., $s-1$) terms that are not forced into the model. These tests use the criterion of rejection appropriate for that order. [For an $s-1$-factor term, compare the test statistic to $\chi^2(1-\gamma_{s-1}, d_{s-2}-d_{s-1})$.] The fact that relatively few lower-order terms will not be forced into the model makes it practical to carry out this procedure.

The end result of Aitkin's model selection method is a collection of minimal adequate models. *These models should be compared for interpretability. The should also be compared to see how well they fit the model assumptions.* In fact, they can even be compared using the Adjusted R^2 or the Akaike information criteria discussed in Section 3.6.

Probably the main fault of Aitkin's method is the backward elimination in its first step. The first step is to decide on an adequate all s-factor model. For example, in a five-factor table, there are 10 three-factor terms. If 1 of the three-factor terms is substantial and the other 9 are not, then a test for all 10 terms will have little power to establish the need for this single three-factor term. We saw an example of this phenomenon earlier with the abortion opinion data. If it is important to pick up individual high-order interactions, Aitkin's method will be problematic. On the other hand, Aitkin's method may provide a good starting point to which an examination of higher-order interactions can be added.

Finally, we discuss the choice of γ_s's. Aitkin suggests choosing the γ_s's so that there is a probability no greater than, say, γ of rejecting the main-effects-only model (i.e., complete independence) when main-effects-only is adequate. Suppose there are t factors in the table. When complete independence is true, the various tests for s-order interactions are asymptotically independent. Thus, asymptotically, the probabilities of not rejecting any test is the product of the probabilities for the individual tests and the γ_i's

should be chosen to satisfy

$$1 - \gamma = \prod_{s=2}^{t} (1 - \gamma_s) \; .$$

(Note that we do not consider testing main effects, i.e., first-order "interactions.")

Particular values of the γ_s's can be chosen by analogy with a balanced t-factor analysis of variance. In a t-factor ANOVA, the number of s-factor effects is $\binom{t}{s}$. In a balanced ANOVA (with known variance), each of these tests might be conducted at some common level α.

Applying this idea to log-linear models, the probability of not rejecting any of the s-factor tests (assuming complete independence holds and tests are independent) is

$$1 - \gamma_s = (1 - \alpha)^{\binom{t}{s}} \; .$$

This determines a specific value for γ_s. The corresponding value of γ can be found using the binomial theorem. Because $2^t = \sum_{2=0}^{t} \binom{t}{s}$, we find that

$$\begin{aligned} \gamma &= 1 - \prod_{s=2}^{t} (1 - \gamma_s) \\ &= 1 - \prod_{s=2}^{t} (1 - \alpha)^{\binom{t}{s}} \\ &= 1 - (1 - \alpha)^{2^t - t - 1} \; . \end{aligned}$$

Aitkin (1979) suggests that it is reasonable to pick an α level that yields a γ between .25 and .5.

In Example 6.4.1, $t = 4$ and the γ_s values were chosen to satisfy

$$1 - \gamma_s = (1 - .05)^{\binom{4}{s}} \; ,$$

so $\gamma_4 = .05$, $\gamma_3 = .185$ and $\gamma_2 = .265$. These yield

$$\begin{aligned} \gamma &= 1 - \{(1 - .05)(1 - .185)(1 - .265)\}, \\ &= .431 \end{aligned}$$

which is in Aitkin's suggested range.

In the discussion of Aitkin's (1978) paper, D.R. Cox suggests that the emphasis on simultaneous testing in Aitkin's method is excessive. A compromise approach that seems intuitively appealing to the current author is, for some value α, to choose $\alpha = \gamma_2 = \cdots = \gamma_t$.

6.5 Model Selection Among Decomposable and Graphical Models

Wermuth (1976) has proposed a backward elimination technique that is restricted to the decomposable models discussed in Section 5.2. She focuses on identifying pairs of factors that can be viewed as conditionally independent. Wermuth's method is most easily understood in terms of graph theory and our general discussion of the method will center on that. However, we begin our discussion with an example that does not use graphical terms or motivations.

EXAMPLE 6.5.1. Consider again the race, sex, opinion, age data with a backward elimination cutoff of $\alpha = .05$. The initial model is the saturated model and we consider all factor pairs for possible conditional independence. This leads to the following lack of fit tests:

Factor Pair	Model	*df*	G^2	*P*
RS	[ROA][SOA]	18	20.45	.3080
RO	[RSA][SOA]	24	38.22	.0329
RA	[RSO][SOA]	30	22.09	.8506
SO	[RSA][ROA]	24	27.91	.2639
SA	[RSO][ROA]	30	11.29	.9989
OA	[RSO][RSA]	28	78.06	.0000

Note that for factor pair RS, the corresponding model [ROA][SOA] has R and S independent given O and A. Similar interpretations hold for the other pairs and models. The largest *P value* among the tests is for [RSO][ROA]. The *P value* is greater than .05, so we take as the model [RSO][ROA]. The conditional independence relation is that S and A are independent given R and O.

We have incorporated into the model the conditional independence of factors S and A. At the next step, we consider models that incorporate conditional independence between another pair of factors. In particular, we consider models that are reduced relative to [RSO][ROA] and incorporate an additional conditional independence. *The one exception is that the factor pair RO is in both terms of the model* [RSO][ROA], *so it cannot be considered.* As will be seen later in this section, incorporating a conditional independence between the pair RO would lead to a model that is not decomposable. For the pair OA, conditional independence is introduced by reducing the term [ROA] into [RO][RA]. The resulting model is [RSO][RO][RA]. The term [RO] is redundant because it is implied by [RSO]; thus, the reduced model is [RSO][RA]. A similar analysis holds for all other factor pairs except RO. The second step of the model selection method requires the following models and statistics:

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[RSO][ROA]	30	11.29	—	—	—
OA	[RA][RSO]	38	80.45	8	69.16	.0000
RA	[OA][RSO]	45	24.77	15	13.48	.5657
OS	[RS][ROA]	34	30.29	4	19.00	.0011
RS	[SO][ROA]	33	23.12	3	11.81	.0083

The tests presented as differences are tests of the given models against the model [RSO][ROA]. The largest *P value* is again greater .05 and belongs to [OA][RSO], so this model is used for the next step. The model [OA][RSO] incorporates the conditional independence of R and A in addition to the conditional independence of S and A obtained from the first step of the selection procedure. The model [OA][RSO] has R and S independent of A given O.

In the next step, we begin with the model [OA][RSO]. The pairs SA and RA have already been identified for conditional independence. *There are no pairs that are contained in both terms of the model, so there are no pairs that cannot be considered for possible conditional independence.* All pairs other than SA and RA are considered for the possibility that they are conditionally independent. To incorporate conditional independence between O and A into the model [OA][RSO], break the term [OA] into [O][A] and use the model [O][A][RSO], which is equivalent to [A][RSO]. To incorporate conditional independence between O and S into [OA][RSO], break the term [RSO] into [RO][RS] and use the model [OA][RO][RS]. Similar models are used for the factor pairs RS and RO. The necessary tests are given below.

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[OA][RSO]	45	24.77	—	—	—
OA	[A][RSO]	55	86.45	10	61.68	.0000
OS	[OA][RS][RO]	47	43.77	2	19.00	.0000
RS	[OA][SO][RO]	48	36.60	3	11.83	.0082
RO	[OA][SO][RS]	49	48.57	4	23.80	.0001

None of the *P values* is greater than .05, so we stop with the model [OA][RSO].

It is interesting to note that this happens to be the same model as achieved by forward selection of multiple effects. Note also that if $\alpha \leq .0082$, we would be led to consider R and S as conditionally independent and thus use the model [RO][SO][OA]. This is the other model achieved by stepwise regression.

In turn, [RO][SO][OA] would lead to taking S and O as conditionally independent and thus using the model [RO][S][OA]. These results can be seen from the following displays.

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[OA][SO][RO]	48	36.60	—	—	—
SO	[RO][OA][S]	50	45.79	2	9.19	.0101
OA	[RO][SO][A]	58	98.29	10	61.69	.0000
RO	[OA][SO][R]	50	50.59	2	13.99	.0009

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[RO][OA][S]	50	45.79	—	—	—
OA	[RO][S][A]	60	107.5	10	61.7	.0000
RO	[OA][R][S]	52	59.78	2	13.99	.0009

The model [RO][S][OA] is one of the minimally adequate models arrived at in our second application of Aitkin's method.

A key feature of Wermuth's method is that a pair of factors contained in more than one term in the model cannot be considered for conditional independence. Edwards and Havranek (1985) suggest dropping this requirement. Doing so changes the method from a search among decomposable models to a search among graphical models.

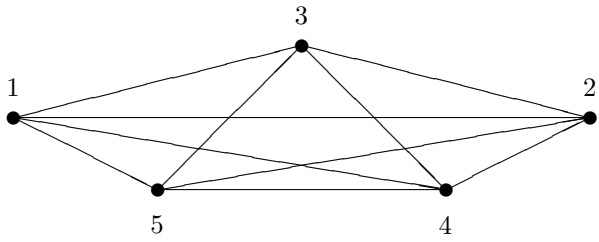
General Discussion

We now present a general discussion of Wermuth's method using graph-theoretic ideas. Recall that graphical models are completely determined by their two-factor effects. The procedure begins with the graphical model that includes all two-factor effects, i.e., the saturated model. Every two-factor effect is considered for deletion. The two-factor effect that generates the largest *P value* is deleted if the *P value* exceeds some cutoff point α.

Whichever effect is dropped, it determines two subsets of factors in which there are effects between every pair of factors. Recall that a subset with all possible two-factor effects is called *complete* and that if a complete subset is not strictly contained in any other complete subset, it is *maximal*. A maximal complete subset is a *clique*. Wermuth's method starts out with the clique based on all factors. Dropping one two-factor effect generates two cliques that each contain all but one factor.

Proceeding inductively, at any stage in Wermuth's method there are two or more cliques available. Two-factor effects that are part of more than one clique are not considered for elimination. It will be seen that the graphical model obtained by eliminating such effects is not decomposable. Among all other two-factor effects, the one with the largest *P value* is eliminated provided that the *P value* exceeds α.

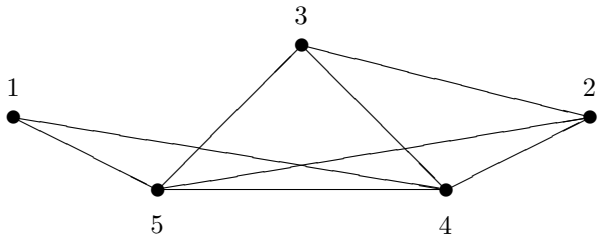
Example 6.5.2. With five factors the initial model is [12345]. This can also be thought of as the initial clique.



All two-factor terms are considered for elimination. There are $\binom{5}{2} = 10$ of these. If, for example, [12] is dropped, it is easily seen from the graph that there are two new cliques, [1345] and [2345]. Thus, the graphical model is [1345][2345]. Another way of establishing the graphical model is to examine the nine remaining two-factor terms: [13], [14], [15], [34], [35], [45], [23], [24], [25]. Of these nine, the first six terms generate [1345] and the last six terms generate [2345]; thus, the model is [1345][2345]. The *P value* for dropping [12] is the *P value* for testing the model [1345][2345] versus [12345].

Having deleted [12], the second stage begins by considering the cliques of the model [1345][2345]. The cliques are [1345] and [2345]. The two-factor effects [34], [35], and [45] are contained in both cliques, so these are not considered for elimination. Among the other two-factor terms, the one with the largest *P value* is eliminated provided the *P value* exceeds α.

In considering whether to drop, say, [13], the test compares the graphical model with both [12] and [13] eliminated to the graphical model in which only [12] is eliminated. As discussed above, when [12] is eliminated, the model is [1345][2345]. If [13] is also eliminated, the clique [1345] breaks up into two complete subsets [145] and [345].

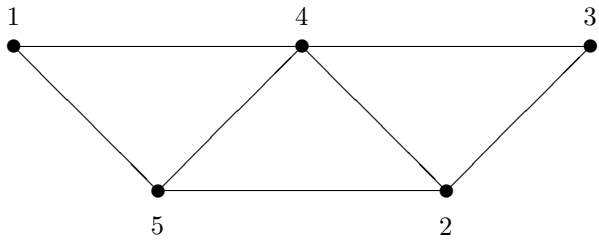


To see this, observe that [1345] is generated by [13], [14], [15], [34], [35], [45]. If [13] is dropped, the complete subsets are based on [14], [15], [45], and [34], [35], [45]. These generate [145] and [345], respectively. Note that [345] is not a clique because it is not maximal; it is contained in the complete subset [2345]. With both [12] and [13] eliminated, the cliques are [145] and

[2345], so the model is [145][2345]. The test for eliminating [13] is the test of [145][2345] versus [1234][2345].

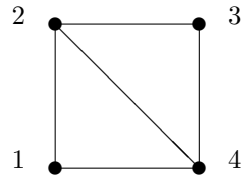
If [12] and [13] have been eliminated in the first two stages, the model is [145][2345]. The only two-factor effect contained in more than one clique is [45]. All other effects are considered for elimination. Suppose [35] is dropped. The resulting model is [145][234][245]. Now both [45] and [24] are in more than one clique, so neither can be eliminated. The process of testing and dropping two-factor terms continues until there are no tests with a *P value* greater than α.

To see that dropping an effect that is in more than one clique destroys decomposability, consider dropping [24] from the model [145][234][245].



From the two cliques involving [24], form the closed chain [34][45][52][23]. This chain of length 4 has one chord, [24]. If [24] is dropped, the model is no longer chordal, hence no longer decomposable. Whenever an effect contained in more than one clique is dropped, this construction of a closed chain of length four with no chords will work. Simply construct a closed chain out of the two vertices in the common effect and two other distinct vertices, one from each clique.

Conversely, if a model is decomposable and dropping a two-factor term makes it nondecomposable, then that two-factor term must be in more than one clique. For a model to become nondecomposable, the term eliminated must be the only chord in a closed chain of length four or more. If a closed chain of length four has only one chord, the chord must be in two complete subsets and thus in two cliques. In particular, if a decomposable model contains the closed chain [12][23][34][41] and contains *only* the one chord [24], then the model includes the complete sets [124] and [234], so [24] is contained in two different complete sets.



Each of the two complete sets must be contained in some clique and these cliques must be distinct. If the sets were in the same clique, then that clique would also have to include the chord [13]. By assumption, this cannot occur. In fact, this argument for closed chains of length four is sufficient for all cases because when a model becomes nondecomposable, the term eliminated must be the only chord in a closed chain of length four. In a decomposable model, any closed chain of length greater than four must have more than one chord because if the length is five or more and there is only one chord, there is a reduced closed chain of length at least four without a chord.

Clearly, Wermuth's method can be generalized to graphical models by removing the restriction that two-factor effects in more than one term are not considered for elimination. At each stage, all two-factor effects that have not been previously deleted can be considered for elimination. The corresponding models are the graphical models determined by the two-factor effects that have not been eliminated. This procedure was apparently first suggested by Edwards and Kreiner (1983). Model selection among graphical models is also discussed by Havranek (1984) and Edwards and Havranek (1985).

With four factors, the difference between Edwards and Kreiner's method and Wermuth's method occurs only at the second stage. This is due to the fact that there is only one graphical but nondecomposable model. As applied to the abortion opinion data, the term [SA] is dropped at the first stage, so at the second stage, Wermuth's method does not allow [RO] to be dropped. The method of Edwards and Kreiner has no such restriction.

For models with more than four factors, the difference between the graphical method and the decomposable method can be substantial. Restricting model search to graphical models seems like a very promising compromise between searching in the very large class of all ANOVA type models and searching in the very restrictive class of decomposable models. However, the difficulty of searching among graphical models should not be underestimated. Good (1975) has shown that for a 10-factor table there are almost 3.5 million graphical models.

With these methods as with all others, it is important to obtain several candidate models and to evaluate the models on grounds other than the

values of their test statistics. Also, effects included because of the sampling design cannot be eliminated.

6.6 Use of Model Selection Criteria

The best approach to model selection for log-linear models would be to search through all models and choose, for closer examination, those with high values of Adj. R^2, low values of AIC, or extreme values of some other model selection criterion, cf. Section 3.6. Such a procedure would require enormous amounts of computation. One possible way to reduce computations would be to base an initial search on models fitted by *weighted least squares* (cf. Sections 4.4 and 10.6) rather than models fitted by maximum likelihood.

At the moment, the author's best suggestion is to use a variety of model-fitting methods with a variety of critical (cutoff) values, and for stepwise methods, use a variety of initial models. By using several methods, we hope to find a wide range of possible models. These models can then be evaluated relative to each other with the help of the model selection criteria. Often, there is no need to decide on one particular model; a small number of alternative models may be more informative. If it is necessary to decide on only one model, the model with the lowest AIC (or highest Adj. R^2) may not be the best choice. Other considerations such as interpretability and consistency with assumptions may dictate choosing a model with a low AIC but not necessarily the lowest.

EXAMPLE 6.6.2. The evidence presented so far in the series of examples on the race, sex, opinion, age data strongly suggests to the author that the best model is [RSO][OA]. A formal comparison of all of the models arrived at by the various methods suggests the same conclusion.

Model	*df*	G^2	$A - q$	R^2	Adj. R^2
[RSO][OA]	45	24.77	−65.23	.80	.72
[RO][SO][OA]	48	36.60	−59.40	.70	.61
[R][S][OA]	52	59.78	−44.22	.51	.41
[RA][S][OA]	47	53.77	−40.23	.56	.42
[R][SO][OA]	50	50.59	−49.41	.58	.48
[RO][S][OA]	50	45.79	−54.21	.62	.53

In a less clear-cut situation, it would be wise to consider many more stepwise procedures than have been illustrated.

One worrisome aspect is that very little consideration has been given to three-factor effects other than RSO. The reader can check that none of the other three-factor effects substantially improves the model.

We have decided on one particularly good candidate model: [RSO][OA]. However, *the analysis of the data does not end with finding an appropriate ANOVA type model; that is just an important first step.* The model indicates that combinations of race and sex are independent of age given opinions about legalized abortions. Thus, we can collapse over some factors to study interrelationships in marginal tables. We can collapse over ages to study the relationships among race, sex, and opinion. We can collapse over race and sex to study the relationship between age and opinion. Cell counts for the collapsed tables need to be examined to study the nature of the relationships. These aspects of the analysis are discussed further in Section 8. It is also necessary to evaluate whether the model really fits the data. To this end, Section 7 contains information on residual analysis and influential observations. Finally, the interpretability of the model should be examined. This model has a very nice interpretation with race and sex independent of age given opinion. Does the interpretation make any sense? As we will see in Section 8, interpretability is probably this model's weakest point. It can be argued that the appropriate analysis of these data involves explaining opinions on the basis of race, sex, and age. In that case, the methods of Chapter 4 should be used on these data.

6.7 Residuals and Influential Observations

In standard regression analysis, it is common practice to use residuals to check whether assumptions made in the model are valid and to detect the presence of observations that are unusually influential on the fit of the model. Three statistics that are commonly used for these purposes are the leverages, the standardized residuals, and Cook's distances. As seen in Chapter 4, residuals are not of much use when dealing with binary (0-1) data. However, in tables with reasonably large counts, residuals can be useful.

In a regression model $Y = X\beta + e$, the leverages are the diagonal elements of the projection matrix $X(X'X)^{-1}X'$. The matrix X is determined by the design of the data, i.e., the values of the predictor variables in the regression. The leverage measures how far the variables of a particular case are from the average of all of the cases. (The projection matrix is often called the "hat" matrix because it changes Y into $\hat{Y}$, i.e., $\hat{Y} = X(X'X)^{-1}X'Y$. Personally, I find this name distasteful. However, given the liberties I am about to take in naming residuals, I am probably not in a position to make a fuss.)

For log-linear models, there is an analogue of the leverage that depends both on the design and the probability of getting observations in a particular cell. Because the probabilities are unknown, they must be estimated; hence, we use estimated leverages. The estimated *leverage* of the ith case

is denoted

$$\hat{a}_{ii} \ .$$

Leverages are discussed in detail in Chapter 10.

The log-linear analogue of a *standardized residual* is

$$r_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i(1-\hat{a}_{ii})}} \ .$$

For a correct model, a large sample approximation for the distribution of r_i is $r_i \sim N(0,1)$. Note that these are very similar to the *Pearson residuals* discussed earlier:

$$\tilde{r}_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i}} \ ;$$

they differ only in that the standardized residuals involve the leverages.

Actually, the residuals are the values $n_i - \hat{m}_i$, the difference between the observed and predicted values. As discussed in Section 2.1, these need to be standardized in some way. The Pearson residuals use a crude standardization, dividing by $\sqrt{\hat{m}_i}$. In Section 10.7, the Pearson residuals are referred to as the *crude standardized residuals* to distinguish them from the standardized residuals that involve a more sophisticated (and proper) standardization. This terminology was chosen by analogy with regression analysis. Unfortunately, it differs from the terminology used by many authors on log-linear models. Often, the Pearson (crude standardized) residuals are referred to as simply the standardized residuals, and the values defined here as the standardized residuals are referred to as the *adjusted* residuals.

Finally, *Cook's distance* for the ith case is a measure of the influence the ith case has on the fit of the model. In the context of fitting log-linear models, we drop each cell from the table, fit the remaining cells, and then estimate an expected cell count for the dropped cell. This is done without reference to any marginal totals that may be fixed by design; hence, it is most appropriate for Poisson sampling. If the model has p degrees of freedom, the analogue of Cook's distance can be written

$$C_i = \sum_{\text{all cells } r} \hat{m}_r \left[\log(\hat{m}_r/\hat{m}_{r(i)})\right]^2 \Big/ p$$

where $\hat{m}_{r(i)}$ is the estimate of the rth cell when the ith cell has been deleted (cf. Section 10.7). For Poisson sampling, these values can be calibrated by comparing them to a $\frac{1}{p}\chi^2(p)$ distribution. If $C_i > \chi^2(.5,p)/p$, cell i has a substantial influence. For multinomial or product-multinomial sampling, the degrees of freedom should be reduced by the number of independent multinomials. Because computation of all the $\hat{m}_{r(i)}$'s would require separate iterative procedures and be very expensive, it is suggested that a one-step estimate be used. Starting with the values $\hat{m}_r$, drop a cell and, rather than

fully iterating, use just one step of the Newton-Raphson algorithm to obtain the $\hat{m}_{r(i)}$'s, cf. Section 10.7.

This definition of the Cook's distances has weaknesses; however, the situation is analogous to that of the Pearson residuals. The definition of the Pearson residuals as standardized residuals is weak, but the Pearson residuals are easy to compute and they contain valuable information. As will be seen below, using standard computer software, the standardized residuals are now easy to compute, so there is little reason to use the Pearson residuals. Similarly, using standard computer software, Cook's distances, as defined here, are easy to compute. Moreover, the author feels that they contain valuable information. Until something better becomes readily available, the author suggests examining these Cook's distances. Anderson (1992) discusses diagnostics for categorical data analysis and Thomas and Cook (1989, 1990) discuss influence for generalized linear models (which include log-linear models, cf. Chapter 9).

6.7.1 Computations

We assume that the reader is capable of fitting an ANOVA model using a regression program. Our log-linear models are ANOVA type models. Good computer programs for doing regression generally provide leverages, standardized residuals, and Cook's distances. Also, they allow for computing weighted regressions.

After fitting the log-linear model, retain the counts for each cell, n_i, and the fitted values for each cell, $\hat{m}_i$. Now use the regression program to refit the ANOVA model, but use weighted regression with

$$\text{weight}_i = \hat{m}_i$$

and a dependent variable

$$Y_i = \log(\hat{m}_i) + (n_i - \hat{m}_i)/\hat{m}_i \ .$$

The leverages given by the program will be the $\hat{a}_{ii}$'s. The standardized residuals reported will be

$$r_i/\sqrt{\text{MSE}}$$

where $\sqrt{\text{MSE}}$ is the estimate of the standard deviation from the regression. Simply multiply the reported standardized residuals by $\sqrt{\text{MSE}}$ to obtain the correct values. The reported values of Cook's distances are

$$C_i/\text{MSE}$$

where C_i is computed using a one-step fit. The reported values C_i need to be multiplied by MSE. For the purpose of comparing the relative magnitudes of the r_i's or C_i's, the multiplication is irrelevant. For comparing

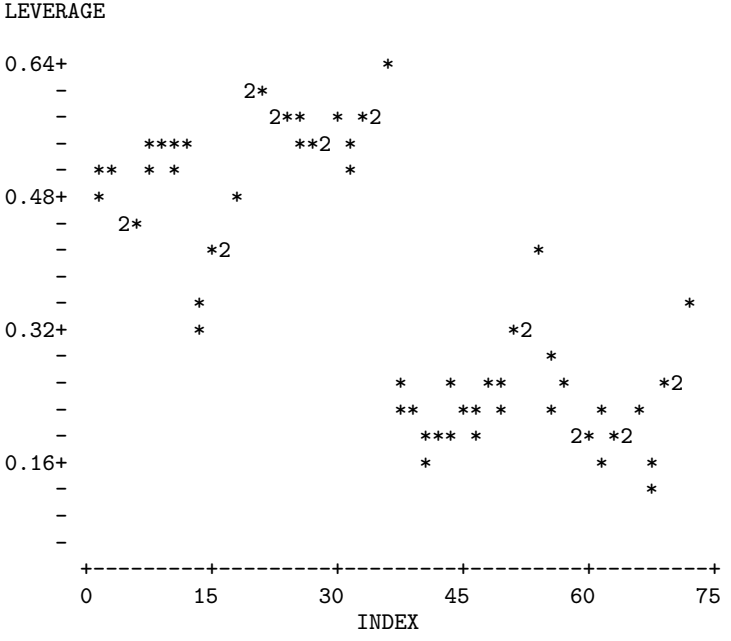


FIGURE 6.1. Leverage–Index Plot

standardized residuals to a $N(0,1)$ distribution or Cook's distances to a χ^2 distribution, the multipliers are important. It should also be noted that for large samples and a correct model, the MSE approaches a χ^2 distribution divided by its degrees of freedom. The large sample expected value for the MSE is 1.

EXAMPLE 6.7.1. We now examine the leverages, standardized residuals, and Cook's distances for the abortion opinion data. In particular, we consider the fit of the model [RSO][OA]. The fitted values are given in Section 8 as Table 6.7.

The leverages are plotted against index values in Figure 6.1. The index values are just values $1, 2, \ldots, 72$ assigned to the cells. The G^2 for the model [RSO][OA] has 45 degrees of freedom. There are 72 cells, so there are $72 - 45 = 27$ degrees of freedom for the model. The sum of the 72 leverages must add up to 27. The average leverage is $\frac{27}{72} = .375$. The largest leverage is about .64, which is less than twice the average. None of the leverages seems excessively large. Leverages are rarely very large in balanced ANOVA type models.

Figures 6.2 and 6.3 contain a box plot and an index plot of the standardized residuals, respectively. The box plot identifies one very large residual and four other large residuals. In the index plot, only two residuals really

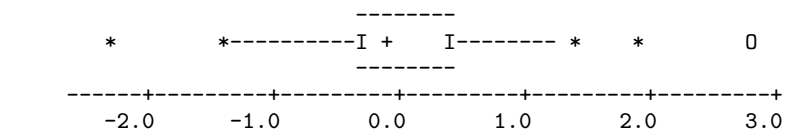


FIGURE 6.2. Standardized Residual–Box Plot

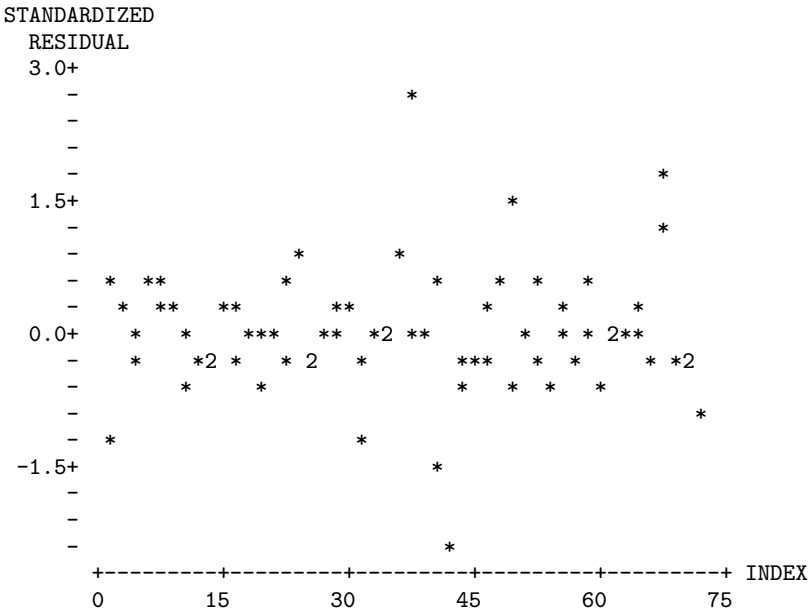


FIGURE 6.3. Standardized Residual–Index Plot

stand out. There were 24 nonwhite males between 18 and 25 years of age who support legalized abortion; the estimated value from the model is only 14.52. This cell has a leverage of .222, a standardized residual of 2.82, and a Cook's distance of .085. The other large standardized residual is from the cell for nonwhite males above 65, who support legalized abortion. The observed value in this cell is 4, the fitted value is 10.90, the leverage is .181, the standardized residual is -2.31, and Cook's distance is .044. Considering that there are 72 cells, these values are not remarkably large. In fact, what is remarkable is that most of the standardized residuals are so tightly packed around zero.

Figure 6.4 contains a normal probability plot of the standardized residuals. If the asymptotic theory is valid, the plot should be approximately linear. It is not. Again, the problem seems to be that there are too many cells fitted too well by the model.

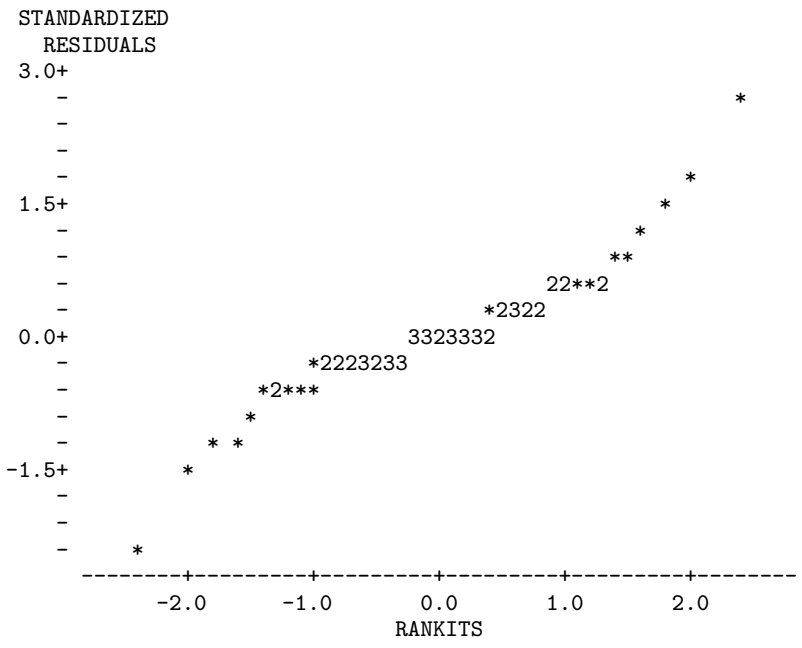


FIGURE 6.4. Standardized Residual–Rankit Plot

The MSE for the regression from which the residuals were obtained was .571. Under the sampling schemes considered, the MSE is a lack of fit test statistic. For large samples and a correct model, the MSE should be near 1. For incorrect models, it should be greater than 1. Again, the model fits better than we have any right to expect.

When examining leverages directly, we look for large individual leverages because they indicate that a cell is unlike the other cells. Large leverages are leverages near 1. Leverages also play a role in standardized residuals. Even if none of the leverages is large, they can be important. In particular, the range of leverages in this example is from .115367 to .632986. The factors $1/\sqrt{1-a_{ii}}$ are used to change Pearson residuals into standardized residuals. These factors range from 1.063 to 1.651, so a Pearson residual with the largest leverage would have a standardized residual more than 1.5 times larger than the same Pearson residual would have with the smallest leverage. This is not an inconsequential difference. Pearson residuals and standardized residuals can lead to very different conclusions.

Figure 6.5 contains a box plot of the Cook distances. Figure 6.6 contains an index plot of the Cook distances. Most cells have very little influence on the fit. The plots identify a number of relatively influential cases, but as compared to the calibrating values of $\frac{1}{27}\chi^2(27)$, none of the distances are substantial. For example, $\frac{1}{27}\chi^2(.5, 27) = 26.3/27 = .974$; none of the

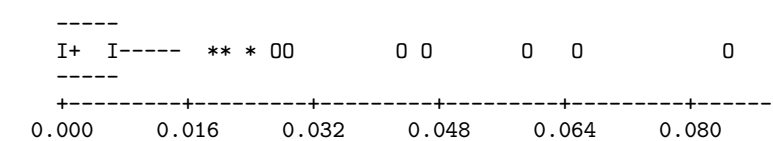


FIGURE 6.5. Cook's Distance–Box plot

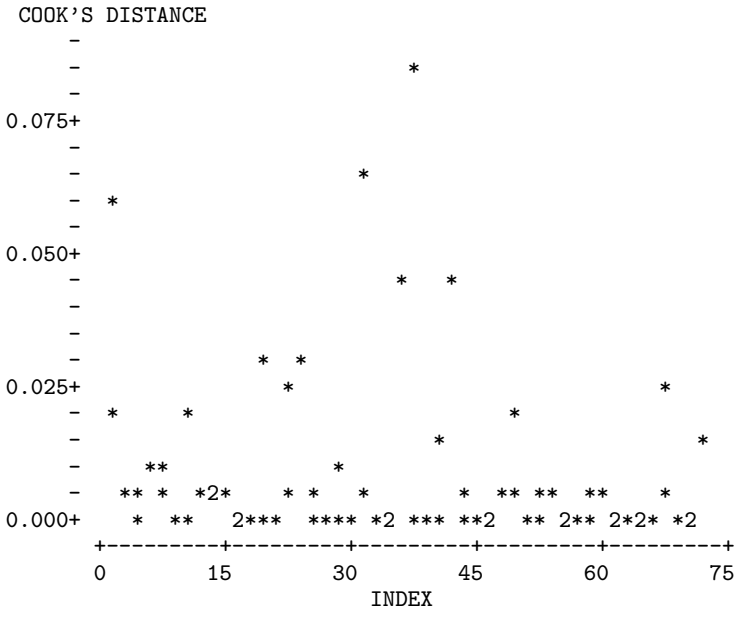


FIGURE 6.6. Cook's Distance–Index Plot

distances are anywhere near .974. The largest Cook's distance is the .085 for young nonwhite males who support legalized abortion.

6.7.2 Computing Commands

Below are commands that give the diagnostics provided by BMDP-4F.

```
/ INPUT     FILE = 'C:\LOGLIN\ABORT.DAT'.
         FORMAT = FREE.
         VARIABLES = 5.
/ VARIABLE  NAMES = R, S, A, O, N.
/ TABLE     INDEX = R, S, A, O.
            COUNT = N.
/ STAT      ALL.
```

```
/ PRINT     LINE = 79.
            LAMBDA.
            BETA.
            VAR.
            STAN.
            CHISQ.
            LRCHI.
/ FIT       MODEL = RSO, OA.
/ END
```

Diagnostics are also available from GLIM. The procedure for fitting loglinear models was illustrated in Subsection 3.7.1. The commands for diagnostics are exactly as in Subsection 4.4.2.

6.8 Drawing Conclusions

We have discussed model interpretation, model selection, and model validation. We have selected a model [RSO][OA] that is reasonably small, fits well, and has a simple interpretation. No cells seem to be unduly influential and no cells have outrageously bad fits. The model indicates that, given Opinion, Age is independent of Race and Sex. As discussed in the introduction to this chapter, the model is a description of the data and can be used to predict behavior in a similarly conducted study. It is not a statement about causation. In fact, it makes little sense to imagine that opinions about abortion cause the relative frequencies of Race and Sex to be independent of the frequencies of Age.

By itself, the model tells us nothing about the relationships among Race, Sex, and Opinion or about the relationship between Age and Opinion. Table 6.7 contains the estimated expected cell counts under the model [RSO][OA]. It is a complicated table, but much could be learned from studying it. Fortunately, there are easier ways to get at this information; we can collapse factors and study marginal tables.

As discussed in Section 5.3, the Race-Sex-Opinion relationships can be examined by collapsing over Age and looking at the Race-Sex-Opinion marginal table. This is given in Table 6.8. We see that whites are more likely to be in the survey than nonwhites. White females are a bit more likely to appear than white males. Among nonwhites, males and females are about equally likely. Ignoring the undecideds, the rate of support for legalized abortion is about the same for white and nonwhite males. It is higher for nonwhite females than white females. It is higher for white females than for white males. All of these things can be examined using odds ratios similar to Section 4.6. Formal inference requires standard errors as discussed in Section 10.2.

TABLE 6.7. Estimated Cell Counts under [RSO][OA]

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	65+
White	Male	Support	105.5	132.1	114.5	76.94	72.81	79.19
		Oppose	41.87	61.93	54.95	47.98	52.34	61.93
		Undec.	1.39	2.50	5.27	5.27	5.54	8.04
	Female	Support	141.0	176.7	153.1	102.9	97.38	105.9
		Oppose	44.61	65.98	58.55	51.11	55.76	65.98
		Undec.	2.43	4.37	9.22	9.22	9.70	14.07
NonWhite	Male	Support	14.52	18.19	15.76	10.59	10.03	10.90
		Oppose	5.61	8.30	7.36	6.43	7.01	8.30
		Undec.	0.84	1.52	3.20	3.20	3.37	4.88
	Female	Support	19.97	25.01	21.67	14.57	13.79	14.99
		Oppose	3.91	5.79	5.14	4.48	4.89	5.79
		Undec.	0.35	0.62	1.32	1.32	1.39	2.01

TABLE 6.8. Race, Sex, Opinion Marginal Table

		Opinion			
Race	Sex	Support	Oppose	Undec.	Totals
White	Male	581	321	28	930
	Female	777	342	49	1168
Nonwhite	Male	80	43	17	140
	Female	110	30	7	147
	Totals	1548	736	101	2385

TABLE 6.9. Opinion, Age Marginal Table

	Age						
Opinion	18-25	26-35	36-45	46-55	56-65	65+	Totals
Support	281	352	305	205	194	211	1548
Oppose	96	142	126	110	120	142	736
Undec.	5	9	19	19	20	29	101
Totals	382	503	450	334	334	382	2385

To examine the relationship between Opinion and Age, we can collapse over Race and Sex. The marginal totals are given in Table 6.9. We see that some age groups are more likely to respond. There is more support than opposition in each age group. Undecideds increase with age. Also, the amount of support seems to decrease with age.

I find myself not really caring about the relative incidences of Race and Sex, but rather am interested in the relative support for legalized abortion among the different groups. This amounts to treating Opinion as a response variable and Race and Sex as explanatory variables; i.e., Race and Sex can be imagined to determine Opinion. In my experience, most contingency tables have one or more factors of particular interest that can be considered as response factors. (Of course, that is only in my experience.) Specific methods for analyzing tables with response factors were examined in Chapter 4.

If O were to be treated as a response and R, S, A as factors explaining that response, Asmussen and Edwards (1983) argue that, of the models listed in Example 6.6.2, only [RA][S][OA] is appropriate. Recall from Section 4.6 their contention that log-linear models are appropriate for response factors only if the model allows for collapsing over the response factors onto the explanatory factors, cf. Section 5.3. For example, the model [RSO][OA] can be collapsed over Race and Sex or collapsed over Age but not over the response factor opinion. Therefore, [RSO][OA] is not a reasonable log-linear model for the response factor O. It is illogical for Race and Sex to be independent of Age given the factor Opinion which is supposed to be a response. The response cannot generate independence between explanatory factors! On the other hand, models such as [RA][S][OA] are reasonable. Recall that [RA][S][OA] is one of the minimally adequate models found by Aitkin's method. However, you should also recall that Aitkin's method totally missed the important [RSO] interaction.

6.9 Exercises

EXERCISE 6.9.1. Reanalyze the auto accident data of Example 4.8.1 without treating any of the factors as a response factor.

EXERCISE 6.9.2. Reanalyze the abortion attitude data of Exercise 4.8.4 without treating any of the factors as a response.

EXERCISE 6.9.3. Using our discussion of graphical models and collapsibility in Chapter 5, argue that when the true model is [123][24][456],

the test of marginal association for the $u_{123(ijk)}$'s does not ignore any vital information. Is the same conclusion appropriate when the true model is [12][13][23][24][456]? What if the true model is [123][124][456], [123][24][456][15], or [123][24][456][15][36]?

7

Models for Factors with Quantitative Levels

Just as in analysis of variance, if the levels of some factors are associated with quantitative values, these values can be used in the analysis. For example, in a two-factor ANOVA where factors are two kinds of fertilizer and levels are different quantitative amounts of fertilizers, an ANOVA would often examine linear and higher-order contrasts in the main effects and polynomial contrasts in the interaction (e.g., the linear-by-linear contrast).

In the analysis of categorical data, it is relatively rare that a factor has truly quantitative values associated with its levels. Often categories are ordered, but are not intrinsically quantitative. This is referred to as having *ordinal factor levels* or simply as having ordinal data. For example, socioeconomic status is often categorized as low, medium, and high. Surely, it would be advantageous to incorporate information about ordering into the analysis. The problem is in finding an appropriate method. The most commonly used method is to assign scores to the three levels of the factor. These scores can be any known numbers, say, x_1, x_2, and x_3. In fact, the most common method is to take $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. The analysis then proceeds as if the scores are true quantitative levels.

Alternatively, a factor might be income and the levels of the factor could be income intervals, say, less than \$20,000, \$20,000-\$40,000, and more than \$40,000. To have quantitative levels, we need one number associated with each level. Such numbers simply do not exist. As a practical matter, we could use the midpoints of the intervals as quantitative levels (scores). We can then develop models based on these approximate scores. Of course, if actual incomes are available for each individual, it would be more suitable to use the incomes in an appropriate regression analysis, cf. Section 4.1.

However, in practice it is not uncommon to encounter factors that have been created by categorizing continuous variables.

Continuous variables that have been categorized present some unique difficulties when assigning scores. With categories that are intervals, using the midpoints of the intervals is simple and appealing. In the income example above, the midpoint scores \$10,000 and \$30,000 may work reasonably well as quantitative levels for the first two income intervals. However there is no natural quantitative level to use for the category "more than \$40,000." Any analysis is dependent on the score that is chosen to represent the third category. Moreover, using midpoints may not be very efficient. Suppose it was known that most people in the less than \$20,000 interval had incomes near \$20,000. It would be better to use a score that was near \$20,000 rather than using the midpoint \$10,000. In practice, the method of determining a score will often depend on additional sources of information. If \$10,000 is not an appropriate score, there must be additional information leading to that conclusion. Use the same additional information to arrive at an alternative score.

In this chapter, we examine models that incorporate the quantitative nature of factor levels. When factor levels are not truly quantitative, the appropriateness of such models will be directly related to the appropriateness of the scores being used. We assume that the quantitative levels (i.e., scores) are known and consider linear models for the log of the expected cell counts (i.e., log-linear models). An alternative approach is to consider the scores as parameters and to estimate the scores. If the scores are parameters, then the models considered are no longer *linear* models for the log of the expected cell counts. Such models are discussed in Section 3.

7.1 Models for Two-Factor Tables

Consider a 3×4 table with quantitative levels x_1, x_2, x_3 and w_1, w_2, w_3, w_4. If the observations in the table were 12 normally distributed values y_{ij}, an analysis of variance would be appropriate. The model with no interaction is

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + e_{ij} \,. \tag{1}$$

In this model, the 2 degrees of freedom for the main effect of Factor 1 can be broken into a linear contrast and a quadratic contrast. The three degrees of freedom for Factor 2 can be broken into a linear contrast, a quadratic contrast, and a cubic contrast. Equivalently, we can rewrite model (1) as a regression model

$$y_{ij} = u + \beta_1 x_i + \beta_2 x_i^2 + \eta_1 w_j + \eta_2 w_j^2 + \eta_3 w_j^3 + e_{ij} \,. \tag{2}$$

Now consider the full interaction model

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + e_{ij} \,. \tag{3}$$

Note that with only one observation per cell, the terms $u_{12(ij)}$ are hopelessly confounded with the errors e_{ij}. Since our goal is only to draw analogies between analysis of variance and log-linear models, we need not concern ourselves with this confounding. To explore the interaction in model (3), we can consider contrasts in the interactions. Using the quantitative factor levels leads to considering things like the linear-by-linear, linear-by-quadratic, and quadratic-by-cubic interaction contrasts. In total, there are $(3-1)(4-1) = 6$ of these linearly independent interaction contrasts. Alternatively, we can rewrite model (3) using regression terms in place of the interactions,

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma_{11}x_i w_j + \gamma_{12}x_i w_j^2 + \gamma_{13}x_i w_j^3 + \gamma_{21}x_i^2 w_j + \gamma_{22}x_i^2 w_j^2 + \gamma_{23}x_i^2 w_j^3 + e_{ij}.$$

This suggests a variety of partial interaction models that can be considered. The simplest of these models is

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma_{11}x_i w_j + e_{ij}\,.$$

This is the model that provides for main effects in each factor, but models the interaction as consisting entirely of linear-by-linear interaction.

7.1.1 Log-Linear Models with Two Quantitative Factors

Exactly the same procedures are used when the data consist of counts. Consider an $I \times J$ table with quantitative levels $x_1, \ldots, x_I$ and $w_1, \ldots, w_J$. The model of independence is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}\,.$$

The model of full interaction (the saturated model) is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}\,.$$

We can structure the interaction by considering a linear-by-linear association model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma x_i w_j\,. \tag{4}$$

The maximum likelihood estimates $\hat{m}_{ij}$ must satisfy

$$\begin{aligned} \hat{m}_{i\cdot} &= n_{i\cdot}\,, \qquad i = 1, \ldots, I, \\ \hat{m}_{\cdot j} &= n_{\cdot j}\,, \qquad j = 1, \ldots, J, \end{aligned}$$

and

$$\sum_{ij} \hat{m}_{ij} x_i w_j = \sum_{ij} n_{ij} x_i w_j\,.$$

(This is easily seen from the results of Chapter 10.) Model (4) can be tested against the saturated model using either G^2 or X^2. The reduced model of independence can be tested against model (4) using either G^2, X^2, or $\hat{\gamma}/\mathrm{SE}(\hat{\gamma})$.

Often, model (4) is written in an equivalent form. If observations in all IJ cells are possible, model (4) is equivalent to

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + \gamma(x_i - \bar{x}.)(w_j - \bar{w}.) \tag{5}$$

where $\bar{x}. = \frac{1}{I}\sum_{i=1}^{I} x_i$ and $\bar{w}. = \frac{1}{J}\sum_{j=1}^{J} w_j$. Frequently, the factor levels are equally spaced. This means that for some constants c and d, $x_{i+1} - x_i = c$, $i = 1, \ldots, I-1$, and $w_{j+1} - w_j = d$, $j = 1, \ldots, J-1$. This special case turns out to be equivalent to taking $x_i = i$, $i = 1, \ldots, I$ and $w_j = j$, $j = 1, \ldots, J$. Model (4) can be rewritten as

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma(i)(j)\,. \tag{6}$$

For equally spaced levels, model (6) is a reparametrization of model (4). The parameter γ in (6) is only identical to the parameter γ in (4) when the original scores are $x_i = i$ and $w_j = j$. In particular, γ in model (6) is equivalent to γdc in model (4). The scores $x_i = i$ and $w_j = j$ are used frequently when the factor levels are ordinal.

Model (4), when applied with scores that are equally spaced, is called the *uniform association* model. The name is apt because under this model, the odds ratios for consecutive table entries are identical. In particular,

$$\frac{m_{ij}m_{i+1\,j+1}}{m_{ij+1}m_{i+1\,j}} = e^{\gamma dc}\,,$$

$i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$. To see this, note that

$$\begin{aligned}
&\log\left(\frac{m_{ij}m_{i+1\,j+1}}{m_{ij+1}m_{i+1\,j}}\right)\\
&\quad= \log m_{ij} - \log m_{ij+1} - \log m_{i+1\,j} + \log m_{i+1\,j+1}\\
&\quad= u + u_{1(i)} + u_{2(j)} + \gamma x_i w_j\\
&\qquad - u - u_{1(i)} - u_{2(j+1)} - \gamma x_i w_{j+1}\\
&\qquad - u - u_{1(i+1)} - u_{2(j)} - \gamma x_{i+1} w_j\\
&\qquad + u + u_{1(i+1)} + u_{2(j+1)} + \gamma x_{i+1} w_{j+1}\\
&\quad= \gamma[x_i w_j - x_i w_{j+1} - x_{i+1} w_j + x_{i+1} w_{j+1}]\\
&\quad= \gamma[x_i(w_j - w_{j+1}) - x_{i+1}(w_j - w_{j+1})]\\
&\quad= \gamma[x_i(-d) - x_{i+1}(-d)]\\
&\quad= \gamma d[x_{i+1} - x_i]\\
&\quad= \gamma dc\,.
\end{aligned}$$

If $x_i = i$ and $w_j = j$, then $d = 1$ and $c = 1$,, so consecutive log odds ratios equal γ.

The case of equal spacings arises very often, so we will always refer to model (4) and its equivalent, model (5), as the model of uniform association. If factor levels are not equally spaced, this terminology is meaningful in the sense that γ is a measure of association that applies uniformly, but any particular odds ratio must also be adjusted for differences in factor levels.

Finally, note that there is much more flexibility available than merely considering the independence model, the uniform association model, and the saturated model. The saturated model is equivalent to

$$\begin{aligned}\log m_{ij} \quad = \quad & u + u_{1(i)} + u_{2(j)} \\ & + \gamma_{1,1}x_i w_j + \gamma_{1,2}x_i w_j^2 + \cdots + \gamma_{1,J-1}x_i w_j^{J-1} \\ & + \gamma_{2,1}x_i^2 w_j + \gamma_{2,2}x_i^2 w_j^2 + \cdots + \gamma_{2,J-1}x_i^2 w_j^{J-1} \\ & \quad \vdots \\ & + \gamma_{I-1,1}x_i^{I-1} w_j + \gamma_{I-1,2}x_i^{I-1} w_j^2 + \cdots + \gamma_{I-1,J-1}x_i^{I-1} w_j^{J-1} .\end{aligned}$$

A wide variety of submodels can be fitted. For example, we could consider a second-order interaction model, say

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma_{1,1}x_i w_j + \gamma_{1,2}x_i w_j^2 + \gamma_{2,1}x_i^2 w_j + \gamma_{2,2}x_i^2 w_j^2 .$$

This model is larger than the uniform association model (5), but smaller than the saturated model. The maximum likelihood estimates for this model satisfy the equations

$$\begin{aligned}\hat{m}_{i\cdot} &= n_{i\cdot} , & i &= 1, \ldots, I, \\ \hat{m}_{\cdot j} &= n_{\cdot j} , & j &= 1, \ldots J, \\ \sum_{ij} x_i w_j \hat{m}_{ij} &= \sum_{ij} x_i w_j n_{ij} , \\ \sum_{ij} x_i w_j^2 \hat{m}_{ij} &= \sum_{ij} x_i w_j^2 n_{ij} , \\ \sum_{ij} x_i^2 w_j \hat{m}_{ij} &= \sum_{ij} x_i^2 w_j n_{ij} , \\ \sum_{ij} x_i^2 w_j^2 \hat{m}_{ij} &= \sum_{ij} x_i^2 w_j^2 n_{ij} .\end{aligned}$$

7.1.2 Models with One Quantitative Factor

Suppose that only the first factor in an $I \times J$ table has quantitative levels. Denote these levels as $x_1, \ldots, x_I$. We still want to consider models that are more general (larger) than the model of independence, but smaller than the saturated model. A frequently used model in this situation is the *column effects* model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \tau_j x_i . \tag{7}$$

This model implies that there is a linear effect on $\log m_{ij}$ from the rows of the table, but that the slope (τ) of this linear effect changes from column to column. In the case of I populations and $J = 2$ responses, model (7) is also a simple linear logistic regression model, cf. Section 2.6.

Although model (7) is appropriate when only one factor is quantitative, it can also be used when both factors are quantitative. Note that model (4) is a reduced model relative to model (7) in which the additional structure $\tau_j = \gamma w_j$ is imposed. Thus, the uniform association model assumes that the slopes change linearly with the columns. Model (7) is more general in that it allows arbitrary changes in the slopes. In particular, model (7) is equivalent to the model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + \gamma_{11} x_i w_j + \gamma_{12} x_i w_j^2 + \cdots + \gamma_{1,J-1} x_i w_j^{J-1} .$$

Maximum likelihood estimates for model (7) must satisfy

$$\begin{aligned} \hat{m}_{i\cdot} &= n_{i\cdot} , & i &= 1, \ldots, I, \\ \hat{m}_{\cdot j} &= n_{\cdot j} , & j &= 1, \ldots, J, \\ \sum_{i=1}^{I} \hat{m}_{ij} x_i &= \sum_{i=1}^{I} n_{ij} x_i , & j &= 1, \ldots, J . \end{aligned}$$

Testing is performed in the usual way.

More generally, we can consider any submodel of the saturated model. The saturated model can be reparametrized as

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \tau_{1,j} x_i + \tau_{2,j} x_i^2 + \cdots + \tau_{I-1,j} x_i^{I-1} .$$

For $J = 2$, this is the $I - 1$-degree polynomial logistic regression model

$$\log \left(m_{i1} / m_{i2} \right) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{I-1} x_i^{I-1} .$$

Of course, if the second factor in the table is quantitative rather than the first factor, we can write the saturated model as

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \eta_{i,1} w_j + \cdots + \eta_{i,J-1} w_j^{J-1}$$

and consider reduced models. The *row effects* model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \eta_i w_j$$

is probably the most frequently used of these.

Although it seems to be done infrequently, there is no mathematical reason not to fit models using regression in place of main effects. For example, a reduced model relative to the uniform association model is

$$\log m_{ij} = u + \beta x_i + \eta w_j + \gamma x_i w_j .$$

This model also implies uniform association (in terms of the odds ratios when levels are equally spaced), but imposes additional constraints.

EXAMPLE 7.1.1. A sample of men between the ages of 40 and 59 was taken from the city of Framingham, Massachusetts. The men were cross-classified by their serum cholesterol and systolic blood pressure. We restrict attention to a subsample that did not develop coronary heart disease during a 6-year follow-up period. The data are given below.

Cholesterol	Blood Pressure (in mm Hg)				
(in mg/100 cc)	<127	127-146	147-166	167+	Totals
<200	117	121	47	22	307
200-219	85	98	43	20	246
220-259	119	209	68	43	439
≥260	67	99	46	33	245
Totals	388	527	204	118	1237

Consider four models:

Abbreviation	Model
[C][P][C_1]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + C_{1i}(j)$
[C][P][P_1]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + P_{1j}(i)$
[C][P][γ]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + \gamma(i)(j)$
[C][P]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)}$.

These are the row effects, column effects, uniform association, and independence models, respectively. The fits for the models relative to the saturated model are

Model	df	G^2	$A - q$
[C][P][C_1]	6	7.404	−4.596
[C][P][P_1]	6	5.534	−6.466
[C][P][γ]	8	7.429	−8.571
[C][P]	9	20.38	2.38

The best fitting model is

$$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + \gamma(i)(j)\,.$$

Using the side conditions $u_{C(1)} = u_{P(1)} = 0$, the parameter estimates and standard errors are

Parameter	Estimate	Standard Error
u	4.614	.0699
$u_{C(1)}$	0	—
$u_{C(2)}$	−0.4253	.1015
$u_{C(3)}$	−0.0589	.1363
$u_{C(4)}$	−0.8645	.1985
$u_{P(1)}$	0	—
$u_{P(2)}$	0.0516	.0965
$u_{P(3)}$	−1.164	.1698
$u_{P(4)}$	−1.991	.2522
γ	0.1044	.0293

The estimated cell counts are

Estimated Cell Counts: Uniform Association

	Blood Pressure			
Cholesterol	<127	127-146	147-166	167+
<200	112.0	131.0	43.1	20.9
200-219	81.3	105.4	38.5	20.8
220-259	130.1	187.4	76.0	45.5
≥260	64.5	103.2	46.4	30.8

These are obtained from the uniform association model, so the odds ratios for consecutive table entries are identical. For example, the odds of blood pressure < 127 relative to blood pressure 127-146 for men with cholesterol < 200 are 1.11 times the similar odds for men with cholesterol of 200-219; up to roundoff error

$$\frac{112.0/131.0}{81.3/105.4} = \frac{112.0(105.4)}{81.3(131.0)} = e^{.1044} = 1.11$$

where $.1044 = \hat{\gamma}$. Similarly, the odds of blood pressure 127-146 relative to blood pressure 147-166 for men with cholesterol < 200 are 1.11 times the odds for men with cholesterol of 200-219:

$$\frac{131.0(38.5)}{105.4(43.1)} = e^{.1044} = 1.11\,.$$

Also, the odds of blood pressure < 127 relative to blood pressure 127-146 for men with cholesterol 200-219 are 1.11 times the odds for men with cholesterol of 220-259:

$$\frac{81.3(187.4)}{130.1(105.4)} = e^{.1044} = 1.11\,.$$

For consecutive categories, the odds of lower blood pressure are 1.11 times greater with lower blood cholesterol than with higher blood cholesterol.

The asymptotic 95% confidence interval for γ has end points $.1044 \pm 1.96(.0293)$. The interval is $(.047, .162)$. The corresponding interval for the odds ratio is $(e^{.047}, e^{.162})$ or $(1.05, 1.18)$. Thus, for consecutive categories, we are 95% confident that the odds of lower blood pressure are between 1.05 and 1.18 times greater with lower cholesterol than with higher cholesterol.

Of course, we can also compare nonconsecutive categories. For categories that are one step away from consecutive, the odds of lower blood pressure are $1.23 = e^{2(.1044)}$ times greater with lower cholesterol than with higher cholesterol. For example, the odds of having blood pressure < 127 compared to having blood pressure of $147 - 166$ with cholesterol < 200 are $1.23 = e^{2(.1044)}$ times those for cholesterol $200 - 219$. To check this, observe that

$$\frac{112.0(38.5)}{81.3(43.1)} = 1.23.$$

Similarly, the odds of having blood pressure < 127 compared to having blood pressure of 127-146 with cholesterol < 200 are 1.23 times those for cholesterol 220-259. Extending this leads to observing that the odds of having blood pressure < 127 compared to having blood pressure of 167+ with cholesterol < 200 are $2.559 = e^{9(.1044)}$ times those for cholesterol ≥ 260.

It is of interest to compare the estimated cell counts obtained under uniform association with the estimated cell counts under independence. The estimated cell counts under independence are

Estimated Cell Counts: Independence

	Blood Pressure			
Cholesterol	<127	127-146	147-166	167+
<200	96.3	130.8	50.6	29.3
200-219	77.2	104.8	40.6	23.7
220-259	137.7	187.0	72.4	41.9
≥260	76.85	104.4	40.4	23.4

With $\gamma > 0$, the uniform association model increases the estimated cell counts (relative to independence) for cells with (a) high cholesterol and high blood pressure and (b) low cholesterol and low blood pressure. Also, the uniform association model decreases the estimated cell counts for cells with (a) high cholesterol and low blood pressure and (b) low cholesterol and high blood pressure.

7.2 Higher-Dimensional Tables

The same basic methods used to incorporate quantitative levels into two-factor models can also be used in higher dimensions. For example, consider

an $I \times J \times K$ table with quantitative levels $x_1, \ldots, x_I$, $w_1, \ldots, w_J$, and $v_1, \ldots, v_K$. Assuming no three-factor interaction, there are three types of models that are particularly useful.

The *homogeneous uniform association* model is a model in which given a level for any factor, the remaining two factors display a uniform association. This model is

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \beta_1 x_i w_j + \beta_2 x_i v_k + \beta_3 w_j v_k \,.$$

Note that if $x_i = i$, $w_j = j$

$$\begin{aligned}
&\log\left(\frac{m_{ijk} m_{i+1\ j+1\ k}}{m_{i+1\ jk} m_{i\ j+1\ k}}\right) \\
&= \quad u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \beta_1 x_i w_j + \beta_2 x_i v_k + \beta_3 w_j v_k \\
&\qquad + u + u_{1(i+1)} + u_{2(j+1)} + u_{3(k)} + \beta_1 x_{i+1} w_{j+1} \\
&\qquad\quad + \beta_2 x_{i+1} v_k + \beta_3 w_{j+1} v_k \\
&\qquad - u - u_{1(i+1)} - u_{2(j)} - u_{3(k)} - \beta_1 x_{i+1} w_j - \beta_2 x_{i+1} v_k - \beta_3 w_j v_k \\
&\qquad - u - u_{1(i)} - u_{2(j+1)} - u_{3(k)} - \beta_1 x_i w_{j+1} - \beta_2 x_i v_k - \beta_3 w_{j+1} v_k \\
&= \quad \beta_1 (x_i w_j - x_{i+1} w_j - x_i w_{j+1} + x_{i+1} w_{j+1}) \\
&= \quad \beta_1 \,.
\end{aligned}$$

Thus, for any level of k, the log odds ratio for consecutive table entries equals β_1. Similarly, for j fixed, consecutive log odds ratios equals β_2; and for i fixed, consecutive log odds ratios equal β_3.

If Factor 3 does not have quantitative levels or if we merely wish to ignore the quantitative nature of the levels of Factor 3, we can write

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \tau_{1k} x_i + \tau_{2k} w_j + \beta x_i w_j \,.$$

For each level of k, this model gives uniform associations. For a fixed level of i or a fixed level of j, odds ratios need not display uniform association.

If neither Factors 2 or 3 has quantitative levels or if we wish to ignore their quantitative nature, we can use the model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} + \tau_{2j} x_i + \tau_{3k} x_i \,.$$

As before, models can be generalized by including powers of the x_i, w_j, and v_k scores. We can also model the three-factor interaction. The models

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + \beta x_i w_j v_k$$

and

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \beta_1 w_j v_k + \beta_2 x_i v_k + \beta_3 x_i w_j + \gamma x_i w_j v_k$$

deal with the three-factor interaction while modeling two-factor interactions in alternative ways. Both of these models could be described as *heterogeneous uniform association* models.

EXAMPLE 7.2.1. In Chapter 6, we found that for the race, sex, opinion, age data, the model [RSO][OA] fits well. The ages are quantitative levels. We consider whether using the quantitative nature of this factor leads to a more succinct model. The age categories are 18-25, 26-35, 36-45, 46-55, 56-65, and 66+. For lack of a better idea, the category scores were taken as 1, 2, 3, 4, 5, and 6. Since the first and last categories are different from the other four, the use of the scores 1 and 6 are particularly open to question. Two models were considered:

Abbreviation	Model
[RSO][OA]	$\log(m_{hijk}) = u_{RSO(hij)} + u_{OA(jk)}$
[RSO][A][O_1]	$\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k$
[RSO][A][O_1][O_2]	$\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k + O_{2j}k^2$

Both of these are reduced models relative to [RSO][OA]. ([RSO][OA] is equivalent to $\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k + O_{2j}k^2 + O_{3j}k^3 + O_{4j}k^4 + O_{5j}k^5$.) To compare models, we need the following statistics

Model	df	G^2
[RSO][OA]	45	24.77
[RSO][A][O_1][O_2]	51	26.99
[RSO][A][O_1]	53	29.33

Comparing [RSO][A][O_1] versus [RSO][OA] gives $G^2 = 29.33 - 24.77 = 4.56$ with degrees of freedom $53 - 45 = 8$. The G^2 value is not significant. Similarly, [RSO][A][O_1][O_2] is an adequate model relative to [RSO][OA]. The test for [O_2] has $G^2 = 29.33 - 27.99 = 1.34$ on 2 *df*, which is not significant. The model with only [O_1] fits the data well.

A primary difficulty with using quantitative factors is the necessity of assigning the factor scores. One way to avoid this problem is to estimate the factor scores. Methods for doing this are discussed in Section 3.

7.2.1 Computing Commands

Models with quantitative factors can be fit easily using several computer packages, e.g., SPLUS, GLIM, GENSTAT, and SAS PROC GENMOD. For example, the model $\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k + O_{2j}k^2$ can be fitted using SAS PROC GENMOD as given below.

```
options ps=60 ls=72 nodate;
data abort;
   infile 'abort.dat';
   input R S A O N;
   A1 = A; A2 = A * A;
proc genmod data=abort;
```

```
   class R S O A;
   model N = R*S*O A O*A1 O*A2 / link=log
                                 dist=poisson;
run;
```

The key difference here from analysis of variance type models is that in the "class" command, age was not specified as a grouping (class) variable. The terms $O_{1j}k$ and $O_{2j}k^2$ are really interactions between the factor O and the predictors variables age and age squared. In the model statement, they are simply specified as interactions.

7.3 Unknown Factor Scores

The next step in generalizing the models of Sections 1 and 2 is to allow the factor scores to be unknown. In place of model (7.1.4), we assume the model

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma\nu_i\omega_j \tag{1}$$

where $\nu_i,\ i = 1, \ldots, I$, and $\omega_j,\ j = 1, \ldots, J$, are unknown parameters with

$$\sum_{i=1}^{I} \nu_i^2 = 1 = \sum_{j=1}^{J} \omega_j^2 \tag{2}$$

and

$$\alpha_\cdot = \beta_\cdot = \nu_\cdot = \omega_\cdot = 0.$$

The side conditions are imposed because the model is no longer log-linear. In general, for nonlinear models, the exact parametrization can be important in determining the properties of the model. The side conditions are necessary to have a well-defined parametrization. In particular, without condition (2), the parameter γ would not be well defined.

Model (1) is not log-linear, so the theoretical results used to justify fitting log-linear models do not apply. A separate theoretical development is required. Moreover, computer programs specifically developed for fitting log-linear models by maximum likelihood cannot be used to obtain the maximum likelihood fit of the model. Chuang (1983) used iteratively reweighted nonlinear least squares to obtain maximum likelihood estimates for models with unknown factor scores.

In addition to model (1), it is of interest to examine reduced models. In particular, the submodel of (1) with column main effects that are linear in the unknown factor scores is

$$\log(m_{ij}) = \mu + \alpha_i + \lambda\omega_j + \gamma\nu_i\omega_j \tag{3}$$

where

$$\alpha_\cdot = \nu_\cdot = \omega_\cdot = 0. \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = \sum_{j=1}^{J} \omega_j^2 = 1 \; .$$

Model (3) can also be written as

$$\log(m_{ij}) = \mu + \alpha_i + \nu_i \omega_j \tag{4}$$

with

$$\alpha_\cdot = \omega_\cdot = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \omega_j^2 = 1 \; .$$

Here, ν_i is equivalent to $\lambda + \gamma\nu_i$ in model (3), which is why the conditions $\nu_\cdot = 0$ and $\sum_{i=1}^{I} \nu_i^2 = 1$ are dropped. We can take model (4) one step further and write it as

$$\log(m_{ij}) = \alpha_{0i} + \alpha_{1i}\omega_j$$

where

$$\omega_\cdot = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \omega_j^2 = 1 \; .$$

The new parametrization is related to model (4) by $\alpha_{0i} \equiv \mu + \alpha_i$ and $\alpha_{1i} = \nu_i$. This version of the linear column effects model has the nice interpretation of *fitting separate lines in the unknown factor scores for each level of i*.

Similarly, if row effects are linear in the factor scores, the appropriate model is

$$\log(m_{ij}) = \mu + \lambda\nu_i + \beta_j + \gamma\nu_i\omega_j$$

with

$$\beta_\cdot = \nu_\cdot = \omega_\cdot = 0 \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = 1 = \sum_{j=1}^{J} \omega_j^2 \; .$$

This is equivalent to

$$\log(m_{ij}) = \mu + \beta_j + \nu_i\omega_j \tag{5}$$

and also to the separate lines model

$$\log(m_{ij}) = \beta_{0j} + \beta_{1j}\nu_i$$

where

$$\nu_\cdot = 0 \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = 1$$

in both models and $\beta_\cdot = 0$ in model (5).

If both main effects are linear, the model is

$$\log(m_{ij}) = \mu + \lambda_1\nu_i + \lambda_2\omega_j + \gamma\nu_i\omega_j$$

with

$$\nu_\cdot = \omega_\cdot = 0 \qquad \text{and} \qquad \sum_{i=1}^{I}\nu_i^2 = \sum_{j=1}^{J}\omega_j^2 = 1\,.$$

An equivalent model is

$$\log(m_{ij}) = \mu + \nu_i + \omega_j + \gamma\nu_i\omega_j \tag{6}$$

where

$$\nu_\cdot = \omega_\cdot = 0\,.$$

The relationship between the models is based on ν_i being equivalent to $\lambda_1\nu_i$, ω_j being equivalent to $\lambda_2\omega_j$, and γ being equivalent to $\gamma/\lambda_1\lambda_2$.

When the factor categories are known to be ordered, a corresponding order can be imposed on the estimated factor scores. For example, models (1), (4), (5), and (6) can be fitted subject to the condition that $\nu_1 \le \nu_2 \le \cdots \le \nu_I$. Similar orderings can be imposed on the column scores. Unfortunately, such constraints cause complications in the numerical procedures required to fit the models. In practice, it seems to be more common to fit the models without imposing order conditions on the scores. One can then verify whether the data are consistent with the a priori ordering.

Model (1) was first proposed by Fienberg (1968). Later, Goodman (1979, 1981), Anderson (1980), and Chuang (1983) extended the use of estimated factor scores. Johnson and Graybill (1972) proposed a model similar to (1) for standard analysis of variance. They built on earlier results in analysis of variance that are also of interest for log-linear models.

Tukey (1949) proposed a 1 degree of freedom test for nonadditivity in a two-way analysis of variance. Mandel (1961, 1971) extended Tukey's results by considering tests for more general models and presented a justification of the models based on unknown factor scores. Mandel's models differ from those considered by Johnson and Graybill in that they use the row and column effects in place of the unknown factor scores. These models and their justification apply equally well to log-linear models.

Mandel's models are

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \delta_i\beta_j, \qquad \alpha_\cdot = \beta_\cdot = 0\,, \tag{7}$$

and

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \phi_j\alpha_i, \qquad \alpha_\cdot = \beta_\cdot = 0\,. \tag{8}$$

The *Tukey model* is

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j, \qquad \alpha_\cdot = \beta_\cdot = 0\,. \tag{9}$$

It is easily seen that model (7) is equivalent to the linear row effects model (4). By equating

$$\beta_j = \omega_j$$

and

$$\delta_i = \nu_i - 1$$

and substituting into (4), we see that model (4) can be written as model (7). Conversely, if model (7) holds, write

$$\omega_j = \beta_j \Big/ \sqrt{\sum \beta_j^2}$$

and

$$\nu_i = (1 + \delta_i)\sqrt{\sum \beta_j^2}$$

to see that model (4) holds. Similarly, models (5) and (8) are equivalent and models (6) and (9) are identical. This establishes the justification for examining models (7), (8), and (9). They are equivalent to models with interesting interpretations in terms of underlying unknown factor scores.

If these models are appropriate and necessary, the maximum likelihood fits should be obtained. Maximum likelihood estimation and (generalized) likelihood ratio tests involving any of the models for unknown factor scores require specialized methods for fitting the log-nonlinear models. Standard programs for fitting log-linear models are not appropriate. However, by analogy with the two-stage fitting procedure commonly used in analysis of variance, a simple method can be derived for evaluating whether these models are necessary. All of the models considered contain the model of complete independence

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j \tag{10}$$

as a submodel. This may not be obvious in models (4), (5), and (6), but it is in their equivalent versions (7), (8), and (9).

Models (7), (8), and (9) can be tested against model (10) in a very simple way. First, write

$$\tau_{ij} = \mu + \alpha_i + \beta_j$$

and note that if model (10) is true, (7) is equivalent to

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \delta_i \tau_{ij}\,, \tag{11}$$

(8) is equivalent to

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \phi_j \tau_{ij}\,, \tag{12}$$

and (9) is equivalent to

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma(\tau_{ij})^2\,. \tag{13}$$

Models (11), (12), and (13) would be log-linear if the τ_{ij}'s were known. Fit model (10) by maximum likelihood to obtain $\hat{\tau}_{ij}$. Substitute $\hat{\tau}_{ij}$ for τ_{ij} in models (11), (12), and (13) so that they are log-linear in the other parameters. Fit the models based on $\hat{\tau}_{ij}$ using standard log-linear methods and test them against model (10) in the usual way. If model (10) is true, these tests have asymptotic chi-squared distributions. For linear models, the validity of tests based on this two-stage fitting procedure was established by Milliken and Graybill (1970) and Rao (1965). For log-linear models, a corresponding result is given by Christensen and Utts (1992).

These models and methods can also be applied to higher-dimensional tables and to logit models. Example 7.4.1 gives the details for fitting a logit model.

EXAMPLE 7.3.1. Wing (1962), Haberman (1974b), and Fienberg (1980) have considered data on the relationship between length of hospitalization and frequency of visits for 132 long-term schizophrenic patients. Length of hospitalization was categorized as over 2 years but under 10, $(2, 10)$, over 10 years but under 20, $(10, 20)$, and over 20 years, 20+. Frequency of visits were regular, irregular (no home visits, hospital visits less than once a month), and never. The data are given in Table 7.1.

TABLE 7.1. Schizophrenic Data

Visitation Frequency (i)	Length of Hospitalization (j) in years $(2, 10)$	$(10, 20)$	20+
Regular	43	16	3
Irregular	6	11	10
Never	9	18	16

Fitting model (10) in the usual way and using the two-stage fitting procedure described above for the other models yields the lack of fit statistics given below.

Model	df	Two-Stage G^2
(10)	4	38.35
(11)	2	1.21
(12)	2	11.18
(13)	3	14.76

Except for model (10), these G^2's are not likelihood ratio lack of fit statistics for the models. However, the G^2 for model (10) can be subtracted from the other G^2's to obtain valid asymptotic χ^2 tests for model (10) versus the other models. The test statistics are as follows:

Model	df	Two-Stage G^2
(11)	2	37.14
(12)	2	27.17
(13)	1	23.59

All of the models fit better than (10), especially model (11).

One set of parameter estimates are given below for the two-stage fit of model (11). Recall that parameter estimates are not uniquely defined.

Parameter	Estimate
μ	-12.91
α_1	0.000
α_2	14.65
α_3	15.23
β_1	0.000
β_2	0.4727
β_3	0.4977
δ_1	5.034
δ_2	0.05197
δ_3	0.000

Observe that $\hat{\delta}_1 > \hat{\delta}_2 > \hat{\delta}_3 = 0$ with $\hat{\delta}_1$ much larger than the others. To draw conclusions about the meaning of the $\hat{\delta}$'s, we need to examine their multipliers in model (11), the $\hat{\tau}_{ij}$'s. The $\hat{\tau}_{ij}$'s are

	i		
j	1	2	3
1	3.305	3.051	2.612
2	2.473	2.220	1.780
3	2.939	2.685	2.246

In each row, the $\hat{\tau}_{ij}$'s are decreasing. With non-negative $\hat{\delta}$'s, as we move to the right in each row, the fitted counts decrease. For patients who are visited irregularly or never, the model indicates little decrease over time due to the interaction because the $\hat{\delta}$ values are near zero. However, $\hat{\delta}$ is large for row 1, so, according to the fitted model, the number of patients who have regular visits decreases dramatically over time.

Exercise 7.1. Show that models (7) and (11) are equivalent. Show that models (9) and (13) are equivalent.

7.4 Logit Models

Results analogous to Section 3 apply to models for the log odds. The example in this section involves a logit model with factors that are assumed to have unknown quantitative category scores.

EXAMPLE 7.4.1. Rosenberg (1962) presents data on the relationships among Religion, Father's Educational Level, and Self-Esteem. The data are given in Table 7.2. Self-Esteem is considered the response.

TABLE 7.2. Data of Rosenberg (1962).

		Father's Educational Level					
Religion	Self-Esteem	8th or less	Some HS	HS Grad	Some Coll	Coll Grad	Post Coll
Catholic	High	245	330	388	100	77	51
	Low	115	152	153	40	37	19
Jewish	High	28	89	102	67	87	62
	Low	11	37	35	18	12	13
Protestant	High	125	234	233	109	197	90
	Low	68	91	173	47	82	32

Chuang (1983) presents a maximum likelihood analysis of Rosenberg's data that includes logit versions of Mandel's models (7.3.7) and (7.3.8) and the Tukey model (7.3.9). The expected cell counts are m_{ijk}, where i denotes religion, j denotes educational level, and k denotes self-esteem. A logit model imposes structure on

$$\tau_{ij} \equiv \log(m_{ij1}/m_{ij2})\,.$$

The logit versions of Mandel's models are

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \delta_i\beta_j \tag{1}$$

and

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \phi_j\alpha_i\,. \tag{2}$$

The logit version of the Tukey model is

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j\,. \tag{3}$$

The additive model is

$$\tau_{Xij} = \mu + \alpha_i + \beta_j\,. \tag{4}$$

In Section 3, the models (7.3.11), (7.3.12), and (7.3.13) were used to simplify the two-stage fitting process. Their logit model analogues are

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \delta_i \tau_{Xij}\,, \tag{5}$$

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \phi_j \tau_{Xij}\,, \tag{6}$$

and

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \gamma(\tau_{Xij})^2\,. \tag{7}$$

As in Section 3, these models are equivalent to models (1), (2), and (3), respectively.

To fit (5), (6), and (7) using maximum likelihood requires the use of something other than a standard log-linear model or logistic model computer program. Chuang (1983) suggests modifying a nonlinear least squares program. The alternative two-stage procedure discussed in Section 3 is easily implemented with standard software. Begin by fitting model (4) to obtain the $\hat{\tau}_{Xij}$'s. The only nonlinear aspect to models (5), (6), and (7) is the presence of the τ_{Xij} parameters, so if these parameters were known, there would be no difficulty in obtaining fits for the models. In the two-stage procedure, the estimates $\hat{\tau}_{Xij}$ are substituted for the parameters. The resulting linearized models are fitted in the usual way with standard software. Test statistics for comparing the models to model (4) are also computed in the usual way.

Table 7.3 presents the results of fitting models (5), (6), and (7) by both maximum likelihood and the two-stage procedure. The G^2 values reported in Table 7.3 for the two-stage fits are not directly applicable because they are statistics for testing the models against the saturated model. The theoretical justification given in Christensen and Utts (1992) applies only to testing models (5), (6), and (7) against the additive model (4). Although the G^2's reported in Table 7.3 do not have a sound theoretical basis as test statistics, they are sometimes a valuable data analytic tool. This is not unreasonable because they are one-to-one functions of test statistics that have a sound basis.

TABLE 7.3. Model Fits

Model	df	MLE G^2	Two-Stage G^2
(5)	8	12.76	16.34
(6)	5	12.07	13.25
(7)	9	25.58	26.34
(4)	10	26.39	—

The results of testing the models with nonadditivity against the additive model are given in Table 7.4. In this example, the two-stage tests appear to be a little less powerful than the generalized likelihood ratio tests, but

the qualitative conclusions about the best-fitting model are identical for the two methods. Model (5), i.e., model (1), appears to be the best-fitting model. Recall from Section 3 that this is the model that, for each religion, fits a line in the unknown factor scores associated with Father's Educational Level.

TABLE 7.4. Tests of Nonadditivity

Model	*df*	MLE G^2	Two-Stage G^2
(5)	2	13.63	10.05
(6)	5	14.32	13.14
(7)	1	0.81	0.05

If one of the log-nonlinear models is to be used in further work, an exact maximum likelihood fit of the model should be used. Nonetheless, the two-stage fitting procedure provides a simple yet valid diagnostic tool for checking whether Mandel's models and the Tukey model require more investigation.

7.5 Exercises

EXERCISE 7.5.1. Reanalyze the Intelligence versus Clothing table of Exercise 2.6.3 using the methods of Section 1. Note that this ignores the potentially complicating factor Standard and the complex sampling scheme.

EXERCISE 7.5.2. For the model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + \gamma_1 \, i \, j + \gamma_2 \, i^2 j \, ,$$

find the log odds ratio

$$\log[m_{ij} m_{i+1\ j+1} / m_{i+1\ j} m_{i\ j+1}]$$

in terms of the model parameters.

EXERCISE 7.5.3. Duncan, Schuman, and Duncan (1973) and Duncan and McRae (1979) present data on evaluations made in 1959 and 1971 of the performance of radio and TV networks. The data are given in Table 7.5. Use the methods of Section 2 to analyze these data.

EXERCISE 7.5.4. Assuming the use of consecutive integer scores, find the log odds ratios

$$\log[m_{ijk} m_{i+1\ j+1\ k} / m_{i+1\ jk} m_{i\ j+1\ k}]$$

TABLE 7.5. Radio and Television Network Performance.

	Respondent's	Performance of Networks		
Year	Race	Poor	Fair	Good
1971	White	158	636	600
	Black	24	144	224
1959	White	54	253	325
	Black	4	23	81

and

$$\log[m_{ijk}m_{i+1\ j\ k+1}/m_{i+1\ jk}m_{i\ j\ k+1}]$$

in terms of the model parameters for the two heterogeneous uniform association models of Section 2.

Exercise 7.5.5. Use the methods of Section 3 to analyze the Intelligence – Clothing table of Exercise 2.6.3.

8
Fixed and Random Zeros

Not infrequently, one encounters a table in which a number of the cell counts are 0s. These cells sometimes cause problems when fitting log-linear models. Recall that in our discussion of multiple logistic regression in Section 4.1, we had a 200×2 table that was riddled with zero counts. Except for the fact that some asymptotic results did not hold, the zeros caused no problems. Cells with zero counts merely have the potential to cause problems.

Cells with zero counts are classified in two ways: fixed and random. *Fixed zeros* are cells in which it is impossible to observe counts. Such cells must always be zero. *Random zeros* are cells that happen to have zero counts, but where it is possible to have positive counts. Tables with fixed zeros are called *incomplete tables*.

8.1 Fixed Zeros

EXAMPLE 8.1.1. Brunswick (1971) reports data on the health concerns of teenagers. These data have also been examined by Grizzle and Williams (1972) and Fienberg (1980). The data are given in Table 8.1. The two zeros in the table are fixed. It is physiologically impossible for males (of whatever age) to have menstrual difficulties. (Technically, I suppose the real issue here is whether males worry about menstrual difficulties. I suppose somewhere, sometime, some teenage male has worried about them, but one must admit that these are very nearly fixed zeros. We will treat them as such.)

TABLE 8.1. Health Concerns of Teenagers

		Health Concerns (k)			
Sex (i)	Age (j)	Sex, Reproduction	Menstrual Problems	How Healthy I Am	Nothing
Male	12-15	4	0	42	57
	16-17	2	0	7	20
Female	12-15	9	4	19	71
	16-17	7	8	10	31

The solution to dealing with fixed zeros is to throw them away. They are cells that do not really exist. One simply ignores those cells and fits a model to the cells that do exist. Thus, the model is fitted to an incomplete table.

In fact, it is impossible to include fixed zeros in a log-linear model. A log-linear model specifies a value of $\log(m_i)$ for every cell i. It is implicit that $\log(m_i)$ is defined. If a cell is a fixed zero, the probability of an observation occurring in that cell is zero, so the expectation, m_i, must also be zero. In such cases, $\log(m_i)$ is undefined. To fit log-linear models, one has to throw fixed zeros away.

If the Newton-Raphson algorithm of Chapter 10 is used to fit log-linear models, throwing out fixed zeros is no problem. The algorithm can easily handle the fact that not all combinations of the factor categories are considered in the model.

When using iterative proportional fitting, the situation is slightly more complex. The algorithm is based on having all combinations of the factor categories defined. Fortunately, there is a simple way around this problem. Recall that it is standard to start iterative proportional fitting with all initial cell estimates equal to 1 and that if the initial values satisfy the constraints of the model, the subsequent iterations also satisfy the constraints of the model. With all initial values of 1, the initial values satisfy the constraints of any interesting ANOVA model. To deal with fixed zeros, take the initial values of the corresponding cells to be 0. It is easily seen that if the initial value is 0, then all subsequent values will also be 0. (We use the definition that $0/0 = 0$.) Moreover, the constraints on the model with fixed zero cells eliminated are identical to the constraints on the complete model when the fixed zero cells are required to have fitted values of 0.

Although one can get the correct fitted values using iterative proportional fitting, the user often must provide the correct degrees of freedom. These are determined by the model with the fixed zero cells eliminated.

EXAMPLE 8.1.2. We illustrate the computation of degrees of freedom using the data of Example 8.1.1. Consider the model

$$\log(m_{ijk}) = M + S_i + A_j + H_k + (SA)_{ij} + (SH)_{ik} + (AH)_{jk}\,. \quad (1)$$

Clearly, the grand mean M can exist and we have data available for each sex, age, and health concern. Computing these degrees of freedom as usual, we have

Term	*df*
M	1
S	1
A	1
H	3

We also have data available for every combination of sex and age, and every combination of age and health concern. Again, computing degrees of freedom in the usual way, we get

Term	*df*
SA	1
AH	3

We do not have data available for every combination of sex and health concern. We have no data for males with menstrual difficulties. In the 2×4 table of sex and health concerns, we have only 7 cells instead of 8. The degrees of freedom for this table must be 7. To fit this table perfectly, we would use the parameters M, S_i, H_k, and $(SH)_{ik}$. The total number of degrees of freedom for these parameters must be 7. Because M has 1 degree of freedom, S has 1, and H has 3, that leaves only 2 degrees of freedom available for SH. Thus, model (1) has

$$1 + 1 + 1 + 3 + 1 + 2 + 3 = 12$$

degrees of freedom. The saturated model has 1 degree of freedom for each cell. There are nominally 16 cells, but 2 are fixed zeros, so the table has only 14 cells. For testing model (1) against the saturated model, the degrees of freedom are $14 - 12 = 2$.

These techniques are easily applied to all models for the health concern data. In testing [SA][H] against the saturated model, we have dropped the SH and AH terms. The degrees of freedom for these terms are put into the test degrees of freedom. Model (1) has 2 degrees of freedom for the test, SH has 2 degrees of freedom, and AH has 3 degrees of freedom, so the test of [SA][H] has $2 + 2 + 3 = 7$ degrees of freedom. Alternatively, we can do the calculation by noting that the saturated model has 14 *df*, while the [SA][H] model has terms and degrees of freedom $M(1)$, $S(1)$, $A(1)$, $SA(1)$, $H(3)$ for a total of 7 degrees of freedom. The test has $14 - 7 = 7$ degrees of freedom.

The fits for these data are summarized below.

Model	*df*	G^2
[SA][SH][AH]	2	2.03
[SH][AH]	3	4.86
[SA][AH]	4	13.45
[SA][SH]	5	9.43
[SA][H]	7	22.03
[SH][A]	6	15.64
[AH][S]	5	17.46
[S][A][H]	8	28.24

The best fitting model is either [SA][SH][AH] or [SH][AH], depending on whether health concerns are treated as a response variable or not.

One final note on fixed zeros. Because fixed zeros are really cells that do not exist, the existence of fixed zeros has no effect on the validity of large sample results. If the counts in the other cells are large, asymptotic results hold.

8.2 Partitioning Polytomous Variables

We now investigate a method that uses incomplete tables to examine category effects.

EXAMPLE 8.2.1. Duncan (1975) presents data on the earth-shattering question, "Who should shovel the snow from sidewalks?" The data are given in Table 8.2. Mothers were asked whether boys, girls, or both should do the shoveling. Mothers never responded that girls alone should do the shoveling, so the data are presented with only two categories. In addition, there are two explanatory factors, the mother's religion (R), Protestant, Catholic, Jewish, or other, and the year (Y) in which the question was asked. (Having lived 36 years in Minnesota and Montana, I am aware that a key factor has been left out. Father does the vast majority of the shoveling.)

We will treat mothers' opinions on shoveling (S) as a response variable and consider only log-linear models that correspond to logit models. The point of this example is to illustrate how to use tables with fixed zeros to answer questions about parameters in log-linear models.

Fitting the standard models gives

Model	*df*	G^2
[RY][RS][YS]	3	0.4
[RY][RS]	4	21.5
[RY][YS]	6	11.2
[RY][S]	7	31.7

TABLE 8.2. Mothers' Opinions on Who Should Shovel Snow

		Shoveling (k)	
Religion (i)	Year (j)	Boy	Both
Protestant	1953	104	42
	1971	165	142
Catholic	1953	65	44
	1971	100	130
Jewish	1953	4	3
	1971	5	6
Other	1953	13	6
	1971	32	23

Clearly, the best fitting model is [RY][RS][YS]. We can write the model as

$$\log(m_{ijk}) = (RY)_{ij} + S_k + (RS)_{ik} + (YS)_{jk}\,, \tag{1}$$

$i = 1,2,3,4$, $j = 1,2$, $k = 1,2$. Because S is a response variable, the $(RY)_{ij}$'s must be in the model. The important terms are S_k, $(RS)_{ik}$, and $(YS)_{jk}$. In a logit model, S_k corresponds to the grand mean; not very interesting. The terms $(YS)_{jk}$ correspond to main effects in years with 1 degree of freedom. The terms $(RS)_{ik}$ correspond to main effects in religion with 3 degrees of freedom. Further analysis must examine the nature of the three degrees of freedom in (RS). We are really asking about relationships among the religion categories.

One way to proceed was illustrated in Section 4.6. We could incorporate constraints on the religions. For instance, we could treat all Protestants and Jews alike, while allowing Catholics and Others to have separate effects on the shoveling response. Such a procedure would involve recoding the indices. We could recode the index i into a new pair of indices g and h as follows:

i	1	2	3	4
(g,h)	(1,1)	(2,1)	(1,2)	(3,1)

where g indicates religion with no difference between Protestants and Jews, while h is simply used to tell Protestants and Jews apart. We can rewrite model (1) as

$$\log(m_{ghjk}) = (RY)_{ghj} + S_k + (RS)_{ghk} + (YS)_{jk}\,. \tag{2}$$

To eliminate differences between Protestants and Jews in (RS), we drop the h, giving

$$\log(m_{ghjk}) = (RY)_{ghj} + S_k + (RS)_{gk} + (YS)_{jk}\,. \tag{3}$$

Both models (2) and (3) are actually models for incomplete tables. The possible values for g are 1, 2, 3. For h, they are 1, 2. For j and k, they are also 1 and 2. This suggests the existence of a $3 \times 2 \times 2 \times 2$ table. However, not all combinations of the indices are possible. Only when g is 1 can h be 2. Any cell $(g, 2, j, k)$ in the $3 \times 2 \times 2 \times 2$ table with $g = 2$ or 3 is a fixed zero.

It is obvious that quite a few questions can be addressed by creative reindexing. Duncan (1975) presented a particular pattern that is very flexible. Transform the index i into (r, s, t, u) where the correspondence is as follows:

i	1	2	3	4
(r, s, t, u)	(1,2,2,2)	(2,1,2,2)	(2,2,1,2)	(2,2,2,1)

Each 4-tuple has three 2s and one 1. If the 1 is in the third place, then the 4-tuple corresponds to $i = 3$, etc. The index r is 1 for Protestant and 2 otherwise. The index s is 1 for Catholic and 2 otherwise. Similarly, $t = 1$ indicates Jewish and $u = 1$ indicates Other. Including the year and shoveling factors, we now have a $2 \times 2 \times 2 \times 2 \times 2 \times 2$ table with many fixed zeros because mothers have only one religion. Using a natural identification, denote the factors corresponding to r, s, t, and u as P, C, J, and O. Model (1) can now be rewritten as

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (PCJOS)_{rstuk} + (YS)_{jk} .$$

Moreover, we can denote this as [PCJOY][PCJOS][YS]. Note that the four factors P, C, J, O taken together are equivalent to the old factor R.

Now consider the model [PCJOY][YS][CS]. Recall that [PCJOY] is included because S is a response factor. The term [YS] seemed important in our original analysis, so it is retained. The new model replaces the terms $(RS)_{ik} = (PCJOS)_{rstuk}$ in model (1) with the terms $(CS)_{sk}$. In particular, the model is

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (CS)_{sk} + (YS)_{jk} .$$

The logit model effect of the four different religions is being replaced with one effect that distinguishes Catholics from all other religions.

Earlier, we considered a model that treated Protestants and Jews the same, but allowed separate effects for Catholics and Others. In Duncan's setup, this is the model [PCJOY][YS][COS]. In the model

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (COS)_{tuk} + (YS)_{jk} ,$$

the effect of religion on mothers' shoveling opinions is taken up by the $(COS)_{tuk}$ terms. Only three such terms exist: $(COS)_{12k}$, an effect for Catholics; $(COS)_{21k}$, an effect for others; and $(COS)_{22k}$. The $(COS)_{22k}$ term does not distinguish between Protestants and Jews.

Moreover, because the $2 \times 2 \times 2 \times 2 \times 2 \times 2$ table is so very incomplete, there is another way to do exactly the same thing. The model [PCJOY][YS][CS][OS] is equivalent to [PCJOY][YS][COS]. The model for [PCJOY][YS][CS][OS] is

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (CS)_{sk} + (OS)_{uk} + (YS)_{jk} \; .$$

Catholics get the effects $(CS)_{1k} + (OS)_{2k}$. Others get the distinct effects $(CS)_{2k} + (OS)_{1k}$. Protestants and Jews both get $(CS)_{2k} + (OS)_{2k}$.

In a complete table, the 15 interactions between S and the religion factors P, C, J, and O would each have 1 degree of freedom. In fact, there are only 3 degrees of freedom for $(RS) = (PCJOS)$. Thus, there are a lot of redundant models. In fact, any term that includes three of P, C, J, and O is equivalent to any other term that includes three, and these are all equivalent to the four-factor term. This follows from the fact that any three of the indices r, s, t, and u completely determine the cell; e.g., if $r = 2$, $s = 2$, $t = 2$, then we must have $u = 1$. Because all terms with three of the factors are the same, models such as [PCJOY][YS][PCJS] and [PCJOY][YS][CJOS] are equivalent to each other and to [PCJOY][YS][PCJOS].

Although Duncan's method of reindexing is flexible, it is not a panacea. There are interesting questions that cannot be addressed using Duncan's method. For example, if we wanted a model that treated Protestant and Jews the same and also treated Catholics and Others the same, we could not arrive at such a model using Duncan's method.

Now let's see where Duncan's method gets us with the shoveling data. Some models and fits are given below:

Model	*df*	G^2
[PCJOY][PCJOS][YS] = [RY][RS][YS]	3	0.4
[PCJOY][PS][YS]	5	4.8
[PCJOY][CS][YS]	5	1.4
[PCJOY][JS][YS]	5	10.9
[PCJOY][OS][YS]	5	9.8
[PCJOY][YS] = [RY][YS]	6	11.2

The models that involve [JS] and [OS] do not fit very well. The model with [JS] only distinguishes between Jews and non-Jews. The relatively poor fit indicates that Protestants, Catholics, and Others do not act the same. Similarly, the relatively poor fit of the model with [OS] indicates that Protestants, Catholics, and Jews do not act the same.

Both the models with [PS] and [CS] fit reasonably well. The model with [CS] fits especially well. This model suggests that Protestants, Jews, and Others act the same, but Catholic mothers have different opinions about who should shovel snow. The model with [PS] indicates that Catholics, Jews, and Others act the same, but Protestant mothers have different opinions. There are so few Jews and Others that it is not surprising that lumping

them with either large category does not substantially hurt the fit. What we really know is that Protestants mothers have different attitudes than Catholic mothers and that if we want to lump Jewish and Other mothers with one of those categories, they seem to fit better with the Protestants than the Catholics. However, we know almost nothing about Jewish mothers and little about Other mothers. From looking at Table 8.2, it is clear that the Catholic mothers are much more egalitarian about shoveling than the Protestants or the Others.

We could go further with this analysis by considering models [RY][YS][CS][PS], [RY][YS][CS][JS], and [RY][YS][CS][OS], but considering what little difference there is between [RY][YS][CS] and [RY][YS][RS], there seems little point in pursuing the analysis further. Note that I have gotten lazy and started writing [RY] for [PCJOY].

8.3 Random Zeros

Random zeros are cells that have positive probability of occurring, but do not occur in the sample at hand. If a large enough sample was taken, these cells should eventually contain positive counts.

Random zeros present three problems. First, they suggest that asymptotic results are invalid. If the sample was large enough for asymptotic approximations, these cells ought not be zero. The discussions of small samples in Section 2.4 and conditional inference in Section 3.5 are relevant to this problem. Second, when a table includes random zeros, maximum likelihood estimates of the parameters may not exist. Finally, there is a practical problem in that some computer programs will give "MLEs" even when they do not exist. In this section, we concern ourselves primarily with problems related to the nonexistence of MLEs.

EXAMPLE 8.3.1. Consider a $4 \times 2 \times 3 \times 3$ table on the results of arthroscopic knee surgery. The four factors and their categories are listed as follows:

Factor Label	Factor Description	Categories
Type	Type of injury	Twist, Direct, Both, No injury
Sex	Sex of patient	Male, Female
Age	Age of patient	11-30, 31-50, 51-91
Result	Outcome of surgery	Excellent, Good, Fair-Poor

The data are given in Table 8.3. First, note that one cannot fit a saturated log-linear model. For a saturated model, $n_{hijk} = \hat{m}_{hijk}$. Because the model is log-linear, $\log(\hat{m}_{hijk})$ must be defined. However, for some cells, $n_{hijk} = 0$, so either $n_{hijk} \neq \hat{m}_{hijk}$ or $\log(\hat{m}_{hijk})$ is not defined.

TABLE 8.3. Data on Arthroscopic Knee Surgery

Type (h)	Sex (i)	Age (j)	Result (k)		
			Ex	Good	F-P
		11-30	21	11	4
	Male	31-50	32	20	5
		51-91	20	12	5
Twist					
		11-30	3	1	0
	Female	31-50	6	5	2
		51-91	6	3	1
		11-30	3	2	2
	Male	31-50	2	4	4
		51-91	0	0	0
Direct					
		11-30	0	1	1
	Female	31-50	0	0	0
		51-91	1	2	3
		11-30	7	1	1
	Male	31-50	11	6	2
		51-91	0	4	6
Both					
		11-30	1	0	0
	Female	31-50	1	1	1
		51-91	2	4	1
		11-30	0	0	0
	Male	31-50	1	2	1
		51-91	3	3	0
No Injury					
		11-30	1	0	0
	Female	31-50	1	2	0
		51-91	1	6	8

We can extend this argument to other models. If we consider models that include the $u_{TSA(hij)}$ terms (i.e., models that include [TSA]), the MLEs must satisfy the condition $n_{hij\cdot} = \hat{m}_{hij\cdot}$ for all h, i, and j. However, $n_{213\cdot} = n_{222\cdot} = n_{411\cdot} = 0$. If MLEs exist, then $\hat{m}_{213\cdot} = 0$, etc.

But, because we have a log-linear model, each term $\hat{m}_{213k}$ must be strictly positive. Summing over k, $\hat{m}_{213\cdot}$ must also be strictly positive. Because $\hat{m}_{213\cdot} = 0$, we have a contradiction. The MLEs must not exist. Therefore, we cannot fit any log-linear models that include [TSA]. Similarly, $n_{4\cdot 13} = n_{4\cdot 12} = 0$, so MLEs do not exist for models that include [TAR]. All other marginal totals are positive, so models with [TSA] or [TAR] are our primary source of concern.

To some extent, these problems can be evaded. Suppose we wish to fit a model with [TSA]. If we think of TSA as one composite factor, we can think of the problem as being one of fitting a $(4 \times 2 \times 3) \times 3$ table, i.e., a 24×3 table. In this table, three of the "rows" have zero totals. If we drop these three rows from the table, we can fit the remaining 21×3 table. Now, suppose we fit the model [TSA][TR][AR]. We have 21 degrees of freedom in the model for fitting [TSA]. Normally, this would be 24 degrees of freedom for fitting a grand mean, main effects T, S, A; two-factor effects (TS), (TA), (SA); and the three-factor effect (TSA). However, because 3 rows are being dropped, we have only 21 degrees of freedom. For fitting [TR], we have an additional 8 degrees of freedom. Adding [TR] involves adding the main effect R with 2 degrees of freedom and the two-factor effect (TR) with 6 degrees of freedom. Finally, adding [AR] is equivalent to adding the two-factor effect (AR) with 4 degrees of freedom. The model has $21+8+4 = 33$ degrees of freedom. The table has $21 \times 3 = 63$ degrees of freedom, so the test of [TSA][TR][AR] has $63 - 33 = 30$ degrees of freedom. (Incidentally, $G^2 = 34.72$, so this is a very good model.)

We now consider the problem of determining the degrees of freedom for testing the more complex model, [TSA][TAR]. Recall that there are marginal totals of 0 associated with both of these terms. Of the 24 marginal totals associated with [TSA] (obtained by summing over R), three are 0. This leads us to fitting the $(24-3) \times 3$, TSA by R table considered above. Of the 36 marginal totals associated with [TAR] (obtained by summing over S), two are 0. This would normally lead us to fitting the $(36-2) \times 2$ TAR by S table. In fact, what we need to do is fit the intersection of these tables. The 21×3 table involves dropping the cells (h, i, j, k) with values (2,1,3,1), (2,1,3,2), (2,1,3,3), (2,2,2,1), (2,2,2,2), (2,2,2,3), (4,1,1,1), (4,1,1,2), and (4,1,1,3). The 34×2 table drops the cells (4,1,1,3), (4,2,1,3), (4,1,1,2), and (4,2,1,2). Note that both tables drop the cells (4,1,1,2) and (4,1,1,3), so a total of 11 cells are being dropped from the $4 \times 2 \times 3 \times 3$ table. We are left with a table that has $72 - 11 = 61$ cells.

We now compute the degrees of freedom for the model [TSA][TAR]. As before, fitting [TSA] alone accounts for 21 degrees of freedom. Determining the degrees of freedom for adding [TAR] is somewhat more complex. Nor-

mally, [TAR] alone would involve 36 degrees of freedom broken down as follows: (grand mean) : (1), T : (3), A : (2), R : (2), (TA) : (6), (TR) : (6), (AR) : (4), and (TAR) : (12). With the marginal zeros, there are only 34 degrees of freedom. The 2 degrees of freedom come out of the (TAR) interaction, so the correct degrees of freedom are (TAR) : (10). Adding [TAR] to a model with [TSA] involves adding the effects R, (TR), (AR), and (TAR). The degrees of freedom for [TSA][TAR] are $21 + 2 + 6 + 4 + 10 = 43$. The lack of fit test for [TSA][TAR] has $61 - 43 = 18$ degrees of freedom; G^2 is 21.90.

The basic approach to dealing with models that have random zeros is to identify all cells that imply that MLEs do not exist. In other words, identify all cells for which the maximum likelihood constraints would imply that $\hat{m} = 0$. Such cells are dropped from the model and MLEs are found for the remaining cells. We are simply treating cells with "$\hat{m} = 0$" as fixed zeros. The degrees of freedom for the table are the number of cells in the full table minus the number of "fixed" zeros. The degrees of freedom for the model are the usual degrees of freedom for the model minus the number of degrees of freedom lost because there is "no information" available on some parameters. From Example 8.3.1, [TSA] usually involves 24 degrees of freedom related to the $4 \times 3 \times 3$ TSA marginal table. The table has $n_{213\cdot} = n_{222\cdot} = n_{411\cdot} = 0$. The nine cells involved in these three marginal totals are being treated like fixed zeros, so those nine cells "do not exist." There is no information available on 3 of the 24 cells of the TSA marginal table. Thus, 3 degrees of freedom are lost to the model because there is no information available.

We now consider one more example to set the ideas and illustrate some additional details.

EXAMPLE 8.3.2. The data in Table 8.4 were adapted from Lee (1980). The table involves four factors related to the survival of patients with stages 3 and 4 melanoma: Gender, Remission, Immunity, and Survival. Remission has three categories: still in remission, relapsed, never in remission. Immunity has three categories that were derived from results on six skin tests. One test score of at least 10 indicates good immunity. No test scores of at least 10 indicates no immunity. If more than half of the test scores are unknown and those that are known are less than 10, the immunity is taken as unknown. Finally, to get more interesting marginal zeros, Lee's count in cell (2,1,3,2) has been changed from 1 to 0.

Denote the factors Gender, Remission, Immunity, and Survival as G, R, I, and S, respectively. Consider fitting the model [GRI][GRS][GIS][RIS]. The likelihood equations are, for all h, i, j, and k,

$$\begin{aligned} n_{hij\cdot} &= \hat{m}_{hij\cdot}\,, \\ n_{hi\cdot k} &= \hat{m}_{hi\cdot k}\,, \\ n_{h\cdot jk} &= \hat{m}_{h\cdot jk}\,, \end{aligned}$$

TABLE 8.4. Melanoma Data

Gender (h)	Remission (i)	Immunity (j)	Survival (k) Dead	Alive
		No Immunity	2	0
	Relapsed	Immunity	4	1
		Unknown	0	0
		No Immunity	0	1
Male	Remission	Immunity	1	10
		Unknown	3	3
		No Immunity	3	1
	None	Immunity	10	5
		Unknown	8	2
		No Immunity	2	0
	Relapsed	Immunity	3	4
		Unknown	0	0
		No Immunity	0	0
Female	Remission	Immunity	0	10
		Unknown	0	4
		No Immunity	2	0
	None	Immunity	6	3
		Unknown	3	8

$$n_{\cdot ijk} \quad = \quad \hat{m}_{\cdot ijk}\,.$$

Many of these marginal tables have zero totals. The RIS marginal table has four zeros: $0 = n_{\cdot 131} = n_{\cdot 132} = n_{\cdot 112} = n_{\cdot 211}$. These involve the eight cells with counts of 0: (1,1,3,1), (2,1,3,1), (1,1,3,2), (2,1,3,2), (1,1,1,2), (2,1,1,2), (1,2,1,1), and (2,2,1,1). The GRI table has three zeros: $0 = n_{113\cdot} = n_{213\cdot} = n_{221\cdot}$. These involve six cases: (1,1,3,1), (1,1,3,2), (2,1,3,1), (2,1,3,2), (2,2,1,1), and (2,2,1,2). The GRS table has $n_{22\cdot 1} = 0$. This total involves the cases (2,2,1,1), (2,2,2,1), and (2,2,3,1). Finally, the GIS table has $n_{2\cdot 12} = 0$, so the cells (2,1,1,2), (2,2,1,2), and (2,3,1,2) are all 0.

Having listed all of the cells that would have to have $\hat{m} = 0$, we see that there are only 12 distinct cells. (These 12 happen to be all of the cells in the entire table with counts of 0, but that fact is irrelevant.) The full table is a $2 \times 3 \times 3 \times 2$ table having 36 cells. If we drop the 12 cells with $\hat{m} = 0$, we have 24 cells remaining in the table. Thus, the table has 24 degrees of freedom.

We now consider the degrees of freedom for the model. The RIS table has four zero totals. The corresponding cells are dropped from the table, so rather than being a 2×18 table, the G by RSI table is a 2×14 table. Thus, fitting [RIS] gives the model 14 degrees of freedom.

We can also compute the degrees of freedom for fitting [RIS] term by term. To compute the degrees of freedom term by term, we need to note that the RI marginal table has $n_{\cdot 13\cdot} = 0$. Thus, this 3×3 table has one dropped cell and only 8 degrees of freedom. The degrees of freedom are allocated: (grand mean) : (1), R : (2), I : (2), and (RI) : (3) rather than the standard value 4. Moving up to the RIS $3 \times 3 \times 2$ table, we have 14 nonempty cells. Because the (RI) interaction has only 3 degrees of freedom, the 14 degrees of freedom correspond to: (grand mean) : (1), R : (2), I : (2), S : (1), (RI) : (3), (RS) : (2), (IS) : (2), which leaves (RIS) : (1) rather than the standard value of 4. Thus, with this pattern of random zeros, the four empty cells cause a reduction of 1 degree of freedom in the (RI) interaction and 3 degrees of freedom in the (RIS) interaction. Note that the only two-factor marginal table that contains a zero is the RI table. Thus, any other reductions in degrees of freedom must occur in three-factor interaction terms.

A similar analysis holds for the GRI marginal table. This $2 \times 3 \times 3$ table has three zeros, so three cells are dropped. There are $18 - 3 = 15$ degrees of freedom. The degrees of freedom are allocated: (grand mean) : (1), G : (1), R : (2), I : (2), (GR) : (2), (GI) : (2); once again (RI) : (3) (rather than 4) which leaves us with (GRI) : (2) (rather than 4).

The model [RIS][GRI] has 14 degrees of freedom for [RIS]. We then add the terms G, (GR), (GI), and (GRI) with $1 + 2 + 2 + 2 = 7$ degrees of freedom. Thus, [RIS][GRI] has $14 + 7 = 21$ degrees of freedom.

The $2 \times 3 \times 2$ GRS table has one zero, hence 11 degrees of freedom. They are allocated: (grand mean) : (1), G : (1), R : (2), S : (1), (GR) : (2), (GS) : (1), (RS) : (2), which leaves (GRS) : (1). Adding [GRS] to [RIS][GRI] involves adding the terms (GS) and (GRS) with $1 + 1 = 2$ degrees of freedom, so [RIS][GRI][GRS] has $21+2 = 23$ degrees of freedom.

Finally, the $2\times3\times2$ GIS table has one cell dropped for 11 degrees of freedom. The 1 degree of freedom lost is taken from the (GIS) interaction, so (GIS) has 1 degree of freedom. The model [RIS][GRI][GRS][GIS] involves adding (GIS) to the model [RIS][GRI][GRS]. The degrees of freedom are $23 + 1 = 24$.

Recall that the degrees of freedom for the table are 24. With 24 degrees of freedom for the model, we get a perfect fit. Having dropped out the cells that require $\hat{m} = 0$, the model [RIS][GRI][GRS][GIS] is a saturated model.

In order to implement this approach to dealing with random zeros, we must be able to identify all cells for which the likelihood equations imply that $\hat{m} = 0$. As in Examples 8.3.1 and 8.3.2, it is frequently easy to identify some cells that have $\hat{m} = 0$. Often, all of the cells with $\hat{m} = 0$ are easily identified. Cells are easy to identify if they correspond to marginal totals that are zero. Unfortunately, sometimes all of the marginal totals can be positive, but cells with $\hat{m} = 0$ still exist.

EXAMPLE 8.3.3. Consider a $2 \times 2 \times 2$ table with $n_{111} = n_{222} = 0$ and $n_{ijk} > 0$ for all other cells. If we fit the model [12][13][23], the likelihood equations are

$$\begin{aligned} n_{ij\cdot} &= \hat{m}_{ij\cdot}\,, \\ n_{i\cdot k} &= \hat{m}_{i\cdot k}\,, \end{aligned}$$

and

$$n_{\cdot jk} = \hat{m}_{\cdot jk}\,.$$

Because n_{111} and n_{222} are the only 0s, all of the marginal totals listed above are positive. Looking at the marginal totals indicates no cause for concern. Nonetheless, these equations imply that $\hat{m}_{111} = \hat{m}_{222} = 0$.

To see this, first note that we must have $n_{\cdots} = \hat{m}_{\cdots}$. Writing out all of the likelihood equations involving n_{111} and n_{222} gives

$$\begin{aligned} n_{111} + n_{112} &= \hat{m}_{111} + \hat{m}_{112}\,, \\ n_{221} + n_{222} &= \hat{m}_{221} + \hat{m}_{222}\,, \\ n_{111} + n_{121} &= \hat{m}_{111} + \hat{m}_{121}\,, \\ n_{212} + n_{222} &= \hat{m}_{212} + \hat{m}_{222}\,, \\ n_{111} + n_{211} &= \hat{m}_{111} + \hat{m}_{211}\,, \\ n_{122} + n_{222} &= \hat{m}_{122} + \hat{m}_{222}\,. \end{aligned}$$

Adding these six equations together, we get

$$2(n_{111} + n_{222}) + n_{\cdots} = 2(\hat{m}_{111} + \hat{m}_{222}) + \hat{m}_{\cdots}\,.$$

Because $n_{...} = \hat{m}_{...}$, we have

$$n_{111} + n_{222} = \hat{m}_{111} + \hat{m}_{222} .$$

With $n_{111} = n_{222} = 0$, we need both $\hat{m}_{111} = 0$ and $\hat{m}_{222} = 0$. These two cells would be dropped from the $2 \times 2 \times 2$ table.

Although the author is unaware of any general method of identifying situations like that in Example 8.3.3, the process of fitting models can give hints as to whether such cells have been overlooked. In particular, if some estimated cell counts seem to be converging to zero or if the iterative estimated cell counts fail to converge, it would be wise to consider the possibility that this is due to cells that need to be dropped because of the pattern of random zeros.

Finally, it is interesting to look at the test of [TSA][TR][AR] versus [TSA][TAR] that can be obtained from Example 8.3.1. The test has $G^2 = 34.72 - 21.90 = 12.82$ with $df = 30 - 18 = 12$. In this example, the degrees of freedom for the test happen to correspond to the usual degrees of freedom for the (TAR) interaction with T at four levels, A at three levels, and R at three levels. However, the 12 degrees of freedom are actually arrived at quite differently. As established earlier, (TAR) has 10 degrees of freedom. The other 2 degrees of freedom in the test come from the fact that [TSA][TR][AR] is fit to a 63-cell table rather than the 61-cell table of [TSA][TAR]. So the 12 degrees of freedom come from 10 degrees of freedom for (TAR) and 2 degrees of freedom for new cells.

This discussion also points out that there are some technical difficulties involved in testing [TSA][TR][AR] versus [TSA][TAR]. We are testing a 63-cell table against a 61-cell table. We have not discussed this sort of thing previously. Obviously, this requires a mathematical theory that embeds both of these within the 72-cell $4 \times 2 \times 3 \times 3$ table allowing for fixed zero cells, log-linear models on the nonzero cells, and reduced models in which fixed zeros are allowed to become unfixed. Moreover, asymptotic theory will be of limited value because all of these problems are being caused by small sample sizes.

8.4 Exercises

EXERCISE 8.4.1. Brown (1980) presents data that are reproduced in Table 8.5, on a cross-classification of 53 prostate cancer patients. The factors are acid phosphatase level in the blood serum, age, stage, grade, x-ray, and nodal involvement. Acid level and age are categorized as high or low. Stage is an indication of size and location of the tumor; a positive value is more serious. The grade and x-ray indicate whether biopsy and x-ray tests are positive for cancer. The final factor is the whether the lymph nodes are involved. Analyze the data using iterative proportional fitting and ANOVA type models.

TABLE 8.5. Nodal Involvement in Prostate Cancer

Low Acid

Grade		−				+			
X−ray		−		+		−		+	
Involvement		No	Yes	No	Yes	No	Yes	No	Yes
Low	−	4	0	1	0	1	1	0	0
Low	+	0	0	0	1	1	0	0	1
High	−	3	0	2	0	1	0	0	0
High	+	2	1	0	0	4	0	0	0
Age	Stage								

High Acid

Grade		−				+			
X−ray		−		+		−		+	
Involvement		No	Yes	No	Yes	No	Yes	No	Yes
Low	−	5	1	1	0	0	1	0	0
Low	+	1	1	0	0	1	2	1	5
High	−	3	0	0	1	0	0	0	1
High	+	2	2	0	1	0	0	0	1
Age	Stage								

EXERCISE 8.4.2. Extend your analysis of the Berkeley graduate admissions data (cf. Exercise 3.8.4) by incorporating the method of partitioning polytomous factors from Example 8.1.2.

EXERCISE 8.4.3. *Partitioning Two-Way Tables.*
Lancaster (1949) and Irwin (1949) present a method of partitioning tables that was used in Exercise 2.7.4. We now establish the validity of this method. Consider a two-dimensional $I \times (J + K - 1)$ table. The partitioning method tests for independence in two subtables. One table is a reduced $I \times K$ table consisting of the last K columns or the full table. The other table is an $I \times J$ table that uses the first $J - 1$ columns of the full table and also includes a column into which the last K columns of the full table have been collapsed. Write the data with three subscripts as n_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, L_j$, where

$$L_j = \begin{cases} 1 & \text{if } j \neq J \\ K & \text{if } j = J. \end{cases}$$

Consider the models

$$\log(m_{ijk}) = \alpha_i + \beta_{jk}, \tag{1}$$

$$\log(m_{ijk}) = \beta_{ij} + \beta_{jk}, \tag{2}$$

and

$$\log(m_{ijk}) = \gamma_{ijk}. \tag{3}$$

Model (3) is the saturated model so

$$\hat{m}_{ijk}^{(3)} = n_{ijk}.$$

Model (1) is the model of independence in the $I \times J + K - 1$ table, so

$$\hat{m}_{ijk}^{(1)} = n_{i..}n_{.jk}/n_{...}.$$

(a) Show that the maximum likelihood estimates for model (2) are

$$\hat{m}_{ijk}^{(2)} = \begin{cases} n_{ijk} & \text{if } j \neq J \\ n_{iJ.}n_{.Jk}/n_{.J.} & \text{if } j = J. \end{cases}$$

(b) Show that both G^2 and X^2 are the same for testing model (2) against model (3) as for testing the reduced table for independence.

(c) Show that both G^2 and X^2 are the same for testing model (1) against model (2) as for testing the collapsed table for independence.

(d) Extend (b) and (c) by showing that all power divergence statistics are the same, cf. Exercise 2.7.8.

EXERCISE 8.4.4. *The Bradley-Terry Model.*
"Let's suppose, for a moment, that you have just been married and that you are given a choice of having, during your entire lifetime, either x or y children. Which would you choose?" Imrey, Johnson, and Koch (1976) report results from asking this question of 44 Caucasian women from North Carolina who were under 30 and married to their first husband. Women were asked to respond for pairs of numbers x and y between 0 and 6 with $x < y$. The data are summarized in Table 8.6. The most basic form for such experiments is to ask each woman to respond for all possible pairs of numbers. If this was done, there is a considerable amount of missing data. For example, in comparing 0 children with 1 child there are only $17+2 = 19$ responses rather than 44.

TABLE 8.6. Family Size Preference

Alternative Choice	Preferred Number of Children						
	0	1	2	3	4	5	6
0	—	17	22	22	15	26	25
1	2	—	19	13	10	9	11
2	1	0	—	11	11	6	6
3	3	1	7	—	6	2	6
4	1	10	12	13	—	4	0
5	1	11	18	15	17	—	11
6	2	13	20	22	14	12	—

This data collection technique is called the method of *paired comparisons*. It is often used for such things as taste tests. Subjects find it easier to

distinguish a preference between two brands of cola than to rank their preferences among half a dozen. David (1988) provides a good survey of the literature on the analysis of preference data along with notes on the history of the subject. One particular model for preference data assumes that each item has a probability π_i of being preferred. Thus, in a paired comparison, the conditional probability that i is preferred to j is $\pi_i/(\pi_i + \pi_j)$. There are many ways to arrive at this model; the one given above is simple but restrictive. In other developments, the parameters π_i need not add up to one, but it is no loss of generality to impose that condition. Bradley and Terry (1952) rediscovered the model and popularized it. The Bradley-Terry model was put into a log-linear model framework by Fienberg and Larntz (1976). For I items being compared, their framework consists of fitting the incomplete $I(I-1)/2 \times I$ table in which the rows consist of all pairs of items and the columns consist of the preferred item. A test of the model $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$ is a test of whether the Bradley-Terry model holds.

(a) Rewrite Table 8.6 in the Fienberg-Larntz form.

(b) Test whether the Bradley-Terry model fits.

(c) Show that under the log-linear main effects model the odds of preferring item j to item j' is the ratio of a non-negative number depending on j and a non-negative number depending on j'. Show that this is equivalent to the Bradley-Terry model.

(d) Estimate the probabilities π_i.

9

Generalized Linear Models

Generalized linear models are a class of models that generalize the linear models used for regression and analysis of variance. They allow for more general mean structures and more general distributions than regression and analysis of variance. Generalized linear models were first suggested by Nelder and Wedderburn (1972). An extensive treatment is given by McCullagh and Nelder (1989). Generalized linear models include logistic regression as a special case. Another special case, Poisson regression, provides the same analysis for count data as log-linear models. The discussion here involves more distribution theory than has been required elsewhere in this book; in particular, it makes extensive use of the exponential family of distributions and the gamma distribution. Information on these distributions can be obtained from many sources, e.g., Cox and Hinkley (1974). Section 1 presents the family of distributions used in generalized linear model theory. Estimation of the linear parameters is dealt with in Section 2. Model fitting and estimation of dispersion are examined in Section 3; both of these topics involve a version of the likelihood ratio test statistic called the *deviance*. Section 4 contains a summary and discussion. We begin with a brief review of some ideas from regression and analysis of variance and three examples of generalized linear models.

Regression and analysis of variance are fundamental tools in statistics. A multiple regression model is

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i,$$

$i = 1, \ldots, n$, where $\mathrm{E}(e_i) = 0$ and, typically, $x_{i1} = 1$, $i = 1, \ldots, n$, so that β_1 is an intercept. The x_{ij}'s are all assumed to be known predictor

variables; the β_j's are fixed unknown parameters. A one-way analysis of variance can be written as

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + e_{ij} \\ &= \mu \cdot (1) + \alpha_1 \delta_{1i} + \cdots + \alpha_t \delta_{ti} + e_{ij} \end{aligned}$$

where $i = 1, \ldots, t$, $j = 1, \ldots, n_i$, $\mathrm{E}(e_{ij}) = 0$, and δ_{hi} is 1 if $h = i$ and 0 otherwise. In the analysis of variance, μ and the α_i's are fixed unknown parameters, while the multiplier 1 for μ and the δ_{hi}'s play roles analogous to the x_{ij}'s in regression.

The key aspect of both regression and ANOVA is that they involve observations whose expected value is a linear combination of known predictor variables. In regression,

$$\mathrm{E}(y_i) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

and in analysis of variance

$$\mathrm{E}(y_{ij}) = \mu \cdot (1) + \alpha_1 \delta_{1i} + \cdots + \alpha_t \delta_{ti} \, .$$

If the observations y in regression and ANOVA have the same variance and are uncorrelated, regression and ANOVA provide the best estimates of (estimable) parameters among all estimators that are unbiased linear functions of the observations. If we go a step further and assume that the observations have independent normal distributions with the same variance, then the usual estimates are the best among all unbiased functions of the observations. In both of these statements, "best" means that the estimates have minimum variance. Under the assumption of independent normal distributions with the same variance, the usual estimates are also maximum likelihood estimates. The current chapter is concerned with finding maximum likelihood estimates for a more general set of models. The models are less restrictive in that they allow more general forms of linearity and distributions other than the normal.

We now set some matrix notation. Both regression and analysis of variance are *linear models*, cf. Christensen (1996b). Let y_i be an observation. It is of no significance how we subscript the observations. Any convenient method of subscripting is acceptable whether it be one subscript as in regression, two subscripts as in one-way analysis of variance, or three subscripts as in two-way ANOVA with replications. Let $x_i' = (x_{i1}, \ldots, x_{ip})$ be a $1 \times p$ row vector of predictor variables. Let $\beta = (\beta_1, \ldots, \beta_p)'$ be a $p \times 1$ column vector of unknown parameters. A typical normal theory linear model assumes

$$y_i \sim N(x_i'\beta, \sigma^2)$$

where $i = 1, \ldots, n$ and the y_i's are independent. For pedagogical reasons, it is advantageous to write

$$y_i \sim N(m_i, \sigma^2), \qquad m_i = x_i'\beta.$$

We begin our discussion of generalized linear models by considering three examples. *In each case, we take* $y_1, \ldots, y_n$ *independent.* The models are specified by their distributions and mean structure.

For normal data,

$$y_i \sim N(m_i, \sigma^2), \qquad \mathrm{E}(y_i) = m_i, \qquad m_i = x_i'\beta.$$

This is the model for analysis of variance and regression.

For Poisson data,

$$y_i \sim \mathrm{Pois}(m_i), \qquad \mathrm{E}(y_i) = m_i, \qquad \log(m_i) = x_i'\beta.$$

This is just a log-linear model for a table containing n cells where the count in each cell has a Poisson distribution with parameter m_i. It is important to note that n is the number of cells and *not* the observation vector as it is elsewhere in this book. As we have mentioned before and as is shown in Chapter 12, under very weak conditions the analysis of a contingency table under Poisson sampling is the same as the analysis under multinomial sampling. It is interesting to note that the framework for generalized linear models assumes independent observations, so it does not apply directly to multinomial sampling or to general product-multinomial sampling. It is the equivalence of the maximum likelihood analyses under the Poisson, multinomial, and product-multinomial sampling schemes that makes generalized linear models a useful tool for contingency tables.

For binomial sampling, we take y_i to be the *proportion* of successes, so

$$N_i y_i \sim \mathrm{Bin}(N_i, p_i), \qquad \mathrm{E}(y_i) = p_i \equiv m_i,$$

$$\log\left(\frac{m_i}{1 - m_i}\right) = \log\left(\frac{p_i}{1 - p_i}\right) = x_i'\beta.$$

Note that N_i is a known quantity and not a parameter. The model is simply a logistic regression or logit model. The data consist of proportions obtained from independent binomial random variables and the mean structure is a linear model in the log odds. Alternative models for binomial regression will be mentioned later. Except in this chapter, y_i for binomial regression is always taken to mean the number of successes, rather than the proportion of successes. The change is made to fit the binomial distribution into the family of generalized linear models.

9.1 Distributions for Generalized Linear Models

The normal, Poisson, and binomial distributions considered above are members of the exponential family of distributions. A random variable y has

a distribution in the exponential family if it has a probability density or mass function that can be written as

$$f(y|\theta) = R(\theta)\exp\left[\sum_{j=1}^{v} q_j(\theta)\, t_j(y)\right] h(y)$$

where $\theta = (\theta_1, \ldots, \theta_s)'$ is a vector of parameters. If $v = 1$, the family is referred to as a one-parameter exponential family; the one parameter can be taken as $q_1(\theta)$.

The theory of generalized linear models requires the distribution of y to be in a subclass of the one-parameter exponential family. The density or mass function must have the form

$$f(y|\theta, \phi; w) = \exp\left\{\frac{w}{\phi}[\theta y - r(\theta)]\right\} h(\phi, y, w) \tag{1}$$

where θ, ϕ, and w are scalars. By assumption, w is a fixed known number. The role of ϕ in this function is curious; it is treated as an unknown constant but not as a parameter. With ϕ constant, the distribution (1) is in the one-parameter exponential family; just take $R(\theta) = \exp[-wr(\theta)/\phi]$, $q(\theta) = w\theta/\phi$, $t(y) = y$, and $h(y) = h(\phi, y, w)$. In practice, ϕ is often an unknown parameter. As such, the distribution need not be in the exponential family relative to the two parameters θ and ϕ because the function $h(\phi, y, w)$ need not satisfy the conditions of a two-parameter exponential family. The value ϕ is simply a *positive* number that is convenient for defining various special cases. The particular form of $f(y|\theta, \phi; w)$ in (1) is chosen so that the maximum likelihood estimate of θ does not depend on ϕ. This will be discussed in more detail in Section 2.

For the family of distributions (1), the expected value of y depends on θ but not on ϕ. For any distribution,

$$1 = \int f(y|\theta, \phi; w)\, dy$$

where it is understood that integration is always replaced by summation when y has a discrete distribution. Taking the derivative with respect to θ on both sides gives

$$0 = \int \dot{f}(y|\theta, \phi; w)\, dy \tag{2}$$

where $\dot{f}$ is the derivative of f with respect to θ and f satisfies conditions so that the derivative can be taken under the integral sign. From the exact form of $f(y|\theta, \phi; w)$ in (1), it is easily seen that (2) is

$$0 = \frac{w}{\phi}\int (y - \dot{r}(\theta))\, f(y|\theta, \phi; w)\, dy$$

where $\dot{r}(\theta)$ is the derivative $dr(\theta)/d\theta$. It follows that

$$\mathrm{E}(y) \equiv m = \dot{r}(\theta).$$

Typically, $\dot{r}$ is an invertible function, so θ is also a function of the mean, say

$$\theta = \dot{r}^{-1}(m).$$

Linear structure for distributions of the form (1) is most naturally specified by

$$\theta = x'\beta \tag{3}$$

where, as usual, β is a vector of unknown parameters and x is fixed and known. Note that with $\theta = \dot{r}^{-1}(m)$, the linear structure $x'\beta$ in equation (3) is also a function of the mean. In fact, the analysis of generalized linear models can be carried through when the linear structure is a more general function of the mean,

$$g(m) = x'\beta,$$

as long as it is possible to write $\theta = g_*(x'\beta)$ for some function $g_*(\cdot)$.

A *generalized linear model* consists of independent observations y_i, $i = 1, \ldots, n$, with

$$y_i \sim f(y_i|\theta_i, \phi, w_i), \qquad \mathrm{E}(y_i) \equiv m_i, \qquad g(m_i) = x_i'\beta.$$

If $g(m_i) = \theta_i$, the model is a *canonical* generalized linear model. In other words, a canonical model has $g(\cdot) \equiv \dot{r}^{-1}(\cdot)$.

Names have been given to the various components of generalized linear models. The linear structure $x'\beta$ is called the *linear predictor*. The function $g(\cdot)$ that specifies the relationship $g(m) = x'\beta$ between the mean and the linear predictor is called the *link function*. If $g(m) = \theta$, the function $g(\cdot)$ is called the *canonical link function*. The density $f(y|\theta, \phi; w)$ is often called the *error function* and the parameter ϕ is often called the *dispersion parameter*.

In Section 3, we will need to know the variance of y when y has a density of the form (1). Taking the second derivative with respect to θ on both sides of

$$1 = \int f(y|\theta, \phi; w)\, dy$$

and assuming that derivatives can be taken under the integral gives

$$\begin{aligned} 0 &= \int \ddot{f}(y|\theta, \phi; w)\, dy \\ &= \frac{w}{\phi} \int \frac{d\big[(y - \dot{r}(\theta))\, f(y|\theta, \phi; w)\big]}{d\theta} dy \\ &= \frac{w^2}{\phi^2} \int (y - \dot{r}(\theta))^2 f(y|\theta, \phi; w)\, dy \\ &\quad + \frac{w}{\phi} \int -\ddot{r}(\theta) f(y|\theta, \phi; w)\, dy \end{aligned}$$

where two dots indicate a second derivative. It follows immediately that $0 = \left[\mathrm{Var}(y)w^2/\phi^2\right] - [\ddot{r}(\theta)w/\phi]$ and, thus,

$$\mathrm{Var}(y) = \ddot{r}(\theta)\phi/w.$$

The function $\ddot{r}(\theta)$ is often written as a function of m,

$$V(m) \equiv \ddot{r}\big(\dot{r}^{-1}(m)\big) = \ddot{r}(\theta).$$

$V(m)$ is generally referred to as the *variance function*.

We now review how the general distribution theory applies to the three examples given earlier: normal, Poisson, and binomial sampling.

If $y \sim N(m, \sigma^2)$, the density for y real is

$$\begin{aligned} f(y|m;\sigma^2) &= \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left[-\frac{(y-m)^2}{2\sigma^2}\right] \\ &= \exp\left(\frac{-m^2}{2\sigma^2}\right)\exp\left(\frac{my}{\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi}\,\sigma}e^{-y^2/2\sigma^2}\right). \end{aligned}$$

To see that the density has the form of equation (1), identify $\theta = m$, $w = 1$, $\phi = \sigma^2$, $r(\theta) = m^2/2$, and $h(\phi, y, w) = \left(1/\sqrt{2\pi}\sigma\right) e^{-y^2/2\sigma^2}$. The expected value of y is m, so the canonical linear structure is $\theta = m = x'\beta$. The canonical link leads to a standard linear model.

For $y \sim \mathrm{Pois}(m)$, the probability mass function on $y = 0, 1, 2, \ldots$ is

$$\begin{aligned} f(y|m) &= \frac{m^y e^{-m}}{y!} \\ &= \exp(-m)\exp(y\,\log(m))\,(1/y!)\,. \end{aligned}$$

Identify $\theta = \log(m)$, $w = 1$, $\phi \equiv 1$, $r(\theta) = m$, and $h(\phi, y, w) = (1/y!)$. It is well known that for a $\mathrm{Pois}(m)$ distribution, the expected value and variance are both m. To see this from the general distribution theory, observe that the mean is $\dot{r}(\theta)$ and with $w = 1$ and $\phi \equiv 1$, the variance is $\ddot{r}(\theta)$. From $\theta = \log(m)$ and $r(\theta) = m$, it follows that $r(\theta) = e^\theta$ and thus $\dot{r}(\theta) = e^\theta$ and $\ddot{r}(\theta) = e^\theta$. Again, using $\theta = \log(m)$ gives $m = \dot{r}(\theta) = \ddot{r}(\theta)$. The expected value of y is m, so the canonical linear structure is $\theta = \log(m) = x'\beta$. The canonical link leads to a standard log-linear model for Poisson data.

For $Ny \sim \mathrm{Bin}(N, p)$ with N known, the mass function on $Ny = 0, \ldots, N$ is

$$\begin{aligned} f(y|p) &= \binom{N}{Ny}p^{Ny}(1-p)^{N-Ny} \\ &= \binom{N}{Ny}(1-p)^N\left(\frac{p}{1-p}\right)^{Ny} \\ &= (1-p)^N\exp\left[Ny\log\left(\frac{p}{1-p}\right)\right]\binom{N}{Ny}. \end{aligned}$$

Identify $\theta = \log\left(\frac{p}{1-p}\right)$, $w = N$, $\phi \equiv 1$, $r(\theta) = -\log(1-p)$, and $h(\phi, y, w) = \binom{N}{Ny}$. The expected value of y is $m \equiv p$, so the canonical linear structure is $\theta = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{m}{1-m}\right) = x'\beta$. The canonical link leads to a standard logistic (logit) model. (See Exercise 9.5.1.)

In general, the inverse of any cumulative distribution function (cdf) $F(\cdot)$ makes a reasonable link function for binomial data, i.e., $g(p) = F^{-1}(p)$. $F(u) = e^u/(1+e^u)$ is the cdf of the logistic distribution and defines logistic regression. Probit regression is the procedure based on taking $F(u) = \Phi(u)$ where $\Phi(u)$ is the cdf of a standard normal distribution. A third example is complementary log-log regression which uses $F(u) = 1 - \exp[1 - \exp(e^u)]$.

As a last example, consider the gamma distribution. The gamma distribution is defined by the probability density function

$$f(y|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda y} y^{\alpha-1}$$

for $y > 0$. The density depends on two parameters, α and λ. The expected value of a gamma distribution is

$$\mathrm{E}(y) \equiv m = \frac{\alpha}{\lambda}$$

and the variance is

$$\mathrm{Var}(y) = \frac{\alpha}{\lambda^2}.$$

To indicate that y has a gamma distribution, write

$$y \sim \mathrm{Gamma}(\alpha, \lambda).$$

Special cases of the gamma distribution include exponential distributions with mean $1/\lambda$, i.e., $\mathrm{Gamma}(1, \lambda)$ and $\chi^2(n)$ distributions, $\mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$'s.

The gamma density can be rewritten as

$$f(y|\alpha, \lambda) = \left(\frac{\lambda}{\alpha}\right)^\alpha \exp\left[\alpha\left(\frac{-\lambda}{\alpha}\right) y\right] \left(\frac{\alpha^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\right).$$

To see the gamma density in the form of equation (1), identify $\theta = -\lambda/\alpha$, $w = 1$, $\phi = 1/\alpha$, $r(\theta) = -\log(\lambda/\alpha)$, and $h(\phi, y, w) = \alpha^\alpha y^{\alpha-1}/\Gamma(\alpha)$. The expected value of y is $m = \alpha/\lambda = -1/\theta$, so the canonical linear structure is $\theta = -1/m = x'\beta$. Note that the distribution is only defined for $y > 0$; thus, for any gamma distribution, $m > 0$. It follows that when using the canonical link, restrictions must be placed on the parameter vector β to ensure that the expected value is positive. (See Exercise 9.5.2.)

The canonical generalized linear model for n independent observations with gamma distributions is

$$y_i \sim \mathrm{Gamma}(\alpha, \lambda_i), \qquad \mathrm{E}(y_i) = m_i = \frac{\alpha}{\lambda_i}, \qquad \frac{-1}{m_i} = x_i'\beta$$

where β is restricted so that $-x_i'\beta > 0$ for all i. Gamma distribution regression is useful for modeling situations in which the coefficient of variation is constant. The coefficient of variation is

$$\frac{\sqrt{\text{Var}(y_i)}}{\text{E}(y_i)} = \frac{\sqrt{\alpha/\lambda_i^2}}{\alpha/\lambda_i} = \frac{1}{\sqrt{\alpha}}.$$

When the data appear to have a constant coefficient of variation, using the gamma distribution is an alternative to doing a standard linear model analysis on the logs of the data, cf. Christensen (1996b, Section 13.7). Note that the constant coefficient of variation does not depend on the choice of link function. Often, noncanonical links such as the identity, $m = x'\beta$, and the log, $\log(m) = x'\beta$, are used with gamma distributed data, cf. McCullagh and Nelder (1989). The identity link requires restrictions on β; the log link does not. When the coefficient of variation is small, the log link analysis is very similar to the linear model analysis on the logs of the data. *The log link is probably the most commonly used for gamma regression. Exponential regression* is the special case with $\phi = 1$ and (usually) a log link.

9.2 Estimation of Linear Parameters

In their most natural form, generalized linear models assume n independent observations with

$$y_i \sim f(y_i|\theta_i, \phi; w_i)$$

and

$$\theta_i = x_i'\beta$$

for some vector of parameters $\beta = (\beta_1, \ldots, \beta_p)'$. The density $f(y_i|\theta_i, \phi; w_i)$ is defined by equation (9.1.1). The linear structure given above uses the canonical link function. More generally, the linear structure can be defined by

$$g(m_i) = x_i'\beta.$$

For this extension, assume $\theta_i = \dot{r}^{-1}(m_i)$ and

$$\theta_i = \dot{r}^{-1}\big(g^{-1}(x_i'\beta)\big) \equiv g_*(x_i'\beta).$$

The likelihood function for the generalized linear model with canonical link function is

$$\begin{aligned} L(\beta;\phi) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi; w_i) \\ &= \prod_{i=1}^{n} f(y_i|x_i'\beta, \phi; w_i). \end{aligned}$$

Using equation (9.1.1), the log-likelihood is

$$\begin{aligned}\ell(\beta;\phi) &\equiv \log[L(\beta;\phi)] \\ &= \sum_{i=1}^{n} \log[f(y_i|x_i'\beta,\phi;w_i)] \qquad (1)\\ &= \sum_{i=1}^{n} \frac{w_i}{\phi}\,[x_i'\beta\, y_i - r(x_i'\beta)] + \sum_{i=1}^{n} \log[h(\phi,y_i,w_i)]\,.\end{aligned}$$

With ϕ fixed, the maximum likelihood estimate of β is obtained by solving for β in the likelihood equations

$$\frac{\partial \ell(\beta;\phi)}{\partial \beta_j} = 0, \qquad (2)$$

$j = 1,\ldots,p$. It is a simple matter to see that taking the partial derivatives $\partial\ell(\beta;\phi)/\partial\beta_j$ leads to likelihood equations of the form

$$\frac{Q_j(\beta)}{\phi} = 0$$

for some functions $Q_j(\cdot)$, $j = 1,\ldots,p$. Obviously, the solution $\hat{\beta}$ to such likelihood equations does not depend on the value of ϕ. Thus, $\hat{\beta}$ is the maximum likelihood estimate for any value of ϕ; i.e., it is the maximum likelihood estimate regardless of the true value of ϕ.

Essentially, the same analysis holds when a linear structure $g(m_i) = x_i'\beta$ is assumed. With $\theta_i = g_*(x_i'\beta)$, simply use $g_*(x_i'\beta)$ in place of $x_i'\beta$ in equation (1). The only problem is that the partial derivatives become slightly more difficult to find.

Maximum likelihood estimates are invariant under transformations of the parameters. In other words, given a maximum likelihood estimate for a parameter, any function of the maximum likelihood estimate is the maximum likelihood estimate for the corresponding function of the parameter. For a discussion of this property see Cox and Hinkley (1974, p. 287). Given a maximum likelihood estimate for β, say $\hat{\beta}$, we immediately obtain an estimate of the expected value m_i, namely

$$\hat{m}_i = g^{-1}(x_i'\hat{\beta}),$$

an estimate of the linear predictor $g(m_i)$, namely

$$g(\hat{m}_i) = x_i'\hat{\beta},$$

and an estimate of θ_i, namely

$$\hat{\theta}_i = g_*(x_i'\hat{\beta}).$$

Solving the likelihood equations (2) is typically accomplished by using the Newton-Raphson algorithm. For generalized linear models, this reduces to performing a series of weighted least squares regressions and is known as *iteratively reweighted least squares*. Sections 10.5 and 11.3 give details for the special cases of log-linear modeling and logistic regression.

Under suitable conditions, the estimate $\hat{\beta}$ and smooth functions of $\hat{\beta}$, e.g., $\hat{m}_i = g^{-1}(x_i'\hat{\beta})$, have asymptotic multivariate normal distributions. Moreover, an estimate of the asymptotic covariance matrix of $\hat{\beta}$ is easily obtained from the iteratively reweighted least squares algorithm. Under suitable conditions, this estimate is consistent and also yields estimated asymptotic covariance matrices for smooth functions of $\hat{\beta}$. Given the estimates and the estimated asymptotic covariance matrices, standard normal theory methods for tests and confidence regions can be applied to yield asymptotic statistical inferences.

This brief discussion of estimation has not addressed several important points. To perform the differentiations, the $x_i'\beta$'s need to define a regression so that the β_j's are well defined. The partial derivatives need to be derived and shown to be of the form $Q_j(\beta)/\phi$, cf. Exercise 9.5.3. A solution to the likelihood equations must be shown to give the maximum of the log-likelihood. Exact conditions for the asymptotic results need to be stated; the necessary conditions may differ for different generalized linear models. For example, in regression analysis, one typically thinks about having the number of observations n go to infinity; however, for contingency table data, the number of cells in the table is n and is typically considered fixed, while the number of counts within the cells is assumed to get large. For more information on many of these issues see McCullagh and Nelder (1989).

9.3 Estimation of Dispersion and Model Fitting

Generalized linear model theory focuses on the estimation of linear parameters. The general theory seems to be less well developed for the purposes of model fitting and dispersion estimation. The basic statistics used in model fitting and estimating functions of the dispersion parameter ϕ are the *deviance* and a generalization of the Pearson test statistic. There are patterns common to the use of these statistics, but specifics vary from case to case. Our discussion focuses on two general asymptotic approaches. In one approach, the number of observations n is allowed to go to infinity. This approach is appropriate for many linear model and logistic regression problems. The second approach fixes n and uses asymptotics based on other aspects of the model. This approach is appropriate for many contingency table problems. We begin by defining the statistics.

The *standardized deviance* is simply the asymptotic form of the likelihood ratio statistic for testing a generalized linear model against the cor-

responding saturated model. Remember that in the likelihood analysis of a generalized linear model, the dispersion parameter ϕ is treated as fixed. A saturated model is simply one in which the number of parameters is so large that the data are fit perfectly. In particular, the model

$$y_i \sim f(y_i|\theta_i, \phi; w_i),$$

with no restrictions on the θ_i's, is saturated because there are as many parameters θ_i as there are observations. The maximum likelihood estimates have $\hat{m}_i = y_i$. This is easily established from the likelihood equations for the θ_i's. Substituting θ_i for $x_i'\beta$ in (9.2.1) and taking partial derivatives with respect to the θ_i's gives $y_i = \dot{r}(\hat{\theta}_i) = \hat{m}_i$ as a solution to the equations. The estimates of the θ_i's for the saturated model are determined by the estimates of the m_i's.

The parameters β and $m = (m_1, \ldots, m_n)'$ are assumed to be interchangeable, so write

$$\ell(m; \phi) \equiv \ell(\beta; \phi).$$

Also, write $y = (y_1, \ldots, y_n)'$. The *standardized deviance* is two times the difference between the maximum of the log-likelihood under the saturated model and the maximum of the log-likelihood under the specified generalized linear model, i.e.,

$$D^*(\hat{m}; \phi) = 2\left[\ell(y; \phi) - \ell(\hat{m}; \phi)\right].$$

Here, y is used in $\ell(y; \phi)$ because y is the maximum likelihood estimate of m for the saturated model. From inspection of (9.2.1), it is easily seen that the standardized deviance can be written as

$$D^*(\hat{m}; \phi) = \frac{D(\hat{m})}{\phi}$$

for a function $D(\hat{m})$ that does not depend on ϕ. Define the function $D(\hat{m})$ to be the *deviance* of the generalized linear model. Recall that in many important special cases, $\phi = 1$. For normal theory linear models, $D(\hat{m})$ is the sum of squares error.

As mentioned before, the likelihood analysis treats ϕ as fixed and ignores the fact that the dispersion ϕ is a parameter. The standardized deviance $D^*(\hat{m}; \phi)$ is only the likelihood ratio test statistic when ϕ is known. When ϕ is unknown, $D^*(\hat{m}; \phi)$ is not even a statistic because it depends on an unknown parameter. $D(\hat{m})$, on the other hand, does not depend on ϕ, so it is a statistic.

Another statistic used to evaluate models and estimate dispersion is the *generalized Pearson statistic*. The Pearson statistic is defined as

$$X^2 = \sum_{i=1}^{n} \frac{w_i\,(y_i - \hat{m}_i)^2}{V(\hat{m}_i)}$$

where $V(\cdot)$ is the variance function defined in Section 1. See Exercise 9.5.6.

The Pearson statistic can be used for consistent estimation of ϕ for large n. The variance of y_i is $V(m_i)\phi/w_i$ so, clearly,

$$\mathrm{E}\left(\frac{w_i\,(y_i - m_i)^2}{V(m_i)}\right) = \phi.$$

By Chebyshev's Weak Law of Large Numbers (cf. Rao, 1973, p. 112), if

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{w_i^2\mathrm{E}(y_i - m_i)^4}{V(m_i)^2} \to 0$$

as $n \to \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{w_i\,(y_i - m_i)^2}{V(m_i)} \xrightarrow{P} \phi.$$

It follows that if $V(\cdot)$ is a continuous function and $\hat{m}_i \xrightarrow{P} m_i$ for all i,

$$\frac{X^2}{n-p} \xrightarrow{P} \phi$$

where p is the number of parameters in β and we are assuming that the $n \times p$ model matrix

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

has $\mathrm{rank}(X) = p$ not depending on n. Typically, we take

$$\hat{\phi} = \frac{X^2}{n-p}.$$

Continuous functions of the dispersion, say $d(\phi)$, are estimated with $d(\hat{\phi})$. If the Pearson estimate is consistent, continuous functions of it are also consistent estimates.

Under some large sample conditions with fixed n, for all values of ϕ the standardized Pearson statistic has the asymptotic distribution

$$\frac{X^2}{\phi} \sim \chi^2(n-p).$$

In these cases, standard methods of variance estimation for normal data can be applied to give asymptotic confidence intervals and tests for ϕ, cf. Christensen (1996a, Sec. 2.6). Using properties of the χ^2 distribution, the asymptotic distribution also leads to the approximation

$$\mathrm{E}\big(X^2\big) \doteq \phi \cdot (n-p).$$

Once again, an obvious estimate of ϕ is

$$\hat{\phi}_P = \frac{X^2}{n-p}.$$

Similarly, under certain conditions with n fixed, the standardized deviance $D^*(\hat{m};\phi)$ has the asymptotic distribution

$$D^*(\hat{m};\phi) \sim \chi^2(n-p)$$

for all values of ϕ. By definition,

$$D(\hat{m}) = \phi D^*(\hat{m};\phi),$$

so, asymptotically,

$$D(\hat{m}) \sim \phi\chi^2(n-p).$$

Again, when the asymptotic distribution is valid, standard methods of variance estimation for normal data can be applied to give asymptotic confidence intervals and tests for ϕ. An obvious point estimate of ϕ is

$$\hat{\phi}_D = \frac{D(\hat{m})}{n-p}.$$

Unfortunately, the deviance-based estimate is frequently inconsistent when $n \to \infty$. Even in the simplest binomial case, $y_i \sim \text{Bin}(1,p)$, this estimate is not consistent. For this case, $\phi \equiv 1$, but for large samples, it is easily seen that

$$\frac{D}{n-1} \xrightarrow{P} -2\left[p\log(p) + (1-p)\log(1-p)\right]$$

which is not typically equal to 1.

Deviances and Pearson statistics can also be used to evaluate the adequacy of generalized linear models. If null distributions are available for the statistics, tests of the adequacy of models can be performed. These can be unconditional, exact conditional, approximate conditional, or asymptotic distributions. If null distributions are not available, the statistics can be used in an exploratory fashion to give rough ideas of model adequacy.

If the data follow a true one-parameter exponential family distribution, the dispersion parameter ϕ is identically constant and, without loss of generality, we can take $\phi \equiv 1$. If the model is correct and the Pearson statistic gives a consistent estimate of ϕ,

$$\frac{X^2}{n-p} \xrightarrow{P} 1.$$

If $X^2/(n-p)$ is substantially larger than 1, it is an indication that the model is incorrect. With $\phi \equiv 1$, the standardized deviance equals the deviance.

If the deviance estimate of ϕ is consistent, the deviance can also be used to evaluate lack of fit. When n is fixed, if the deviance has an asymptotic χ^2 distribution, the deviance is a lack of fit test statistic that can be used in a formal asymptotic test of the generalized linear model against the saturated model. Similarly, if X^2 has an asymptotic χ^2 distribution, the Pearson statistic can be used in a lack of fit test.

To test a model $g(m_i) = x_i'\beta$ against a reduced model, say $g(m_i) = x_{0i}'\gamma$ in a one-parameter family, simply compare the difference in the deviances to an appropriate χ^2 distribution. In particular, the asymptotic generalized likelihood ratio test rejects the adequacy of the reduced model at the α level if

$$D(\hat{m}_0) - D(\hat{m}) > \chi^2(1-\alpha, p-p_0).$$

Here, $\hat{m}_0$ and $D(\hat{m}_0)$ are the maximum likelihood estimate of m and the deviance under the reduced model. The model is a reduced model in the sense that $X_0 = XB$ for some matrix B where

$$X_0 = \begin{bmatrix} x_{01}' \\ \vdots \\ x_{0n}' \end{bmatrix}$$

and

$$\operatorname{rank}(X_0) = p_0.$$

Such reduced model tests tend to be asymptotically valid under weaker conditions than general lack of fit tests. In particular, the tests are often valid under both asymptotic approaches discussed here. Less formally, if $(D(\hat{m}_0) - D(\hat{m}))/(p-p_0)$ is a credible estimate of $\phi \equiv 1$ under the reduced model, it makes sense to reject the reduced model whenever the estimate is much larger than 1.

Both Poisson regression and logistic regression fit into this one-parameter framework. For Poisson regression, the deviance is G^2 and often can be used for lack of fit tests. In both Poisson and logistic regression, the asymptotic χ^2 approximation for the test of a model against a reduced model is often valid. However, we have seen that the lack of fit statistic for logistic regression is typically *not* asymptotically χ^2. Care must be used in applying the asymptotic results given above. The specifics of each situation must be considered.

For generalized linear models with a nontrivial dispersion parameter, we can only test reduced models against larger models. An appealing asymptotic test is to reject the adequacy of the reduced model at the α level if

$$\frac{\big(D(\hat{m}_0) - D(\hat{m})\big)/(p-p_0)}{D(\hat{m})/(n-p)} > F(1-\alpha, p-p_0, n-p).$$

This relies not only on $\big(D(\hat{m}_0) - D(\hat{m})\big)/\phi$ and $D(\hat{m})/\phi$ being asymptotically χ^2 but also on them being asymptotically independent. As discussed

earlier, the χ^2 approximation to the distribution of $D(\hat{m})/\phi$ frequently requires asymptotics based on fixed n. For normal theory models, this is the usual F test.

If (a) n is large, (b) $D(\hat{m})/(n-p)$ is a consistent estimate of ϕ, and (c) $\big(D(\hat{m}_0) - D(\hat{m})\big)/\phi$ is asymptotically χ^2, we get the asymptotic null distribution

$$\frac{\big(D(\hat{m}_0) - D(\hat{m})\big)/(p-p_0)}{D(\hat{m})/(n-p)} \sim \frac{\chi^2(p-p_0)}{p-p_0}$$

with a corresponding test. If $X^2/(n-p)$ is a consistent estimate of ϕ, the Pearson estimate can be used in the denominator of the asymptotic test. This is of particular importance when $D(\hat{m})/(n-p)$ is not consistent but $X^2/(n-p)$ is. Again, a less formal evaluation can be made if $\big(D(\hat{m}_0) - D(\hat{m})\big)/(p-p_0)$ is a plausible estimate of ϕ under the reduced model. The reduced model is called in question when the test statistic is much larger than 1.

Note that as $n-p$ approaches infinity, the $F(p-p_0, n-p)$ distribution approaches a $\chi^2(p-p_0)/(p-p_0)$. Even though the appropriate asymptotic distribution for large n is a rescaled χ^2, for data analytic purposes it may not be unreasonable to use F tables instead.

Normal linear models and gamma distribution regression both fit into the nontrivial dispersion parameter framework. As always, appropriate conditions must be met for the asymptotic results to be valid. For normal theory linear models, the deviance and Pearson statistic both equal the error sum of squares and the F distribution is exact. It does not rely on any asymptotic arguments.

9.4 Summary and Discussion

Without a doubt, iteratively reweighted least squares for generalized linear models is a remarkably useful computing device. A review of the use of iterative generalized least squares in statistical estimation is given by del Pino (1989). Iteratively reweighted least squares is a special case of iterative generalized least squares.

Generalized linear models are designed to treat independent observations that have a distribution in the one-parameter exponential family. They also provide maximum likelihood estimates of appropriate functions of the β parameters when each observation has a distribution in a particular family of two-parameter distributions. This two-parameter family is chosen so that a trick commonly used in estimation for normal theory linear models works for the entire family. The trick is that maximum likelihood estimates of β can be found easily for any value of ϕ. These estimates do not depend on ϕ; therefore, the estimates must be maximum likelihood even when ϕ is an unknown parameter. Given the maximum likelihood estimates of β,

finding the maximum likelihood estimate of ϕ requires solving one equation: $d\ell(\hat{\beta};\phi)/d\phi = 0$. This method is used by Christensen (1996b, Section 2.4) to find maximum likelihood estimates for normal theory linear models. Maximum likelihood estimation of ϕ seems preferable to the essentially ad hoc methods that are illustrated above.

Of the examples that we have considered, Poisson regression and logistic regression are generalized linear models for one-parameter exponential families. Poisson sampling seems to be relatively uncommon for contingency tables; the standard sampling schemes are multinomial and product-multinomial. However, under very mild conditions, maximum likelihood estimates for Poisson sampling are also maximum likelihood estimates for the other sampling schemes. Of course, logistic regression can also be viewed as a special case of product-multinomial log-linear modeling. In regard to Poisson sampling, Santner and Duffy (1989, Problem 3.3) present an interesting data set. The data, originally given in Quine (1975), are on the number of absences of 113 Australian school children. The data are categorized using four factors: age at three levels, sex, cultural background (aboriginal, white), and learning ability (slow, average). The number of absences for different children might be considered as observations on independent Poisson random variables. Note that the number of cross-classifications from the four factors is $3 \times 2 \times 2 \times 2 = 24$, but there are 113 Poisson observations. The analysis of such data would be analogous to a four-factor analysis of variance with unequal numbers of replications on the various treatments. Moreover, Santner and Duffy (1989, p. 135) suggest that the data suffer from *overdispersion*, i.e., $\phi > 1$. It is interesting to note that an observed value of $X^2/(n-p)$ much larger than 1 can indicate *either* lack of fit *or* overdispersion. See McCullagh and Nelder (1989, Sections 4.5, 5.5) for discussion of overdispersion.

The other two examples considered in this chapter, normal theory linear models and gamma distribution regression, involve the two-parameter family of distributions that was used in the basic theory. Generalized linear model methods can be used to analyze other useful models; see McCullagh and Nelder (1989) for a broad range of applications.

While generalized linear models are a useful idea and provide an excellent computing device, care must be taken in their application. For log-linear models, Poisson sampling does not always lead to the same analysis as multinomial and product-multinomial sampling. The distinctions as well as the similarities must be kept in mind. The validity of asymptotic distributions must also be examined carefully. As seen in Section 11.2, a careful analysis of asymptotic issues can be quite complicated. Some extensions of the basic theory such as overdispersion, e.g., allowing ϕ to be a nondegenerate parameter in binomial and Poisson sampling, and *quasi-likelihood* methods have been proposed. Such extensions are widely accepted as providing valuable data analytic tools; however, many people have difficulty in understanding the theoretical basis for them.

9.5 Exercises

EXERCISE 9.5.1. Show that for the binomial model of Section 1, $r(\theta) = \log(1 + e^\theta)$ and that $\dot{r}(\theta) = p$ and $\ddot{r}(\theta) = p(1 - p)$.

EXERCISE 9.5.2. For the gamma model of Section 1, use the definition of θ and $r(\theta)$ to show that the mean and variance are as given.

EXERCISE 9.5.3. Show that the likelihood equations (9.2.2) have the form $Q_j(\beta)/\phi = 0$, $j = 1, \ldots, p$.

EXERCISE 9.5.4. Show that if $f(y_i|\theta, \phi; w)$ from (9.1.1) is the common density of independent observations y_i, $i = 1, \ldots, n$, then $\sum_{i=1}^n y_i$ has a density $f(y_i|\theta_*, \phi_*, w_*)$ for some θ_*, ϕ_*, and w_*.

EXERCISE 9.5.5. Let $Y = (y_1, \ldots y_n)'$. Show that a generalized linear model with canonical link has $X'Y$ as a sufficient statistic.

EXERCISE 9.5.6. Using the definitions of this chapter, find the Pearson statistic (defined in Section 3) for Poisson and binomial regression in terms of the y_i's and m_i's. Show that these are identical to the Pearson statistics defined in Chapter 2.

10

The Matrix Approach to Log-Linear Models

Analysis of variance and regression analysis are both branches of linear model theory. Regression analysis and linear model theory are usually taught using matrices. It is less common to teach analysis of variance with matrices. Although standard log-linear model theory is analogous to analysis of variance, the basic results are more easily stated in matrix notation. It is assumed that the reader is familiar with the basics of using matrices.

We begin with some simple examples of writing log-linear models with matrices.

EXAMPLE 10.0.1. Consider a 3×4 table. The log-linear model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}\,, \qquad i = 1, \ldots 3\,,\ j = 1, \ldots, 4\,, \tag{1}$$

can be written in matrix form as

$$\begin{bmatrix} \log(m_{11}) \\ \log(m_{12}) \\ \log(m_{13}) \\ \log(m_{14}) \\ \log(m_{21}) \\ \log(m_{22}) \\ \log(m_{23}) \\ \log(m_{24}) \\ \log(m_{31}) \\ \log(m_{32}) \\ \log(m_{33}) \\ \log(m_{34}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{1(3)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{2(4)} \end{bmatrix}.$$

The log-linear model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{2}$$

can be written in matrix form as

$$\begin{bmatrix} \log(m_{11}) \\ \log(m_{12}) \\ \log(m_{13}) \\ \log(m_{14}) \\ \log(m_{21}) \\ \log(m_{22}) \\ \log(m_{23}) \\ \log(m_{24}) \\ \log(m_{31}) \\ \log(m_{32}) \\ \log(m_{33}) \\ \log(m_{34}) \end{bmatrix} = X \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{1(3)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{2(4)} \\ u_{12(11)} \\ u_{12(12)} \\ u_{12(13)} \\ u_{12(14)} \\ u_{12(21)} \\ u_{12(22)} \\ u_{12(23)} \\ u_{12(24)} \\ u_{12(31)} \\ u_{12(32)} \\ u_{12(33)} \\ u_{12(34)} \end{bmatrix}$$

where

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The uniform association model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma x_i w_j$$

can be written

$$\begin{bmatrix}\log(m_{11})\\ \log(m_{12})\\ \log(m_{13})\\ \log(m_{14})\\ \log(m_{21})\\ \log(m_{22})\\ \log(m_{23})\\ \log(m_{24})\\ \log(m_{31})\\ \log(m_{32})\\ \log(m_{33})\\ \log(m_{34})\end{bmatrix} = \begin{bmatrix}1&1&0&0&1&0&0&0&x_1w_1\\ 1&1&0&0&0&1&0&0&x_1w_2\\ 1&1&0&0&0&0&1&0&x_1w_3\\ 1&1&0&0&0&0&0&1&x_1w_4\\ 1&0&1&0&1&0&0&0&x_2w_1\\ 1&0&1&0&0&1&0&0&x_2w_2\\ 1&0&1&0&0&0&1&0&x_2w_3\\ 1&0&1&0&0&0&0&1&x_2w_4\\ 1&0&0&1&1&0&0&0&x_3w_1\\ 1&0&0&1&0&1&0&0&x_3w_2\\ 1&0&0&1&0&0&1&0&x_3w_3\\ 1&0&0&1&0&0&0&1&x_3w_4\end{bmatrix} \begin{bmatrix}u\\ u_{1(1)}\\ u_{1(2)}\\ u_{1(3)}\\ u_{2(1)}\\ u_{2(2)}\\ u_{2(3)}\\ u_{2(4)}\\ \gamma\end{bmatrix}.$$

A matrix with only one column will be referred to as a vector. Let $x = (x_1, \ldots, x_q)'$ be a vector. Define

$$\log(x) = (\log(x_1), \log(x_2), \ldots, \log(x_q))'.$$

Consider a table with any number of dimensions that has q cells in it. For a 3×4 table, $q = 12$. For an $I \times J \times K$ table, $q = IJK$. The expected cell counts are denoted by the vector $m = (m_1, \ldots, m_q)'$. A log-linear model is a model

$$\log(m) = X\beta$$

where $\log(m)$ is a $q \times 1$ vector of unknown parameters, X is a $q \times p$ matrix with known values (often X consists entirely of 0s and 1s), and β is a $p \times 1$ vector of unknown parameters. In Example 10.0.1, the log-linear model (1) has an X matrix with 12 rows and 8 columns that consists entirely of 0s and 1s. The β vector was the 8×1 matrix $(u, u_{1(1)}, u_{1(2)}, u_{1(3)}, u_{2(1)}, u_{2(2)}, u_{2(3)}, u_{2(4)})'$. For model (2), the X matrix has 12 rows and 20 columns. The β vector is a 20×1 matrix that contains u, the $u_{1(i)}$'s, the $u_{2(j)}$'s, and the $u_{12(ij)}$'s.

EXAMPLE 10.0.2. Consider a $2 \times 3 \times 2$ table. The model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

can be written

$$\begin{bmatrix} \log(m_{111}) \\ \log(m_{112}) \\ \log(m_{121}) \\ \log(m_{122}) \\ \log(m_{131}) \\ \log(m_{132}) \\ \log(m_{211}) \\ \log(m_{212}) \\ \log(m_{221}) \\ \log(m_{222}) \\ \log(m_{231}) \\ \log(m_{232}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{3(1)} \\ u_{3(2)} \end{bmatrix}$$

$$\log(m) \quad = \quad X \quad \beta\,.$$

The model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

can be written

$$\begin{bmatrix} \log(m_{111}) \\ \log(m_{112}) \\ \log(m_{121}) \\ \log(m_{122}) \\ \log(m_{131}) \\ \log(m_{132}) \\ \log(m_{211}) \\ \log(m_{212}) \\ \log(m_{221}) \\ \log(m_{222}) \\ \log(m_{231}) \\ \log(m_{232}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ u_{1(1)} \\ u_{1(2)} \\ u_{2(1)} \\ u_{2(2)} \\ u_{2(3)} \\ u_{3(1)} \\ u_{3(2)} \\ u_{23(11)} \\ u_{23(12)} \\ u_{23(21)} \\ u_{23(22)} \\ u_{23(31)} \\ u_{23(32)} \end{bmatrix}$$

$$\log(m) \quad = \quad X \quad \beta\,.$$

EXERCISE 10.1. For a 3×4 table, write model (7.1.7) in the form $\log(m) = X\beta$.

EXERCISE 10.2. For a $2 \times 3 \times 2$ table, write the models [2][13], [13][23], [12][23][13], and [123] in the form $\log(m) = X\beta$.

One advantage of establishing results for a general log-linear model $\log(m) = X\beta$ is the flexibility of the model. Results apply to ANOVA type models for any number of dimensions. The X matrix can be any known matrix, so models that incorporate known scores for ordered categories or use predictor variables to model interactions are also special cases of the general log-linear model.

In this chapter, we present a summary of some basic results in maximum likelihood theory for log-linear models. Most of the results are presented more rigorously in Chapter 12. In Sections 1 and 2, results are presented for multinomial sampling. In Section 3, the extension to product-multinomial sampling is discussed. Section 4 discusses drawing inferences about model parameters. Section 5 examines the Newton-Raphson alternative to iterative proportional fitting for finding MLEs. Section 6 discusses the GSK method of fitting log-linear models. Section 7 considers residual analysis.

10.1 Maximum Likelihood Theory for Multinomial Sampling

Suppose we have a table with q cells, observations $n = (n_1, \ldots, n_q)'$, and the log-linear model $\log(m) = X\beta$ holds. Under multinomial sampling, the likelihood function is

$$L(p) = \frac{n.!}{\prod_{i=1}^{q} n_i!} \prod_{i=1}^{q} p_i^{n_i}$$

where $p = (p_1, \ldots, p_q)'$. Equivalently, we can write this as a function of m because $m_i = n.p_i$ and $n.$ is the known sample size. In terms of the m_i's, the likelihood becomes

$$L(m) = \frac{n.!}{\prod_{i=1}^{q} n_i!} \prod_{i=1}^{q} (m_i/n.)^{n_i} .$$

Estimation

Maximum likelihood estimates (MLEs) are values $\hat{m}_i$ that maximize $L(m)$ subject to the constraints of our model. There are two constraints on the model: One is the log-linear structure

$$\log(m) = X\beta \quad \text{for some} \quad \beta \tag{1}$$

and the other relates to the fact that with multinomial sampling, $1 = \sum_{i=1}^{q} p_i$. This second condition is equivalent to $n. = m.$. Let J be a $q \times 1$ vector consisting entirely of 1s. The condition $n. = m.$ can be written as

$$n'J = m'J . \tag{2}$$

Rather than maximizing $L(m)$, it is simpler to maximize the log of $L(m)$,

$$\log L(m) = \log(n.!) - \sum_{i=1}^{q} \log(n_i!) + \sum_{i=1}^{q} n_i \log(m_i) - \sum_{i=1}^{q} n_i \log(n.)\,.$$

The only term that involves the m_i's is $\sum_{i=1}^{q} n_i \log(m_i) = n' \log(m)$, so it is enough to maximize

$$\ell(m) \equiv n' \log(m)\,.$$

The MLE, $\hat{m}$, is the value that maximizes $\ell(m)$ subject to conditions (1) and (2). In other words, $\hat{m}$ must have the properties that

$$\log(\hat{m}) = X\hat{\beta} \quad \text{for some } \hat{\beta}\,, \tag{3}$$

$$n'J = \hat{m}'J \tag{4}$$

and if $\tilde{m}$ is any other vector with $\log(\tilde{m}) = X\tilde{\beta}$ and $n'J = \tilde{m}'J$, then

$$\ell(\tilde{m}) \leq \ell(\hat{m})\,.$$

It turns out that for a broad class of possible X matrices, the maximization can be performed without imposing condition (2). As will be discussed below, this occurs because the maximum of $\ell(m)$ subject only to condition (1) automatically satisfies condition (2). A standard method for finding the maximum of $\ell(m)$ subject to condition (1) is by setting appropriate partial derivatives equal to zero. It can be shown that the partial derivatives are zero at the point $\hat{m}$ that satisfies

$$n'X = \hat{m}'X, \tag{5}$$

cf. Chapter 12. Moreover, by considering the matrix of second partial derivatives, it can be shown that if $\ell(m)$ achieves its maximum, subject to the constraint $\log(m) = X\beta$, then it will be at the unique value $\hat{m}$ that satisfies conditions (3) and (5). In other words, *any value $\hat{m}$ that satisfies the (marginal) constraints (5) and the model (3) is the maximum likelihood estimate of m provided a maximum exists.* This point was made repeatedly in Section 3.2.

If X is chosen appropriately, then any $\hat{m}$ that satisfies conditions (3) and (5) automatically satisfies condition (2), i.e., satisfies (4). Before examining this claim, we introduce a very useful concept in log-linear model theory, the column space of X. The column space of X is defined to be the set

$$C(X) = \{\mu | \mu = X\beta \quad \text{for some} \quad \beta\}\,.$$

Thus, $C(X)$ consists of all of the possible values for $\log(m)$ that satisfy the log-linear model. Earlier, we discussed the fact that the models

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \tag{6}$$

and

$$\log(m_{ij}) = u_{12(ij)} + u_{13(ik)} \tag{7}$$

are equivalent. If we write the model in equation (6) as

$$\log(m) = X_1\beta_1$$

and the model in equation (7) as

$$\log(m) = X_2\beta_2\,,$$

it is not difficult to show that $C(X_1) = C(X_2)$. In other words, the possible values for $\log(m)$ are identical in models (6) and (7). That is why the models are equivalent.

We now return to the claim that if $\hat{m}$ satisfies conditions (3) and (5), then for an appropriate X matrix, condition (4) is also satisfied. Suppose that $J \in C(X)$. In other words, for some vector b, $J = Xb$. If $\hat{m}$ satisfies condition (5), then it follows that

$$n'J = n'Xb = \hat{m}'Xb = \hat{m}'J;$$

hence, condition (4) is satisfied. Thus, if $J \in C(X)$ and if the MLE of m exists, then we can find the MLE by finding $\hat{m}$ that satisfies conditions (3) and (5).

We will not give a detailed discussion concerning when MLEs exist; for such a discussion, see Haberman (1974a). However, we will mention one result. It is an immediate consequence of Theorem 12.2.1 that if $n_i > 0$ for all $i = 1, \ldots, q$, then the MLEs exist.

The condition imposed above on X, i.e., $J \in C(X)$, is not an onerous condition. It simply means that the model has a parameter u (with no subscripts) or that the model is equivalent to a model that contains a u term. The condition $J \in C(X)$ is also necessary for the asymptotic results discussed in Section 2 and Chapter 12. We will henceforth always assume that $J \in C(X)$.

One final point: The MLE of m does not really depend on X, it depends on $C(X)$. Any two parametrizations $\log(m) = X_1\beta_1$ and $\log(m) = X_2\beta_2$ with $C(X_1) = C(X_2)$ have exactly the same MLE of m.

EXAMPLE 10.1.1. One version of the three-dimensional saturated model is $\log(m_{ijk}) = u_{123(ijk)}$. If this is written in matrix form, $X = I_q$ where I_q is the $q \times q$ identity matrix and $q = IJK$. The conditions for MLEs become

$$\log(\hat{m}) = I_q\hat{\beta} = \hat{\beta} \quad \text{for some} \quad \hat{\beta}$$

and

$$n'I_q = \hat{m}'I_q\,.$$

Clearly, $\hat{m} = n$ satisfies the second of these equations and $\hat{\beta} = \log(n)$ satisfies the first equation. Thus, the MLE of m is $\hat{m} = n$ in a three-dimensional saturated model. In fact, this argument is valid for any saturated model. The idea of a saturated model is that there are enough parameters to explain the data perfectly. This translates to the idea that $C(X) = C(I_q) = \mathbf{R}^q$.

EXAMPLE 10.1.2. Consider a three-dimensional table and the model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} .$$

If one writes out the matrix X (cf. Example 10.0.2), it is easily seen that the condition $n'X = m'X$ is precisely $n_{\cdots} = \hat{m}_{\cdots}$, $n_{i\cdot\cdot} = \hat{m}_{i\cdot\cdot}$, $n_{\cdot j\cdot} = \hat{m}_{\cdot j\cdot}$, $n_{\cdot\cdot k} = \hat{m}_{\cdot\cdot k}$, $n_{\cdot jk} = \hat{m}_{\cdot jk}$. Several of these conditions are redundant. It is sufficient to have $n_{i\cdot\cdot} = \hat{m}_{i\cdot\cdot}$ and $n_{\cdot jk} = \hat{m}_{\cdot jk}$. The other relationships follow from these two. If the model is reparametrized as

$$\log(m_{ijk}) = u_{1(i)} + u_{23(jk)} ,$$

then the condition $n'X = m'X$ for the new X matrix gives only the conditions $n_{i\cdot\cdot} = \hat{m}_{i\cdot\cdot}$ and $n_{\cdot jk} = \hat{m}_{\cdot jk}$. The MLEs of the m_{ijk}'s are precisely the values $\hat{m}_{ijk}$ that satisfy $n_{i\cdot\cdot} = \hat{m}_{i\cdot\cdot}$ and $n_{\cdot jk} = \hat{m}_{\cdot jk}$ and can be written as

$$\log(\hat{m}_{ijk}) = \hat{u}_{1(i)} + \hat{u}_{23(jk)}$$

for some values $\hat{u}_{1(i)}$ and $\hat{u}_{23(jk)}$. It is easily seen that if $\hat{m}_{ijk} = n_{i\cdot\cdot} n_{\cdot jk}/n_{\cdots}$, these conditions are satisfied.

EXERCISE 10.3.

(a) For a 3×4 table, find the conditions that the MLEs must satisfy in model (7.1.4).

(b) Repeat (a) for model (7.1.7).

Testing Hypotheses

One of the tests that is allowed for three-way tables is the test of [1][2][3] versus [1][23]. If these models are written as $\log(m) = X_0\beta_0$ and $\log(m) = X\beta$, respectively, the fact that [1][23] is a larger model is reflected in the fact that $C(X_0) \subset C(X)$. Recall that for a $2 \times 3 \times 2$ table, X_0 and X where presented in Example 10.0.2.

In general, if we assume that $\log(m) = X\beta$ holds and that $C(X_0) \subset C(X)$, we can test the hypothesis

$$H_0 : \log(m) = X_0\beta_0$$

against the hypothesis

$$H_A : (H_0 \text{ is not true}) .$$

The likelihood ratio test statistic is

$$G^2 = -2[\log L(\hat{m}_0) - \log L(\hat{m})]$$

where $\hat{m}_0$ is the MLE of m under the assumption that H_0 is true and $\hat{m}$ is the MLE under the "unrestricted" model. However, in this case, the "unrestricted" model is that $\log(m) = X\beta$. It is easily seen that

$$\begin{aligned} G^2 &= -2[\ell(\hat{m}_0) - \ell(\hat{m})] \\ &= -2[n' \log(\hat{m}_0) - n' \log(\hat{m})] \\ &= 2n'[\log(\hat{m}) - \log(\hat{m}_0)] \\ &= 2\sum_{i=1}^{q} n_i \log(\hat{m}_i/\hat{m}_{0i}) \end{aligned}$$

where again $\hat{m} = (\hat{m}_1, \ldots, \hat{m}_q)'$ is the MLE of m for $\log(m) = X\beta$ and $\hat{m}_0 = (\hat{m}_{01}, \ldots, \hat{m}_{0q})'$ is the MLE of m under the restriction that $\log(m) = X_0\beta_0$.

In fact, G^2 can be written as

$$G^2 = 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{0i}),$$

which is our usual form. To see this equivalence, note that $\log(\hat{m}) = X\hat{\beta}$ for some $\hat{\beta}$, and because $C(X_0) \subset C(X)$, we can write $\log(\hat{m}_0) = X_0\hat{\beta}_0 = X\hat{\gamma}$ for some $\hat{\beta}_0$ and $\hat{\gamma}$. Recall that $\hat{m}'X = n'X$. By substitution,

$$\begin{aligned} G^2 = 2n'[\log(\hat{m}) - \log(\hat{m}_0)] &= 2n'[X\hat{\beta} - X\hat{\gamma}] \\ &= 2n'X[\hat{\beta} - \hat{\gamma}] \\ &= 2\hat{m}'X[\hat{\beta} - \hat{\gamma}] \\ &= 2\hat{m}'[\log(\hat{m}) - \log(\hat{m}_0)] \\ &= 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{0i})\,. \end{aligned}$$

10.2 Asymptotic Results

This section presents a few of the primary asymptotic results for log-linear models under multinomial sampling and mentions some applications of those results. More precise versions of these results are available in Chapter 12.

We begin by setting some notation. Let x be a $q \times 1$ vector. $D(x)$ is used to denote the $q \times q$ diagonal matrix

$$D(x) = [d_{ij}] \qquad \text{where } d_{ii} = x_i\,,\ d_{ij} = 0, \qquad i \neq j\,.$$

One diagonal matrix is used often and has a special notation:

$$D \equiv D(p)\,.$$

Recall that J is a $q \times 1$ vector of 1s. Define

$$A = X(X'DX)^{-1}X'D$$

and

$$A_z = J(J'DJ)^{-1}J'D$$

where it is assumed (but not really necessary) that a parametrization $\log(m) = X\beta$ has been chosen so that the inverse of $X'DX$ exists. Note that because $D(m) = n.D$, D can be replaced by $D(m)$ in A and A_z without changing the resulting matrices. Note that A and A_z depend on the unknown parameters p. We can estimate A and A_z simply by estimating p.

Rather than frequently writing $\log(m)$, let

$$\mu \equiv \log(m)\,.$$

If $\hat{m}$ is the MLE of m, $\hat{\mu} = \log(\hat{m})$ is the MLE of the μ. This follows from the *invariance of maximum likelihood estimates*; for any parameter θ and MLE $\hat{\theta}$, the MLE of a function of θ, say $f(\theta)$, is the corresponding function of the MLE, $f(\hat{\theta})$, cf. Cox and Hinkley (1974, p. 287).

The key asymptotic results about MLEs are given in the following subsections. Throughout, let $N \equiv n..$

Estimation

We begin with results about the large sample distribution of the maximum likelihood estimates.

Theorem 10.2.1. Let $\mu = X\beta$ be a log-linear model for a table with q cells. Let n be the result of a multinomial sample of N observations:

(a) For N sufficiently large, $\hat{\mu} - \mu$ has the approximate distribution $N(0, [A - A_z]D^{-1}(m))$.

(b) As N gets large, $\hat{\mu} - \mu$ converges (in probability) to zero; i.e., $\hat{\mu} - \mu \xrightarrow{P} 0$.

(c) For N sufficiently large, $\hat{m} - m$ has the approximate distribution $N(0, D(m)[A - A_z])$.

(d) As N gets large, $\hat{m}/N$ converges (in probability) to p, i.e., $N^{-1}\hat{m} \xrightarrow{P} p$.

Technically, (a) and (c) deal with convergence in distribution and are similar in spirit to the Central Limit Theorem. In Chapter 11, we will have occasion to write such results as (a) $N^{\frac{1}{2}}(\hat{\mu}-\mu) \xrightarrow{L} N(0,[A-A_z]D^{-1})$ and (c) $N^{-1/2}(\hat{m}-m) \xrightarrow{L} N(0,D[A-A_z])$. The symbol $\xrightarrow{L}$ indicates convergence in distribution. The L comes from the fact that a distribution is sometimes referred to as a distributional law or simply as a law.

One interesting aspect of Theorem 10.2.1 is that, although $\hat{\mu}-\mu$ converges to zero, $\hat{\mu}$ by itself does not converge to anything. As N gets large, $\mu = \log(m) = \log(Np)$ also gets large. Although the difference $\hat{\mu}-\mu$ gets small, we cannot say that $\hat{\mu}$ converges to μ because μ *changes with* N.

Corollary 10.2.2. If the inverse of $(X'DX)$ exists and $\hat{\beta}$ satisfies $\hat{\mu} = X\hat{\beta}$, then $\hat{\beta}-\beta$ converges (in probability) to zero.

Consider the problem of drawing asymptotic inferences about a particular cell. The parameters of interest are $p_i,\ m_i$, and μ_i. We will start from the premise that estimates of the m_i's are available. These can be obtained from iterative proportional fitting as discussed in Section 3.3 or from use of the Newton-Raphson algorithm as discussed later in Section 5. Recall that

$$\hat{m} = (\hat{m}_1, \ldots, \hat{m}_q)',$$

$$\hat{\mu} = \log(\hat{m}) = (\log(\hat{m}_1), \ldots, \log(\hat{m}_q))',$$

and

$$\hat{p} = \frac{1}{N}\hat{m} = (\hat{m}_1/N, \ldots, \hat{m}_q/N)'.$$

To use Theorem 10.2.1, we need one key result. If Y is a $q \times 1$ vector with a multivariate normal distribution, i.e.,

$$Y \sim N(\xi, \Sigma),$$

and if ρ is a $q \times 1$ vector, then the scalar random variable $\rho'Y$ has a (univariate) normal distribution. In particular,

$$\rho'Y \sim N(\rho'\xi, \rho'\Sigma\rho).$$

Let $e_i' = (0, \cdots, 0, 1, 0, \cdots, 0)$ where the 1 is in the ith place. It follows that

$$\begin{aligned} \hat{\mu}_i - \mu_i &= e_i'(\hat{\mu}-\mu), \\ \hat{m}_i - m_i &= e_i'(\hat{m}-m), \end{aligned}$$

and

$$\hat{p}_i - p_i = e_i'(\hat{p}-p).$$

Applying Theorem 10.2.1, for N large we get the approximations

$$\begin{aligned} \hat{\mu}_i - \mu_i &\sim N\left(0, e_i'[A - A_z]D^{-1}(m)e_i\right) , \\ \hat{m}_i - m_i &\sim N\left(0, e_i'D(m)[A - A_z]e_i\right) , \end{aligned}$$

and

$$\hat{p}_i - p_i \sim N\left(0, e_i'\frac{1}{N^2}D(m)[A - A_z]e_i\right) .$$

In order to use these results, we need to be able to find or at least estimate the variances. We begin with $e_i'[A - A_z]D^{-1}(m)e_i$. This value is the ith diagonal element of $A\,D^{-1}(m)$ minus the ith diagonal element of $A_zD^{-1}(m)$. To find $e_i'AD^{-1}(m)e_i$, note that

$$D^{-1}(m)e_i = \left(\frac{1}{m_i}\right)e_i,$$

so

$$e_i'AD^{-1}(m)e_i = \frac{1}{m_i}\,e_i'Ae_i\,.$$

The value $e_i'Ae_i$ is just a_{ii}, the ith diagonal element of A. This is precisely the leverage of the ith case. Leverages were introduced in Section 6.7 and methods for estimating them were given. The maximum likelihood estimate of

$$e_i'AD^{-1}(m)e_i = a_{ii}/m_i$$

is

$$\hat{a}_{ii}/\hat{m}_i\,.$$

The computation of $e_i'A_zD^{-1}(m)e_i$ is even simpler. For multinomial sampling,

$$A_z \equiv J(J'DJ)^{-1}J'D = JJ'D\,.$$

This follows because $J'DJ = p_\cdot = 1$. Moreover,

$$\begin{aligned} A_zD^{-1}(m) &= JJ'DD^{-1}(m) \\ &= JJ'\left(\frac{1}{N}\right)D(m)D^{-1}(m) \\ &= \left(\frac{1}{N}\right)JJ', \end{aligned}$$

so

$$e_i'A_zD^{-1}(m)e_i = \frac{1}{N}\,.$$

Combining results, we see that

$$e_i'[A - A_z]D^{-1}(m)e_i = \frac{a_{ii}}{m_i} - \frac{1}{N};$$

thus,

$$\hat{\mu}_i - \mu_i \sim N\left(0, \frac{a_{ii}}{m_i} - \frac{1}{N}\right).$$

Estimating the variance leads to the approximation

$$(\hat{\mu}_i - \mu_i) \Big/ \sqrt{\frac{\hat{a}_{ii}}{\hat{m}_i} - \frac{1}{N}} \sim N(0,1).$$

Large sample confidence intervals for μ_i follow immediately, e.g., a 95% confidence interval has the end points

$$\hat{\mu}_i \pm 1.96\sqrt{\frac{\hat{a}_{ii}}{\hat{m}_i} - \frac{1}{N}}.$$

The $\alpha = .10$ large sample test of $H_0 : \mu_i = \mu_{i0}$ versus $H_A : \mu_i \neq \mu_{i0}$ rejects when

$$|\hat{\mu}_i - \mu_{i0}| \Big/ \sqrt{\frac{a_{ii}}{m_i} - \frac{1}{N}} > 1.645.$$

Similar arguments lead to the asymptotic results

$$\mathrm{Var}(\hat{m}_i) = m_i a_{ii} - m_i^2/N$$

and

$$\mathrm{Var}(\hat{p}_i) = p_i a_{ii}/N - p_i^2/N.$$

Estimating the variances yields to the large sample distributions

$$\frac{\hat{m}_i - m_i}{\sqrt{\hat{m}_i \hat{a}_{ii} - \hat{m}_i^2/N}} \sim N(0,1)$$

and

$$\frac{\hat{p}_i - p_i}{\sqrt{\hat{p}_i(\hat{a}_{ii} - \hat{p}_i)/N}} \sim N(0,1).$$

Given the distributions, inferential procedures follow in the usual way.

Just as in regression analysis, leverages fall between zero and one and the sum of all of the leverages is precisely the degrees of freedom for the model, i.e., the rank of X. The first of these facts implies that an upper bound on the variance can always be obtained by taking $a_{ii} = 1$. This is convenient because when iterative proportional fitting has been used, finding $\hat{a}_{ii}$ requires the computation of an auxiliary regression analysis. Assuming $a_{ii} = 1$ can be highly conservative because the true a_{ii} value may be much less than one. The second fact gives some idea of the extent of overestimation using $a_{ii} = 1$. If the table has $q = 24$ cells and the model has 12 degrees of freedom, the average size of the a_{ii}'s is $12/24 = \frac{1}{2}$. Thus, the variance terms based on $a_{ii} = 1$ tend to be about twice as large as they are using the estimates $\hat{a}_{ii}$.

It is interesting to note that using the upper bound $a_{ii} = 1$ is equivalent to computing the variance under the saturated model. In the saturated model, we can take $X = I$. This implies that

$$A = I(IDI)^{-1}ID = I\,.$$

Thus, for all i under the saturated model, $a_{ii} = 1$. Clearly, the use of reduced models serves to reduce the variance of estimated cell parameters.

EXAMPLE 10.2.3. In the abortion opinion data of Chapter 3 with the model [RSO][OA] (cf. Table 6.7), the cell for nonwhite males between 18 and 25 years of age who support abortion has $\hat{m}_i = 14.52$ and $\hat{a}_{ii} = .222$. The asymptotic standard error for $\hat{m}_i$ is $\sqrt{14.52(.222) - (14.52)^2/2385} = \sqrt{3.2234 - .0884} = 1.77$. An asymptotic 95% confidence interval for m_i has end points

$$14.52 \pm 1.96(1.77)\,.$$

The interval is $(11.05, 17.99)$. Similar computations lead to a 95% confidence interval for μ_i with end points

$$2.68 \pm 1.96(.123)$$

and a 90% confidence interval for p_i with end points

$$.0061 \pm 1.645(.000742)\,.$$

Besides the parameters for individual cells, the parameters of primary interest are contrasts in the μ_i's. Contrasts in the μ_i's correspond to vectors ρ in which the elements of ρ add up to zero, i.e., $\rho' J = 0$. The simplest such contrasts are log odds, but log odds ratios, the log of ratios of odds ratios, and so on, are also contrasts in the μ_i's. All of these correspond to functions $\rho'\mu$ in which ρ has a very simple structure. Given the $\hat{m}_i$'s, there is no problem in computing $\rho'\hat{\mu} = \rho' \log(\hat{m})$. The problem is in computing the variance. Finding variances for estimated contrasts is more complicated than finding them for estimates of cell parameters because contrasts involve the covariances between the estimated cell parameters. However, the fact that we are dealing with contrasts leads to one simplification based on $\rho' J = 0$.

$$\begin{aligned}
\mathrm{Var}(\rho'\hat{\mu}) &= \rho'(A - A_z)D^{-1}(m)\rho \\
&= \rho' A D^{-1}(m)\rho - \rho' A_z D^{-1}(m)\rho \\
&= \rho' X(X'D(m)X)^{-1}X'\rho - \frac{1}{N}\rho' JJ'\rho \\
&= \rho' X(X'D(m)X)^{-1}X'\rho\,.
\end{aligned}$$

Computation of the variance requires fitting the model using the Newton-Raphson algorithm, cf. Section 5. Newton-Raphson can either be used exclusively or, if the initial fit was performed using iterative proportional fitting, the auxiliary regression model of Section 6.7 can be used. Recall that the auxiliary model requires that an ANOVA type model be reparametrized as a regression model. This is so that appropriate matrix inverses can be taken. If traditional ANOVA type models are used, a simple way to generate a regression model is to drop all u terms involving index values of 1.

EXAMPLE 10.2.4. In Example 3.2.4, we examined data on automobile injuries. We found that the model of no three-factor interaction,

$$\mu_{ijk} = \log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}\,,$$

fit the data very well. Below are given the data and the estimated expected cell counts based on the model.

$n_{ijk}(\hat{m}_{ijk})$		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver	No	350 (350.49)	150 (149.51)	60 (59.51)	112 (112.49)
Ejected (i)	Yes	26 (25.51)	23 (23.49)	19 (19.49)	80 (79.51)

The regression parametrization based on dropping u terms in which any of i, j, or k equal 1 is

$$\begin{bmatrix} \hat{\mu}_{111} \\ \hat{\mu}_{121} \\ \hat{\mu}_{112} \\ \hat{\mu}_{122} \\ \hat{\mu}_{211} \\ \hat{\mu}_{221} \\ \hat{\mu}_{212} \\ \hat{\mu}_{222} \end{bmatrix} = \begin{bmatrix} \log(350.49) \\ \log(149.51) \\ \log(\ 59.51) \\ \log(112.49) \\ \log(\ 25.51) \\ \log(\ 23.49) \\ \log(\ 19.49) \\ \log(\ 79.51) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma} \\ \hat{\gamma}_{1(2)} \\ \hat{\gamma}_{3(2)} \\ \hat{\gamma}_{2(2)} \\ \hat{\gamma}_{13(22)} \\ \hat{\gamma}_{12(22)} \\ \hat{\gamma}_{23(22)} \end{bmatrix}.$$

Because there are 8 cells and $8 - 1$ terms in the model, there is a very simple form to the matrix necessary for obtaining asymptotic variances:

$$X\left(X'D(\hat{m})X\right)^{-1}X' =$$

$$D^{-1}(\hat{m}) - (5.52816)D^{-1}(\hat{m}) \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} [1,-1,-1,1,-1,1,1,-1]\, D^{-1}(\hat{m}). \quad (1)$$

The equality can be verified by direct computation of both sides. This approach requires a good matrix manipulation computer package for computing the left-hand side. The equality can also be verified by hand using orthogonality, projection operators, and the fact that $\text{rank}(X) = 7 = q - 1$. This approach requires facility with vector space concepts, cf. Christensen (1996b, p. 276).

In Example 3.2.4, for both collisions and rollovers, we were interested in the ratio of the odds of a nonsevere injury when the driver was not ejected relative to the odds of a nonsevere injury when the driver was ejected. We proceed to find a 95% confidence interval for

$$\log(m_{11k}m_{22k}/m_{12k}m_{21k})\,.$$

Recall that, based on the model of no three-factor interaction, this log odds ratio does not depend on k. The estimate of the log odds ratio is

$$.77 = \log(2.16)\,.$$

This can be arrived at in either of two ways. For $k = 1$, define the vector $\rho_1' = (1, -1, 0, 0, -1, 1, 0, 0)$ so that

$$\begin{aligned}\rho_1'\hat{\mu} &= \rho_1'\log(\hat{m}) \\ &= \log(\hat{m}_{111}\hat{m}_{221}/\hat{m}_{121}\hat{m}_{211}) \\ &= \log\left[(350.49)(23.49)/(149.51)(25.51)\right] \\ &= \log(2.16)\,.\end{aligned}$$

Otherwise, for $k = 2$, let $\rho_2' = (0, 0, 1, -1, 0, 0, -1, 1)$ so that

$$\begin{aligned}\rho_2'\hat{\mu} &= \log(\hat{m}_{112}\hat{m}_{222}/\hat{m}_{122}\hat{m}_{212}) \\ &= \log\left[(59.51)(79.51)/(112.49)(19.49)\right] \\ &= \log(2.16)\,.\end{aligned}$$

The estimated variance is

$$\rho_j' X \left(X' D(\hat{m}) X\right)^{-1} X' \rho_j = .045.$$

This can be computed directly using matrix manipulations, or it can be computed by hand using equation (1), or it can be computed from the reported standard error of $\hat{\gamma}_{12(22)}$ using the auxiliary regression (which will be reviewed in the next example). The 95% confidence interval for the log odds ratio with k fixed has the end points

$$.77 \pm 1.96\sqrt{.045}$$

and is the interval $(.35, 1.19)$. If we exponentiate the end points, we get a 95% confidence interval for the odds ratio of

$$(1.4, 3.3)\,.$$

Thus, the evidence indicates that the odds of a nonsevere injury when the driver is not ejected are between, roughly, one and a half to three times the odds of a nonsevere injury when the driver is ejected.

EXAMPLE 10.2.5. Consider a $2 \times 3 \times 2$ table and the model

$$\mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} .$$

The design matrix X for this model is given in Example 10.0.2. Suppose we are interested in the log odds

$$\log(m_{1jk}/m_{2jk}) = \mu_{1jk} - \mu_{2jk} = u_{1(1)} - u_{1(2)} .$$

Note that in this model, the odds are the same for any values of j and k. Using the notation in Example 10.0.2, write $\rho' = (1, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0)$ so that

$$\rho'\mu = \mu_{111} - \mu_{211} = u_{1(1)} - u_{1(2)} .$$

The estimate is

$$\rho'\hat{\mu} = \log(\hat{m}_{111}) - \log(\hat{m}_{211}) = \log(\hat{m}_{111}/\hat{m}_{211}),$$

but this estimate does not depend on the last two subscripts. For any j and k,

$$\rho'\hat{\mu} = \log(\hat{m}_{1jk}/\hat{m}_{2jk}) .$$

The difficult part of the analysis is in finding the variance. The variance is most easily computed by setting the problem up as a regression analysis. Write

$$\begin{bmatrix} \mu_{111} \\ \mu_{112} \\ \mu_{121} \\ \mu_{122} \\ \mu_{131} \\ \mu_{132} \\ \mu_{211} \\ \mu_{212} \\ \mu_{221} \\ \mu_{222} \\ \mu_{231} \\ \mu_{232} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_{1(2)} \\ \gamma_{2(2)} \\ \gamma_{2(3)} \\ \gamma_{3(2)} \\ \gamma_{23(22)} \\ \gamma_{23(32)} \end{bmatrix}$$

$$\mu \quad = \quad W \quad \gamma .$$

The new design matrix was arrived at by eliminating every column of the old design matrix that corresponded to a u term involving $i = 1$, $j = 1$, or $k = 1$. The estimates of the γ's are the estimates of the u's subject to the

side conditions $0 = u_{1(1)} = u_{2(1)} = u_{3(1)} = u_{23(1k)} = u_{23(j1)}$ for all j and k, and

$$\rho'\mu = \mu_{111} - \mu_{211} = \begin{cases} \rho' W\gamma = -\gamma_{1(2)} \\ \rho' X\beta = u_{1(1)} - u_{1(2)} \,. \end{cases}$$

Let

$$\begin{aligned} K &= \rho'[\hat{A} - A_z]D^{-1}(\hat{m})\rho \\ &= \rho' X(X'D(\hat{m})X)^{-1}X'\rho \\ &= \rho' W(W'D(\hat{m})W)^{-1}W'\rho \end{aligned}$$

so that, asymptotically,

$$\mathrm{Var}(\rho'\hat{\mu}) = K \,.$$

The value K is easily obtained by performing an auxiliary regression, as discussed in Section 6.7. In particular, fitting

$$Y = W\gamma + e$$

with weights $\hat{m}_i$ and dependent variable

$$y_i = \log(\hat{m}_i) + (n_i - \hat{m}_i)/\hat{m}_i,$$

the regression program *will report*

$$\mathrm{SE}(\hat{\gamma}_{1(2)}) = \sqrt{\mathrm{MSE}}\, K \,.$$

Dividing by $\sqrt{\mathrm{MSE}}$ gives the correct asymptotic standard error.

Almost any good regression program allows the user to print out the matrix

$$\mathrm{Cov}(\hat{\gamma})/\mathrm{MSE} = (W'D(\hat{m})W)^{-1} \,.$$

This is the key to obtaining asymptotic variances for log-linear models. Consider the log odds ratio

$$\begin{aligned} \log(m_{i21}m_{i32}/m_{i22}m_{i31}) &= \mu_{i21} - \mu_{i22} - \mu_{i31} + \mu_{i32} \\ &= u_{23(21)} - u_{23(22)} - u_{23(31)} + u_{23(32)} \,. \end{aligned}$$

This log odds ratio does not depend on the value of i. Picking $i = 1$ for convenience, let

$$\rho' = (0, 0, 1, -1, -1, 1, 0, 0, 0, 0, 0, 0),$$

so

$$\rho'\mu = \rho' X\beta = u_{23(21)} - u_{23(22)} - u_{23(31)} + u_{23(32)} \,.$$

In the $\mu = W\gamma$ parametrization, this becomes

$$\rho'\mu = \rho' W\gamma = \gamma_{23(32)} - \gamma_{23(22)} \,.$$

There are two ways to arrive at this result. First, one can substitute the appropriate functions of the γ's in place of the μ_{ijk}'s. This leads to

$$\begin{aligned}\rho'\mu &= \mu_{121} - \mu_{122} - \mu_{131} + \mu_{132} \\ &= [\gamma_0 + \gamma_{2(2)}] - [\gamma_0 + \gamma_{2(2)} + \gamma_{3(2)} + \gamma_{23(22)}] \\ &\quad - [\gamma_0 + \gamma_{2(3)}] + [\gamma_0 + \gamma_{2(3)} + \gamma_{3(2)} + \gamma_{23(32)}] \\ &= \gamma_{23(32)} - \gamma_{23(22)}\,.\end{aligned}$$

Second, one can notice that

$$\rho'W = (0, 0, 0, 0, 0, -1, 1),$$

so that

$$\rho'\mu = \rho'W\gamma = \gamma_{23(32)} - \gamma_{23(22)}\,.$$

If we write

$$\lambda' = \rho'W,$$

then the estimated large sample variance is

$$\begin{aligned}\rho'X(X'D(\hat{m})X)^{-1}X'\rho &= \rho'W(W'D(\hat{m})W)^{-1}W'\rho \\ &= \lambda'(W'D(\hat{m})W)^{-1}\lambda\end{aligned}$$

which is easily computed if the regression program provides $(W'D(\hat{m})W)^{-1}$. ∎

Variances for other estimated log odds ratios are computed in a similar manner. Because of the model, any log odds ratio with either j or k fixed, e.g., $\log(m_{1j1}m_{2j2}/m_{1j2}m_{2j1})$, is zero by assumption. Estimates of log odds in the j or k indices, e.g., $\log(m_{ij1}/m_{ij2})$, can also be estimated and large sample variances computed. However, because of the existence of the u_{23} interaction, the log odds will depend on the value of j. These issues are considered in more detail in the next example.

EXAMPLE 10.2.6. Consider again the data on classroom behavior used in Examples 3.0.1 and 3.2.2. The data and estimated expected cell counts for the model in which behavior is independent of risk and adversity are given below.

n_{ijk} ($\hat{m}_{ijk}$)		Adversity (k) Low		Medium		High	
Risk (j)		N	R	N	R	N	R
Classroom Behavior (i)	Non.	16 (14.02)	7 (6.60)	15 (14.85)	34 (34.64)	5 (4.95)	3 (4.95)
	Dev.	1 (2.98)	1 (1.40)	3 (3.15)	8 (7.36)	1 (1.05)	3 (1.05)

The ANOVA type model is

$$\mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}\,.$$

Except for the fact that this is a $2 \times 2 \times 3$ table instead of a $2 \times 3 \times 2$ table, the model is exactly as in the previous example.

Dropping all u terms with i, j, or k equal to 1 leads to

$$\begin{bmatrix} \hat{\mu}_{111} \\ \hat{\mu}_{121} \\ \hat{\mu}_{112} \\ \hat{\mu}_{122} \\ \hat{\mu}_{113} \\ \hat{\mu}_{123} \\ \hat{\mu}_{211} \\ \hat{\mu}_{221} \\ \hat{\mu}_{212} \\ \hat{\mu}_{222} \\ \hat{\mu}_{213} \\ \hat{\mu}_{223} \end{bmatrix} = \begin{bmatrix} \log(14.02) \\ \log(6.60) \\ \log(14.85) \\ \log(34.64) \\ \log(4.95) \\ \log(4.95) \\ \log(2.98) \\ \log(1.40) \\ \log(3.15) \\ \log(7.36) \\ \log(1.05) \\ \log(1.05) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma} \\ \hat{\gamma}_{1(2)} \\ \hat{\gamma}_{2(2)} \\ \hat{\gamma}_{3(2)} \\ \hat{\gamma}_{3(3)} \\ \hat{\gamma}_{23(22)} \\ \hat{\gamma}_{23(23)} \end{bmatrix}$$

or, alternatively,

$$\hat{\mu} = W\hat{\gamma}\,.$$

In particular, any weighted or unweighted regression analysis provides

$$\begin{aligned} \hat{\gamma} &= 2.640\,, \\ \hat{\gamma}_{1(2)} &= -1.548\,, \\ \hat{\gamma}_{2(2)} &= -0.754\,, \\ \hat{\gamma}_{3(2)} &= 0.057\,, \\ \hat{\gamma}_{3(3)} &= -1.041\,, \\ \hat{\gamma}_{23(22)} &= 1.601\,, \\ \hat{\gamma}_{23(23)} &= 0.754\,. \end{aligned}$$

(Actually, these are based on more significant digits for the $\hat{m}_{ijk}$'s than were reported above.) The matrix of asymptotic variances and covariances for the $\hat{\gamma}$'s is obtained from doing the appropriate auxiliary regression. It is

	$\hat{\gamma}$	$\hat{\gamma}_{1(2)}$	$\hat{\gamma}_{2(2)}$	$\hat{\gamma}_{3(2)}$	$\hat{\gamma}_{3(3)}$	$\hat{\gamma}_{23(22)}$	$\hat{\gamma}_{23(23)}$
$\hat{\gamma}$	.0610	−.0125	−.0588	−.0588	−.0588	−.0588	−.0588
$\hat{\gamma}_{1(2)}$	−.0125	.0713	.0000	.0000	.0000	.0000	.0000
$\hat{\gamma}_{2(2)}$	−.0588	.0000	.1838	.0588	.0588	−.1838	−.1838
$\hat{\gamma}_{3(2)}$	−.0588	.0000	.0588	.1144	.0588	−.1144	−.0588
$\hat{\gamma}_{3(3)}$	−.0588	.0000	.0588	.0588	.2255	−.0588	−.2255
$\hat{\gamma}_{23(22)}$	−.0588	.0000	−.1838	−.1144	−.0588	.2632	.1838
$\hat{\gamma}_{23(23)}$	−.0588	.0000	−.1838	−.0588	−.2255	.1838	.5172

This matrix is the basis for all subsequent variance estimates.

The estimate of the log odds of nondeviant behavior is

$$\begin{aligned}\log(\hat{m}_{1jk}/\hat{m}_{2jk}) &= \hat{\mu}_{1jk} - \hat{\mu}_{2jk} \\ &= \hat{u}_{1(1)} - \hat{u}_{2(1)} \\ &= -\hat{\gamma}_{1(2)} \\ &= 1.548\,.\end{aligned}$$

The equivalence between the u parametrization and the γ parametrization is easily obtained by inspection of $W\gamma$. The asymptotic standard error is $\sqrt{.0713}$, so a 90% confidence interval has end points

$$1.548 \pm 1.645\sqrt{.0713}\,.$$

The odds of having a home situation that is not at risk depend on the adversity level. The log odds satisfy

$$\begin{aligned}\log(\hat{m}_{i1k}/\hat{m}_{i2k}) &= \hat{u}_{2(1)} + \hat{u}_{23(1k)} - \hat{u}_{2(2)} - \hat{u}_{23(2k)} \\ &= \begin{cases} -\hat{\gamma}_{2(2)}\,, & k=1 \\ -\hat{\gamma}_{2(2)} - \hat{\gamma}_{23(22)}\,, & k=2 \\ -\hat{\gamma}_{2(2)} - \hat{\gamma}_{23(23)}\,, & k=3. \end{cases}\end{aligned}$$

The estimated value when $k = 2$ is

$$.754 - 1.601 = -.847\,.$$

With

$$\mathrm{Var}(-\hat{\gamma}_{2(2)} - \hat{\gamma}_{23(22)}) = \mathrm{Var}(\hat{\gamma}_{2(2)}) + 2\mathrm{Cov}(\hat{\gamma}_{2(2)}, \hat{\gamma}_{23(22)}) + \mathrm{Var}(\hat{\gamma}_{23(22)}),$$

the asymptotic estimated variance is

$$.1838 - 2(.1838) + .2632 = .0794\,.$$

The interesting odds ratios involve the changes in the odds as k changes.

$$\begin{aligned}\log(m_{i11}m_{i22}/m_{i21}m_{i12}) &= u_{23(11)} - u_{23(21)} - u_{23(12)} + u_{23(22)} \\ &= \gamma_{23(22)}\,, \\ \log(m_{i11}m_{i23}/m_{i21}m_{i13}) &= \gamma_{23(23)}\,, \\ \log(m_{i12}m_{i23}/m_{i22}m_{i13}) &= \gamma_{23(23)} - \gamma_{23(22)}\,.\end{aligned}$$

The last of these has an estimate of

$$.754 - 1.601 = -.847$$

and an estimated asymptotic variance of

$$.5172 - 2(.1838) + .2632 = .4128\,.$$

A 95% confidence interval for the log odds ratio is

$$(-.2.106, .412).$$

Transforming to the original scale gives an interval for the odds ratio of

$$(.12, 1.5).$$

The odds of being not at risk for medium-adversity schools is between .12 and 1.5 times those for high-adversity schools. The large interval is related to the very small numbers available at high risk. The result does not depend on the classroom behavior.

As a matter of fact, the need for including the u_{23} terms in the model is driven by the fact that

$$\hat{\gamma}_{23(22)} = 1.601$$

with an asymptotic standard error of

$$\sqrt{.2632} = .513\,.$$

Thus, there is clear evidence that the odds of being not at risk are higher for low-adversity schools than for high-adversity schools. In fact, the odds are roughly between 2 and 13 times larger with 95% confidence.

As we have seen, there is a problem with the output from standard regression software. Using a regression parametrization

$$\mu = W\gamma,$$

the key matrix to be obtained is

$$\hat{A}D^{-1}(\hat{m}) = W(W'D(\hat{m})W)^{-1}W'$$

which does not depend on the choice of W. Unfortunately, most regression software does not report $\hat{A}D^{-1}(\hat{m})$; it only reports

$$(W'D(\hat{m})W)^{-1}\,.$$

(The fact that there are good reasons for doing this makes it no less unfortunate for our purposes.) If the software allows computation of $\hat{A}D^{-1}(\hat{m})$, then the simple structure of the ρ vectors allows simple computation of the estimated variance $\rho'W(W'D(\hat{m})W)^{-1}W'\rho$. If the software does not allow direct computation of $\hat{A}D^{-1}(\hat{m})$, then it is necessary to compute the vector $\lambda' = \rho'W$. In other words, the simple function $\rho'\mu$ must be reparametrized into $\lambda'\gamma$.

Given λ' and $(W'D(\hat{m})W)^{-1}$, the variance $\lambda'(W'D(\hat{m})W)^{-1}\lambda$ is easily computed. The problem is in identifying λ, i.e., identifying the function of γ that is equivalent to $\rho'\mu$. Even though the interesting functions $\rho'\mu$ are simple, the functions of γ get progressively more complex as the model, (i.e., the matrix W) gets more complex. For example, the asymptotic variance of a log odds in terms of model parameters gets progressively more complicated as the model involves more higher-order interactions, even though the log odds is an extremely simple function of μ. With a good matrix manipulation package, keeping track of the parameters can be accomplished numerically.

Asymptotic Variances for Saturated Models

In Chapter 2, an asymptotic standard error was presented for estimated log odds ratios. The standard error is a consequence of applying Theorem 10.2.1a to a saturated model. Generally, standard errors for contrasts in the μ_i's are easily obtained for saturated models. Recall that for a saturated model, $\hat{\mu} = \log(n)$. Applying Theorem 10.2.1a, $\log(n) - \mu$ is approximately $N(0, [A - A_z]D^{-1}(m))$. We wish to characterize $[A - A_z]D^{-1}(m)$. The model is saturated, i.e.,

$$A = I(IDI)^{-1}ID = I,$$

so $AD^{-1}(m) = D^{-1}(m)$. For multinomial sampling (regardless of the log-linear model),

$$A_z D^{-1}(m) = \frac{1}{N} JJ'.$$

Thus, for a saturated model with a large multinomial sample, we have the approximation

$$\log(n) - \mu \sim N\left(0, D^{-1}(m) - \frac{1}{N} JJ'\right).$$

Let $\rho = (\rho_1, \ldots, \rho_q)'$ be a vector with $\rho'J = 0$, i.e., $\rho_\cdot = 0$, so $\rho'\mu$ is a contrast in the μ_i's. The large sample distribution of $\rho' \log(n)$ is

$$\rho' \log(n) - \rho'\mu \sim N(0, \rho'[D^{-1}(m) - (1/n_\cdot)JJ']\rho).$$

With $\rho'J = 0$, we have

$$\rho' \log(n) - \rho'\mu \sim N(0, \rho' D^{-1}(m)\rho)$$

or, equivalently,

$$\frac{\rho' \log(n) - \rho'\mu}{\sqrt{\rho' D^{-1}(m)\rho}} \sim N(0, 1).$$

For this distribution to be useful in drawing inferences about $\rho'\mu$, an estimate of the unknown standard deviation $\sqrt{\rho' D^{-1}(m)\rho}$ must be incorporated. By Theorem 10.2.1d, the vector n/N converges to the vector p, so $\rho' D^{-1}(m)\rho/\rho' D^{-1}(n)\rho = \rho' D^{-1}(p)p/\rho' D^{-1}(n/N)\rho$ converges to 1 and $\sqrt{\rho' D^{-1}(m)\rho}/\sqrt{\rho' D^{-1}(n)\rho}$ converges to 1. Hence, for large samples,

$$\frac{\rho'\log(n) - \rho'\mu}{\sqrt{\rho' D^{-1}(n)\rho}} = \frac{\rho'\log(n) - \rho'\mu}{\sqrt{\rho' D^{-1}(m)\rho}} \frac{\sqrt{\rho' D^{-1}(m)\rho}}{\sqrt{\rho' D^{-1}(n)\rho}} \sim N(0,1)\,.$$

This result can be very useful, especially for examining odds ratios.

EXAMPLE 10.2.7. For the $2\times2\times2$ table of Example 3.2.4 concerning auto injuries, we were interested in whether the odds ratios $p_{111}p_{221}/p_{121}p_{211}$ and $p_{112}p_{222}/p_{122}p_{212}$ were equal. Because

$$\frac{p_{11k}p_{22k}}{p_{12k}p_{21k}} = \frac{m_{11k}m_{22k}}{m_{12k}m_{21k}},$$

the log odds ratios are

$$\log\left(\frac{m_{11k}m_{22k}}{m_{12k}m_{21k}}\right) = \mu_{11k} - \mu_{12k} - \mu_{21k} + \mu_{22k}\,.$$

The odds ratios are equal if and only if the contrast in the μ_{ijk}'s

$$(\mu_{111} - \mu_{121} - \mu_{211} + \mu_{221}) - (\mu_{112} - \mu_{122} - \mu_{212} + \mu_{222})$$

equals zero. [Note that if $\mu = (\mu_{111}, \mu_{112}, \mu_{121}, \mu_{122}, \mu_{211}, \mu_{212}, \mu_{221}, \mu_{222})'$, then $\rho' = (1, -1, -1, 1, -1, 1, 1, -1)$.] The estimated odds ratios were

$$\begin{aligned}\hat{p}_{111}\hat{p}_{221}/\hat{p}_{121}\hat{p}_{211} &= 350(23)/26(150)\\ &= 2.064\end{aligned}$$

and

$$\begin{aligned}\hat{p}_{112}\hat{p}_{222}/\hat{p}_{122}\hat{p}_{212} &= 60(80)/19(112)\\ &= 2.256\,.\end{aligned}$$

The estimate of the contrast is

$$\log(2.064) - \log(2.256) = -0.089\,.$$

The standard error for the estimate is

$$\begin{aligned}\sqrt{\rho' D^{-1}(n)\rho} &= \sqrt{\frac{1}{350} + \frac{1}{23} + \frac{1}{26} + \frac{1}{150} + \frac{1}{60} + \frac{1}{80} + \frac{1}{19} + \frac{1}{112}}\\ &= .4268\,.\end{aligned}$$

We can now test the hypothesis that the contrast is zero. The test statistic is $-.089/.4268 = -.21$. For an α level two-sided test, $|-.21|$ is compared to $z(1-\alpha/2)$. The hypothesis that the contrasts are equal is not rejected for any reasonable size of α. A 95% confidence interval for the contrast has limits

$$-.089 \pm (1.96)(.4268)\,.$$

The test based on the asymptotic standard error is an alternative to the likelihood ratio and Pearson chi-squared tests for no three-factor interaction.

To examine an individual cell, the term $A_z D^{-1}(m)$ must be accounted for in the covariance matrix. It is easily seen that for large samples, the appropriate distribution for $\hat{p}_{ijk}$ is

$$\frac{\hat{p}_{ijk} - p_{ijk}}{\sqrt{\hat{p}_{ijk}(1-\hat{p}_{ijk})/N}} \sim N(0,1)\,.$$

Similar results hold for $\hat{m}_{ijk}$ and $\hat{\mu}_{ijk}$.

Testing Models

Consider the problem of testing a model $\mu = X_0\beta_0$ against a larger model $\mu = X\beta$. In particular, assume that $\mu = X\beta$ is valid and examine the test of

$$H_0 : \mu = X_0\beta_0 \quad \text{for some} \quad \beta_0$$

versus

$$H_A : \mu \neq X_0\beta_0 \quad \text{for any} \quad \beta_0\,,$$

where $C(X_0) \subset C(X)$, i.e., $X_0 = XB$ for some matrix B. Let $\hat{m}$ be the MLE of m under the assumption that $\mu = X\beta$ and let $\hat{m}_0$ be the MLE of m under the assumption that $\mu = X_0\beta_0$. The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{0i})\,.$$

The Pearson test statistic is

$$X^2 = \sum_{i=1}^{q} (\hat{m}_i - \hat{m}_{0i})^2/\hat{m}_{0i}\,.$$

The main asymptotic results for testing hypotheses are given in the following theorem.

Theorem 10.2.8. Let $r = \text{rank}(X)$ and $r_0 = \text{rank}(X_0)$.

(a) If H_0 is true and $N \equiv n_{\cdot}$ is large, the following distributions are approximately valid:

$$G^2 \sim \chi^2(r - r_0)$$

and

$$X^2 \sim \chi^2(r - r_0)\,.$$

Moreover,

$$G^2 - X^2 \xrightarrow{P} 0\,.$$

(b) If H_0 is not true, then both G^2 and X^2 get arbitrarily large as the sample size increases.

It is interesting to note that the degrees of freedom for the test are $\text{rank}(X) - \text{rank}(X_0)$. This is the reason that degrees of freedom are computed exactly as in analysis of variance. In both cases, it is simply the linear structure of the model that determines the degrees of freedom.

10.3 Product-Multinomial Sampling

With a few minor changes, all of the results of Sections 1 and 2 hold for product-multinomial sampling. Suppose that we have t multinomial populations instead of just one. We can write the observations as n_{ij}, $i = 1, \ldots, t$, $j = 1, \ldots, s_i$, where s_i is the number of categories in the ith multinomial. (Note that $q = \sum_{i=1}^{t} s_i$.) The probabilities and expected cell counts can be written similarly as p_{ij} and m_{ij}, respectively.

In place of the condition from multinomial sampling that all the probabilities in the table add to 1, cf. equation (10.1.2), we now have

$$p_{i\cdot} = 1, \qquad i = 1, \ldots, t,$$

and because $m_{ij} = n_{i\cdot} p_{ij}$, we have

$$m_{i\cdot} = n_{i\cdot}\,, \qquad i = 1, \ldots, t\,.$$

Write the vectors $n = (n_{11}, n_{12}, \ldots, n_{ts_t})'$ and $m = (m_{11}, m_{12}, \ldots, m_{ts_t})'$. Let Z be a $q \times t$ matrix of indicator variables for the t samples. Specifically, each column of Z corresponds to a different multinomial. A particular column of Z, say the ith column, has ones in the rows corresponding to $n_{i1}, \ldots, n_{is_i}$ and zeros in all other rows. (Note that if $n_{ij} = \mu_i + e_{ij}$ was a one-way ANOVA, Z would be the design matrix for the linear model.) With this definition of Z, the condition $m_{i\cdot} = n_{i\cdot}$, $i = 1, \ldots, t$, becomes

$$n'Z = m'Z\,.$$

Suppose now that we have the log-linear model $\log(m) = \mu = X\beta$. It can be shown that maximizing the log-likelihood under product multinomial sampling is equivalent to maximizing $\ell(m) = n' \log(m)$, cf. Chapter 12.

The MLE of m must maximize $\ell(m)$ subject to the conditions

$$\log(m) = X\beta \tag{1}$$

and

$$n'Z = m'Z\,. \tag{2}$$

Just as in Section 1, if $\hat{m}$ maximizes $\ell(m)$ subject only to condition (1), then $\hat{m}$ must satisfy

$$n'X = \hat{m}'X\,. \tag{3}$$

In order to get condition (2) satisfied, we restrict our attention to models $\log(m) = X\beta$ in which $C(Z) \subset C(X)$. For such models, $Z = XB$ for some matrix B; hence, (3) implies that

$$n'Z = n'XB = \hat{m}'XB = \hat{m}'Z\,.$$

The assumption that $C(Z) \subset C(X)$ is not difficult to deal with. For an $I \times J$ table in which rows are independent multinomial samples, the condition $C(Z) \subset C(X)$ is the requirement that every log-linear model include (the equivalent of) $u_{1(i)}$ terms for rows. In an $I \times J \times K$ table in which there is an independent multinomial sample for each combination of row and layer, the condition $C(Z) \subset C(X)$ is the requirement that every log-linear model include $u_{13(ik)}$ terms or their equivalent. Note that, for example, the models

$$\log(m_{ijk}) = u_{13(ik)} + u_{123(ijk)}$$

and

$$\log(m_{ijk}) = u_{123(ijk)}$$

are equivalent models, so in spite of the fact that $\log(m_{ijk}) = u_{123(ijk)}$ does not contain $u_{13(ik)}$ terms, it does contain the equivalent of $u_{13(ik)}$ terms.

Under product-multinomial sampling, the asymptotic results of Section 2 change very little. The matrix $D(p)$ is no longer of interest. Instead, define $m^* = (m^*_{11}, \ldots, m^*_{ts_t})$ where $m^*_{ij} = n_{i\cdot}p_{ij}/n_{\cdot\cdot}$. Redefine

$$D = D(m^*)\,.$$

The matrix A is defined as before except that the new version of D is used. Also, redefine A_z as

$$A_z = Z(Z'DZ)^{-1}Z'D\,.$$

For asymptotic results, *let $N = n_{\cdot\cdot}$ get large and let $n_{i\cdot}/n_{\cdot\cdot}$ remain fixed for each i.* Write $N_i = n_{i\cdot} = m_{i\cdot}$.

Before restating the asymptotic results, note that multinomial sampling is just a special case of product-multinomial sampling. In particular, it has $t = 1$, $Z = J$ (J is a $q \times 1$ vector of 1s), $N = n_{\cdot} = n_{1\cdot} = n_{\cdot\cdot}$, and $m^* = p$.

Theorem 10.3.1. For multinomial or product-multinomial sampling,

the following distributions are approximately valid when $N_1, \ldots, N_t$ are large:

(a) $\hat{\mu} - \mu \sim N(0, (A - A_z)D^{-1}(m))$,

(b) $\hat{m} - m \sim N(0, D(m)(A - A_z))$.

In addition,

(c) $\hat{\mu} - \mu \xrightarrow{P} 0$,

(d) $N^{-1}\hat{m} \xrightarrow{P} m^*$.

Estimation for product-multinomial sampling is similar to that for multinomial sampling. Theorem 10.3.1 looks identical to Theorem 10.2.1. The difference is that A_z stands for something different. Again, the only problem is in computing asymptotic variances. Under product-multinomial sampling, the variance of $\rho'\hat{\mu}$ is $\rho'[A - A_z]D^{-1}(m)\rho$. Note that

$$\begin{aligned} Z'D(m)Z &= D(N_1, \cdots, N_t)\,, \\ (Z'D(m)Z)^{-1} &= D\left(\frac{1}{N_1}, \cdots, \frac{1}{N_t}\right), \end{aligned}$$

and

$$A_z D^{-1}(m) = ZD\left(\frac{1}{N_1}, \cdots, \frac{1}{N_t}\right) Z'\,.$$

The variance of $\rho'\hat{\mu}$ is

$$\rho' A D^{-1}(m)\rho - \rho' Z D\left(\frac{1}{N_1}, \cdots, \frac{1}{N_t}\right) Z'\rho\,.$$

The second term can be computed exactly. The first term must be estimated and, even then, requires a computer to evaluate. For example, taking $\rho' = e'_{ij} = (0, \cdots, 0, 1, 0, \cdots, 0)$ with the 1 in the column corresponding to the ij cell, Theorem 10.3.1 yields

$$\hat{\mu}_{ij} - \mu_{ij} = e'_{ij}(\hat{\mu} - \mu) \sim N\left(0, \frac{a_{ij,ij}}{m_{ij}} - \frac{1}{N_i}\right)$$

where $a_{ij,ij}$ is the diagonal element of A corresponding to the ij cell.

Similarly,

$$\mathrm{Var}(\hat{m}_{ij} - m_{ij}) = m_{ij}a_{ij,ij} - m_{ij}^2/N_i$$

and with $p_{ij} = m_{ij}/N_i$,

$$\begin{aligned} \mathrm{Var}(\hat{p}_{ij} - p_{ij}) &= p_{ij}a_{ij,ij}/N_i - p_{ij}^2/N_i \\ &= p_{ij}(a_{ij,ij} - p_{ij})/N_i\,. \end{aligned}$$

If $\rho'\mu$ is a log odds or a log odds ratio that happens to be computed entirely within a particular multinomial, then $\rho'Z = 0$ and

$$\rho'[A - A_z]D^{-1}(m)\rho = \rho' AD^{-1}(m)\rho\,.$$

This is computed exactly as in Section 2. Unfortunately, it again requires a computer to evaluate.

For testing hypotheses, the large sample results appropriate for product-multinomial sampling are given in the following theorem.

Theorem 10.3.2. Assume $\mu = X\beta$ and let X_0 be a matrix with $C(Z) \subset C(X_0) \subset C(X)$. Let $\text{rank}(X) = r$ and $\text{rank}(X_0) = r_0$. For testing $H_0 : \mu = X_0\beta_0$ for some β_0 versus $H_A : \mu \neq X_0\beta_0$ for any β_0, under multinomial or product-multinomial sampling, if N is large, then the following approximate distributions hold:

(a) if H_0 is true, $G^2 \sim \chi^2(r - r_0)$,

(b) if H_0 is true, $X^2 \sim \chi^2(r - r_0)$,

also,

(c) if H_0 is true, $G^2 - X^2 \xrightarrow{P} 0$,

(d) if H_0 is false, G^2 and X^2 tend to infinity as N gets large.

Note that by (c), if H_0 is true, the difference between G^2 and X^2 can be used as an indication of how good the large sample approximation is. If H_0 is not true, then G^2 and X^2 need not be equivalent in large samples.

10.4 Inference for Model Parameters

Thus far, we have been primarily concerned with estimation of m and μ. It may be of interest to estimate the parameter vector β in the log-linear model $\mu = X\beta$. Estimates of β are obtained as in analysis of variance and regression, except that instead of performing operations on the data (y values), the operations are performed on $\hat{\mu}$.

Suppose that $\text{rank}(X) = p$ so that $\mu = X\beta$ is a regression model. $\hat{\beta}$ satisfies

$$\hat{\mu} = X\hat{\beta},$$

so

$$(X'X)^{-1}X'\hat{\mu} = (X'X)^{-1}X'X\hat{\beta} = \hat{\beta}\,.$$

The MLE of $\hat{\beta}$ is obtained by performing a regression on $\hat{\mu}$. (In fact, any weighted regression will give the same $\hat{\beta}$.)

Essentially the same argument holds for ANOVA type models. If one imposes side conditions on the parameters (something the author is loathe to do), then estimates of the parameters in ANOVA models are available. For example, in the model $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$ if the side conditions $u_{1(\cdot)} = u_{2(\cdot)} = u_{12(i\cdot)} = u_{12(\cdot j)} = 0$ are imposed and if we denote $w_{ij} = \hat{\mu}_{ij}$, then

$$\begin{aligned} \hat{u} &= \bar{w}_{\cdot\cdot}\,, \\ \hat{u}_{1(i)} &= \bar{w}_{i\cdot} - \bar{w}_{\cdot\cdot}\,, \\ \hat{u}_{2(j)} &= \bar{w}_{\cdot j} - \bar{w}_{\cdot\cdot}\,, \\ \hat{u}_{12(ij)} &= w_{ij} - \bar{w}_{i\cdot} + \bar{w}_{\cdot j} + \bar{w}_{\cdot\cdot}\,. \end{aligned}$$

Again, these are precisely the estimates obtained by doing an ANOVA on $\hat{\mu}$.

Tests and confidence intervals for functions $\rho' X\beta$ can be obtained from the asymptotic distribution

$$\frac{\rho'\hat{\mu} - \rho' X\beta}{\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}} \sim N(0,1)\,.$$

For example, an asymptotic 95% confidence interval for $\rho' X\beta$ has limits $\rho'\hat{\mu} \pm 1.96\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}$ and an $\alpha = .05$ test of $H_0 : \rho' X\beta = 0$ versus $H_A : \rho' X\beta \neq 0$ rejects if

$$\frac{\rho'\hat{\mu}}{\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}} > 1.96$$

or if

$$\frac{\rho'\hat{\mu}}{\sqrt{\rho'(A - A_z)D^{-1}(n)\rho}} < -1.96\,.$$

Example 10.4.1. In this and the previous three sections, a lot of machinery has been developed for analyzing log-linear models. In this example, we apply the matrix approach to the analysis of model (6.2.2) in Example 6.2.6. Our analysis also employs the data from Example 6.2.1 as summarized in Example 6.2.5.

In matrix form, model (6.2.2) can be written as

$$\begin{bmatrix} \log(m_{1111}) \\ \log(m_{1112}) \\ \log(m_{1121}) \\ \log(m_{1122}) \\ \log(m_{1211}) \\ \log(m_{1212}) \\ \log(m_{1221}) \\ \log(m_{1222}) \\ \log(m_{2111}) \\ \log(m_{2112}) \\ \log(m_{2121}) \\ \log(m_{2122}) \\ \log(m_{2221}) \\ \log(m_{2222}) \end{bmatrix} = X \begin{bmatrix} \lambda \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \\ \lambda_{23} \\ \lambda_{24} \\ \lambda_{34} \\ \lambda_{123} \\ \lambda_{124} \\ \lambda_{134} \\ \lambda_{234} \\ \lambda_{1234} \end{bmatrix}$$

where

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix}.$$

The columns of the design matrix X can be identified as $X_0, \ldots, X_{1234}$ with the subscript of X identical to the subscript of the corresponding λ term. (X_0 corresponds to λ.) Note that, say, X_{12} can be obtained by multiplying together the elements of X_1 and X_2. Similarly, the elements of X_{134} can be obtained by multiplying together the elements of X_1, X_3, and X_4. In fact, any column with more than one subscript can be obtained by multiplying together the appropriate columns with one subscript.

Another important fact is that any two columns, for example X_{12} and X_{134}, have the property that $X'_{12}X_{134} = 0$. Any column, say X_{12}, has $X'_{12}X_{12} = 16 = q$, so we have $\frac{1}{16}X'_{12}X\beta = \lambda_{12}$. The estimate of λ_{12} is $\frac{1}{16}X'_{12}X\hat{\beta} = (1/16)X'_{12}\hat{\mu} = (1/16)(\hat{\mu}_{11..} - \hat{\mu}_{12..} - \hat{\mu}_{21..} + \hat{\mu}_{22..}) = 4(\bar{w}_{11..} - \bar{w}_{12..} - \bar{w}_{21..} + \bar{w}_{22..})$ where $w_{hijk} = \log(n_{hijk})$ because the model is saturated.

The variance of $\frac{1}{16}X'_{12}\hat{\mu}$ is $\left(\frac{1}{16}\right)^2 X'_{12}[D^{-1}(m) - A_z D^{-1}(m)]X_{12}$ for large samples. If the parameter λ_{12} is not forced into the model to deal with product-multinomial sampling, then $X'_{12}A_z D^{-1}(m)X_{12} = 0$, so the asymp-

totic variance is

$$\left(\frac{1}{q}\right)^2 X_{12}' D^{-1}(m) X_{12} = \left(\frac{1}{q}\right)^2 \sum_{hijk} \left(\frac{1}{m_{hijk}}\right)$$

where $q = 16$. The estimated asymptotic variance of $\hat{\lambda}_{12}$ is

$$\widehat{\text{Var}}(\hat{\lambda}_{12}) = \left(\frac{1}{q}\right)^2 \sum_{hijk} \left(\frac{1}{n_{hijk}}\right) .$$

In fact, the same asymptotic variance applies to all of the λ terms that are not forced into the model.

Using the asymptotic distribution

$$\frac{\hat{\lambda}_{12} - \lambda_{12}}{\sqrt{\left(\frac{1}{q^2}\right) \sum_{hijk} \left(\frac{1}{n_{hijk}}\right)}} \sim N(0,1),$$

a test of $H_0 : \hat{\lambda}_{12} = 0$ is based on comparing the test statistic

$$\frac{\hat{\lambda}_{12} - 0}{\sqrt{\left(\frac{1}{q^2}\right) \sum_{hijk} \left(\frac{1}{n_{hijk}}\right)}}$$

to a $N(0,1)$ distribution. Using the numbers in Example 6.2.5, we see that

$$|\hat{\lambda}_{TW}| = 0.914/16 = .0571\,,$$

the standard error is

$$1.307/16 = .0817\,,$$

and the test statistic is

$$\frac{.0571 - 0}{.0817} = 0.70\,,$$

just as reported in Example 6.2.5. There is very little evidence that $\lambda_{TW} \neq 0$.

Similarly, an asymptotic 95% confidence interval for λ_{TW} has end points

$$.0571 \pm 1.96(.0817)\,.$$

10.5 Methods for Finding Maximum Likelihood Estimates

In general, some sort of iterative technique is necessary to find MLEs for log-linear models. The two commonly used methods are *iteratively reweighted least squares* and iterative proportional fitting. Iterative proportional fitting was discussed in Section 3.3. It works only for ANOVA type models. Fitting of general log-linear models is usually performed using iteratively reweighted least squares.

Iteratively Reweighted Least Squares

Maximum likelihood estimates for log-linear models can be found by performing a sequence of weighted linear regressions. This method is an application of the *Newton-Raphson algorithm*.

Given a vector function $f(\beta)$, Newton-Raphson is a method for finding a solution to $f(\beta) = 0$. It begins with an initial guess of β, say β_0. Newton-Raphson then defines a sequence of β's, say $\beta_1, \beta_2, \ldots$, that converge to a value $\hat{\beta}$ that satisfies $f(\hat{\beta}) = 0$. The sequence is defined recursively; we begin with an initial value β_0 and define β_{t+1} given the value of β_t. Specifically, let $df(\beta)$ be the matrix of partial derivatives of the vector-valued function $f(\beta)$. By Taylor's theorem, if β_t and β_{t+1} are close to each other and $\delta_t = \beta_{t+1} - \beta_t$, then the approximate equality

$$f(\beta_{t+1}) \doteq f(\beta_t) + [df(\beta_t)]\delta_t$$

holds. We are seeking a zero of $f(\beta)$ so Newton-Raphson sets

$$0 = f(\beta_t) + [df(\beta_t)]\delta_t$$

so that

$$\delta_t = -[df(\beta_t)]^{-1} f(\beta_t)\,.$$

With $\delta_t = \beta_{t+1} - \beta_t$, we have

$$\beta_{t+1} = \beta_t + \delta_t\,.$$

Consider a log-linear model $\mu = X\beta$ where X is a $q \times p$ matrix with $\text{rank}(X) = p$. Note that any log-linear model can be reparametrized so that $\text{rank}(X) = p$. The MLE of m will be the same regardless of the parametrization. We wish to find the maximum of the function $\ell(m)$. In particular, this can be done by setting appropriate partial derivatives of $\ell(m)$ equal to zero. The Newton-Raphson method can be used to find the zero of the partial derivative vector.

Before applying the Newton-Raphson method, we set some notation. If $x = (x_1, \ldots, x_q)'$, write $e^x = (e^{x_1}, \ldots, e^{x_q})'$. With $\log(m) = X\beta$, m is a function of β. Write $\log(m(\beta)) = X\beta$ and $m(\beta) = e^{X\beta}$. In applying Newton-Raphson, we find $\hat{\beta}$ with $f(\hat{\beta}) = 0$ where

$$f(\beta) = d\ell(e^{X\beta})$$

and $d\ell(e^{X\beta})$ is the matrix of partial derivatives of $\ell(e^{X\beta})$ with respect to the vector β. It follows that $\hat{m} = e^{X\hat{\beta}}$ will maximize $\ell(m)$ subject to the constraint that $\log(\hat{m}) = X\hat{\beta}$ for some $\hat{\beta}$.

It is shown in Chapter 12 that $f(\beta_t) = X'(n - m(\beta_t))$ and that $df(\beta_t) = -X'D(m(\beta_t))X$; thus,

$$\delta_t = [X'D(m(\beta_t))X]^{-1}X'(n - m(\beta_t))$$

and

$$\beta_{t+1} = \beta_t + [X'D(m(\beta_t))X]^{-1}X'(n - m(\beta_t))\,.$$

The value β_{t+1} can be obtained from β_t simply by doing a weighted regression analysis. Let

$$Y \equiv X\beta_t + [D(m(\beta_t))]^{-1}(n - m(\beta_t))\,. \tag{1}$$

If we fit the regression model

$$Y = X\beta + e\,,\ E(e) = 0\,,\ \mathrm{Cov}(e) = [D(m(\beta_t))]^{-1}\,,$$

the estimate of β is

$$\begin{aligned}\beta_{t+1} &= [X'D(m(\beta_t))X]^{-1}X'D(m(\beta_t))Y \\ &= \beta_t + \delta_t\,.\end{aligned}$$

The matrix $D(m(\beta_t))$ is diagonal, so this is the simplest form of weighted regression and can be performed on most standard regression programs. The weights are simply the individual values of the vector $m(\beta_t)$.

This method of finding MLEs, because it consists of a series of weighted regressions in which the weights continually change, is called *iteratively reweighted least squares*. The method does not depend on any particular choice of X except for the condition that $\mathrm{rank}(X) = p$. Any log-linear model can be reparametrized so that X has full column rank, i.e., $\mathrm{rank}(X) = p$, so the method is perfectly general.

10.6 Regression Analysis of Categorical Data

In this section, we present an alternative to maximum likelihood, namely the *weighted least squares* method of fitting log-linear models. This method was introduced by Grizzle, Starmer, and Koch (1969). It consists of fitting a linear model (regression model) to the logs of the counts while also using the counts as weights. We begin by explaining and illustrating the method. Mathematical justifications are given at the end of the section.

Recall that for a saturated model, $\hat{m} = n$ is the MLE. For large samples, Theorem 10.3.1 applies and, because $A = I$, we have the approximation

$$\log(n) \sim N(\mu, D^{-1}(m) - A_z D^{-1}(m))\,.$$

If we assume a log-linear model

$$\mu = X\beta\,,$$

then

$$\log(n) \sim N(X\beta, D^{-1}(m) - A_z D^{-1}(m))\,,$$

which can be rewritten as

$$\log(n) = X\beta + e, \qquad e \sim N(0, D^{-1}(m) - A_z D^{-1}(m)). \tag{1}$$

This is just a linear model, but it has an unusual covariance matrix for the errors. Most commonly in regression analysis, it is assumed that $\mathrm{Cov}(e) = \sigma^2 I$. Courses on applied regression analysis often deal with weighted least squares, where $\mathrm{Cov}(e) = \sigma^2 D(w)$ and w is some $q \times 1$ vector of known constants. This covariance structure can be handled very easily. In particular, most computer programs for doing regression analysis can handle this form of weighted regression. Unfortunately, the covariance matrix for model (1) is more complicated.

As will be discussed later in the subsection on Mathematical Justifications, estimates in model (1) are precisely the same as estimates in

$$\log(n) = X\beta + e, \qquad e \sim N(0, D^{-1}(m)). \tag{2}$$

This is much closer to the standard form of $\sigma^2 D(w)$. There are two differences. One is that in model (2) there is no variance σ^2 to be estimated; we know that $\sigma^2 = 1$. The second difference is that w is supposed to be known but m is not known. This problem is evaded by estimating m from the saturated model. Thus, the regression method is to fit the model

$$\log(n) = X\beta + e, \qquad e \sim N(0, D^{-1}(n)). \tag{3}$$

This procedure has essentially the same asymptotic properties as maximum likelihood estimation.

Although we are using model (3) as a device for fitting the log-linear model, our real model is model (1). Model (3) gives a valid estimate for β, but it cannot be used for the entire analysis. Fortunately, when considering the most interesting parameters in β, model (3) can be used to construct asymptotically valid tests and confidence intervals. In particular, this works for parameters that are not forced into the model to account for the sampling scheme. Remember, we assume that $C(Z) \subset C(X)$ where Z is the matrix of indicators for the product-multinomial samples, cf. Section 3. Any log-linear model can be reparametrized so that $X = [Z, X_1]$ and $\beta' = [\alpha', \beta_1']$. The parameter vector α consists of parameters that are forced into the model to account for the sampling scheme. For drawing inferences about β_1, model (3) gives valid tests and confidence intervals.

Because $\sigma^2 = 1$, when drawing inferences about model (3) one uses tests based on the normal distribution and the chi-square distribution rather than the t distribution and the F distribution. When performing chi-square tests, the test statistic is the numerator sum of squares from the usual F statistic with the appropriate number of degrees of freedom. Again, inferences must be restricted to parameters that are not forced into the model.

EXAMPLE 10.6.1. *Drug Comparisons.*
The hypothetical data presented below has been analyzed in Koch, Imrey,

Freeman, and Tolley (1976). They also mention other references. Three drugs A, B, and C were given to each of 46 subjects. The response of each subject to each drug was noted as favorable (F) or unfavorable (U). Assume a multinomial sampling scheme. The data are

	Drug B	F		U	
	Drug C	F	U	F	U
Drug A	F	6	16	2	4
	U	2	4	6	6

First, consider fitting the log-linear model [AB][C] by fitting the corresponding linear model

$$\begin{bmatrix} \log(6) \\ \log(16) \\ \log(2) \\ \log(4) \\ \log(2) \\ \log(4) \\ \log(6) \\ \log(6) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \beta \\ (\alpha\beta) \\ \gamma \end{bmatrix} + e \tag{4}$$

using the weights (6,16,2,4,2,4,6,6). The parameters α, β, $\alpha\beta$, and γ can be described as main effects for drugs A and B, the $A \times B$ interaction, and the drug C main effect. A regression program gave the following results for fitting this model:

Regression Output

	Coefficient	Std Error	t
μ	1.5662	.1335	11.73
α	.1436	.1300	1.11
β	.1436	.1300	1.11
$\alpha\beta$	.5128	.1294	3.96
γ	−.3055	.1201	−2.54

Sum of squared errors (SSE) = 1.7348
Degrees of freedom error (dfE) = 3
Mean squared error (MSE) = .5783

As discussed above, the regression program acts as if there is a scale parameter σ that needs to be estimated. For log-linear models, the scale parameter is one, so the regression output must be modified to remove the adjustments for scale. This consists of dividing the regression standard errors and multiplying the t values by $(\text{MSE})^{1/2}$. Doing this gives

	GSK Estimates		
Coefficient	Estimate	Std Error	z
μ	1.5662	—	—
α	.1436	.1710	.844
β	.1436	.1710	.844
$\alpha\beta$	.5128	.1702	3.011
γ	−.3055	.1579	−1.932

The z values can be compared to the standard normal distribution for an asymptotic test of whether the coefficients are zero. The fact that no standard error is reported for μ is due to the fact that μ is forced into the model by the multinomial sampling.

SSE is not used in the standard errors of coefficients, but it is used for testing different models. For example, fitting the model [A][B], i.e.,

$$\begin{bmatrix} \log(6) \\ \log(16) \\ \log(2) \\ \log(4) \\ \log(2) \\ \log(4) \\ \log(6) \\ \log(6) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \beta \end{bmatrix} + e \tag{5}$$

with the counts as weights, gives

$$\text{SSE[model (5)]} = 14.017$$

with 5 degrees of freedom. To test model (5) against model (4) [i.e., to test $H_0 : \alpha\beta = \gamma = 0$], compare the difference in the error sums of squares, $14.0173 - 1.7348 = 12.2825$, to a chi-square distribution with $5 - 3 = 2$ degrees of freedom. Of course, to test the significance of one parameter, either the chi square or the normal test can be used. The tests are identical.

Saturated models present some different features. Saturated models must fit perfectly; so for such a model, the SSE is zero. The fact that the saturated model has SSE = 0 causes a problem in finding standard errors and z values for regression coefficients. The regression output will try to use a scale parameter of zero, so the regression standard errors will all be reported as zero and the t values will be reported as infinite. It also follows that for any model other than the saturated model, the SSE reported in the regression output provides a direct test of lack of fit, i.e., a test of the model against the saturated model, when compared to a chi-square distribution with dfE degrees of freedom. For example, in the model [AB][C], comparing 1.7348 to a $\chi^2(2)$ provides a test for lack of fit.

Finally, many regression programs give additional output on the sums of squares for the different coefficients such as Sum of Squares explained by

each variable in the order they are entered into the model. For model (4), this is

Due to	*df*	SS
Regression	4	18.3901
α	1	4.4353
β	1	1.6723
$\alpha\beta$	1	8.5381
γ	1	3.7444

The test of $H_0 : \gamma = 0$ can be performed by comparing 3.7444 to a chi-squared distribution with 1 degree of freedom. The test of $H_0 : \gamma = \alpha\beta = 0$ can be performed by comparing $8.5381 + 3.7444 = 12.2825$ to a chi square with 2 degrees of freedom. Both of these tests are equivalent to previously discussed versions of the tests.

Three things should be noted about the estimation technique of Grizzle, Starmer, and Koch (GSK). First, the method consists of performing one step of the Newton-Raphson algorithm. If the initial guess in the Newton-Raphson equation (10.5.1) is taken as $X\beta_0 = \log(n)$, then one iteration gives the GSK estimate. Second, the GSK method of estimation depends on an asymptotic result. It is only asymptotically that model (1) is valid. Maximum likelihood, on the other hand, is a valid method of estimation for any sample size. Similarly, likelihood ratio statistics are reasonable statistics on which to base tests for any sample size. With maximum likelihood, all procedures will be based on sufficient statistics. Only the distributions depend on large samples. Finally, the GSK method has trouble with observations that are zero. Taking the log of zero is usually a problem.

Koch et al. (1976) propose a compromise between maximum likelihood and weighted least squares. Suppose we wish to fit some model that cannot be conveniently fitted by iterative proportional fitting, say

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + \gamma_i k\,. \tag{6}$$

If software is available for iterative proportional fitting but not for iteratively reweighted least squares, perform the maximum likelihood fit of a slightly larger ANOVA type model, say

$$\log(m_{ijk}) = u_{12(ij)} + u_{13(ik)}\,.$$

The estimated vector $\hat{m}$ obtained from this can be used in place of n in the GSK procedure. The analysis follows the standard GSK methods. The compromise essentially provides GSK with better starting values.

Mathematical Justifications

There are two things that require justification. First, that estimates are the same in models (1) and (2) and, second, that estimates of functions of β_1 have the same variance in models (1) and (2).

Why do models (1) and (2) have the same estimates? Note that

$$\begin{aligned} A_z D^{-1}(m) &= Z(Z'D(m)Z)^{-1}Z'D(m)D^{-1}(m) \\ &= Z(Z'D(m)Z)^{-1}Z'\,. \end{aligned}$$

Thus, the covariance matrix in model (1) is

$$D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z'\,.$$

The covariance matrix in model (2) can be written

$$D^{-1}(m) = [D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z'] + Z(Z'D(m)Z)^{-1}Z'\,.$$

Because $C(Z) \subset C(X)$, this is precisely the condition needed to apply Theorem 10.1.3 in Christensen (1996b). The theorem implies that best linear unbiased estimates in models (1) and (2) are identical.

The idea behind Christensen's theorem is that because $C(Z) \subset C(X)$, for some (non-negative definite) matrix B, the covariance matrix of (1) can be written

$$D^{-1}(m) - XBX'\,.$$

Model (2) is equivalent to model (1) but with an additional independent error term added in. In particular, model (2) is equivalent to

$$\begin{aligned} \log(n) &= X\beta + (e + e_0)\,, \\ e &\sim N(0, D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z')\,, \\ e_0 &\sim N(0, Z(Z'D(m)Z)^{-1}Z')\,, \end{aligned} \tag{7}$$

where e and e_0 are independent. The covariance matrix for the entire error is

$$\begin{aligned} \mathrm{Cov}(e + e_0) &= D^{-1}(m) - Z(Z'D(m)Z)^{-1}Z' + Z(Z'D(m)Z)^{-1}Z' \\ &= D^{-1}(m)\,. \end{aligned}$$

The trick involves the covariance matrix associated with e_0. With probability one, $e_0 \in C(Z(Z'D(m)Z)^{-1}Z') \subset C(Z) \subset C(X)$, cf. Christensen (1996b, Lemma 1.3.5). Because $e_0 \in C(X)$, we are adding error that we cannot distinguish from the mean $X\beta$. The only thing an unbiased estimate can do to such error is ignore it. Thus, the estimates with the additional error e_0 and the estimates without the additional error are identical.

We now examine the fact that estimates of estimable functions of β_1 have the same variance in models (1) and (2).

Estimates are the same in models (1) and (2), so using model (2) and standard linear model results, we have

$$\hat{\mu} = X\hat{\beta} = A\log(n)$$

where $A = X(X'D(m)X)^{-1}X'D(m)$. A is a projection operator onto $C(X)$; this means that $AX = X$, so, in particular, $AA = A$ and, because $C(Z) \subset C(X)$, $AZ = Z$ and $AA_z = A_z$.

Again, write $\mu = X\beta = Z\alpha + X_1\beta_1$. Consider an estimable function of β_1, say $\lambda'\beta_1$. For this to be estimable, by definition we must have $\lambda'\beta_1 = \rho'\mu = \rho'Z\alpha + \rho'X_1\beta_1$ for some $q \times 1$ vector ρ. In particular, we must have $\rho'Z = 0$ and $\rho'X_1 = \lambda'$. Now consider the asymptotic variance of $\lambda'\hat{\beta}_1$ under model (1),

$$\begin{aligned}
\mathrm{Var}(\lambda'\hat{\beta}_1) &= \mathrm{Var}(\rho'\hat{\mu}) \\
&= \mathrm{Var}(\rho'A\log(n)) \\
&= \rho'A[D^{-1}(m) - A_zD^{-1}(m)]A'\rho \\
&= \rho'AD^{-1}(m)A'\rho - \rho'AA_zD^{-1}(m)A'\rho \\
&= \rho'AD^{-1}(m)A'\rho\,.
\end{aligned}$$

The last equality follows from the fact that

$$\rho'AA_z = \rho'A_z = \rho'Z(Z'D(m)Z)^{-1}Z'D(m) = 0$$

because $\rho'Z = 0$.

The variance of $\lambda'\hat{\beta}_1$ under model (2) is

$$\begin{aligned}
\mathrm{Var}(\lambda'\hat{\beta}_1) &= \mathrm{Var}(\rho'A\log(n)) \\
&= \rho'A[D^{-1}(m)]A'\rho,
\end{aligned}$$

so the variances are the same under models (1) and (2). Thus, for estimable functions of β_1, the standard error reported from fitting model (2) is identical to the true standard error which is computed using model (1). For estimating functions that involve α, model (2) cannot be used to obtain standard errors.

In practice, neither model (1) nor (2) can be used because the covariance matrices involve the unknown parameter vector m. Model (3) substitutes the estimate n for m in the covariance matrix of (2). The same substitution in model (1) gives the most proper usable form for the GSK analysis. As above, the two models give the same estimate and the same standard errors for estimates of β_1. For drawing inferences about $\lambda'\beta_1$, use the approximate distribution

$$\frac{\lambda'\hat{\beta}_1 - \lambda'\beta_1}{\sqrt{\rho'D^{-1}(n)\rho}} \sim N(0,1)\,.$$

In particular, a large sample 95% confidence interval for $\lambda'\beta_1$ has end points

$$\lambda'\hat{\beta}_1 \pm 1.96\sqrt{\rho' D^{-1}(n)\rho}\,.$$

An α level test of $H_0 : \lambda'\hat{\beta}_1 = 0$ rejects H_0 when

$$\frac{|\lambda'\hat{\beta}_1|}{\sqrt{\rho' D^{-1}(n)\rho}} > z(1-\alpha/2)\,.$$

In summary, model (3) can be fitted very simply because it has a diagonal covariance matrix. Model (3) gives valid estimates of β. Model (3) yields valid estimates of the variance for parameters that are not forced into the model to deal with the sampling scheme. However, there are constraints on the forced parameters due to the sampling scheme that do not appear in model (3). These constraints reduce the variability to which the forced parameters are subject. Thus, instead of a covariance matrix $D^{-1}(n)$, the appropriate covariance matrix has a term subtracted from $D^{-1}(n)$ to reduce certain aspects of the variability.

10.7 Residual Analysis and Outliers

Residuals are used in regression analysis to check normality, look for serial correlation, examine possible lack of fit, look for heteroscedasticity of variances, identify outliers, and generally to examine whether the assumptions of the regression model appear to be appropriate. Addressing many of these issues is somewhat less appropriate in analysis of variance. For example, appropriate tests for lack of fit are readily available and the question of serial correlation comes up less frequently.

For log-linear models, we will be interested in residuals primarily for identifying outliers and checking approximate normality. We define residuals by analogy with regression analysis.

In a regression model

$$Y = X\beta + e, \tag{1}$$

the residuals $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)$ are defined as the difference between the observations and their estimated expected values. Symbolically,

$$\hat{e} = Y - X\hat{\beta} = (I - H)Y,$$

where $\hat{\beta} = (X'X)^{-1}X'Y$ is the least squares estimate of β and $H = X(X'X)^{-1}X'$. If

$$e \sim N(0, \sigma^2 I)\,,$$

then

$$\hat{e} \sim N(0, \sigma^2(I - H))\,. \tag{2}$$

In most applications of residual analysis, the residuals are standardized before they are used. For example, in checking for outliers, we are checking for residuals that have unusually large absolute values. How large does a residual have to be before it is large enough to cause concern? If we standardize residuals so that they have a variance of about one, then we have a handle on what it means to have a large residual.

There are two methods of standardizing residuals that have been commonly used. One method is a crude standardization

$$\tilde{r}_i = \hat{e}_i/\hat{\sigma},$$

where $\hat{\sigma}^2$ is the mean squared error from fitting model (1). Recall that the object of standardization is to make the variance of the residual about 1. Clearly, we should be dividing the residual by an estimate of its standard deviation. The standard deviation of $\hat{e}_i$ is $\sigma\sqrt{1-h_{ii}}$, where h_{ii} is the ith diagonal element of H. The problem with the crude standardized residuals is that they ignore H. Instead of using the correct distribution (2), crude standardized residuals behave as if $\hat{e} \sim N(0, \sigma^2 I)$. This is the correct distribution for $e = Y - X\beta$, but it ignores the fact that $\hat{\beta}$ is estimated in $\hat{e} = Y - X\hat{\beta}$. In other words, the crude standardized residuals give just that: a very crude standardization. The only advantage to the crude standardized residuals is that they do not require the computation of the h_{ii} values.

The second method of standardizing residuals consists simply of doing it right. The standard deviation of $\hat{e}_i$ is $\sigma\sqrt{1-h_{ii}}$, so define the standardized residual as

$$r_i = \hat{e}_i \Big/ \hat{\sigma}\sqrt{1-h_{ii}}\,.$$

We now argue similarly for log-linear models. Again define the residuals as the difference between the observations and their estimated expected values. Symbolically, the residuals are

$$\hat{e}_i = n_i - \hat{m}_i\,. \tag{3}$$

The need for standardizing these residuals is so glaring that it is almost unheard of to define residuals as in (3). To repeat an intuitive argument given earlier, suppose we have a cell in which $n_i = 7$ and $\hat{m}_i = 2$, then $\hat{e}_i = 7 - 2 = 5$, which is not a very good fit. Now, suppose $n_i = 107$ and $\hat{m}_i = 102$. Again, $\hat{e}_i = 5$, but $\hat{m}_i$ seems to fit n_i quite well. With our standard sampling schemes, variability tends to be large when the numbers n_i and $\hat{m}_i$ are large. For example, under multinomial sampling, for each i, $\mathrm{Var}(n_i) = Np_i(1-p_i) = m_i(N-m_i)/N$. Unless p_i is very close to zero or one, the variance is large when n_i, and implicitly N, are large.

In order to standardize the residuals, we need a relationship similar to (2) for log-linear models. This relationship is essentially that, for large samples, the approximate distribution is

$$n - \hat{m} \sim N(0, D(m)(I - A))\,.$$

A more formal statement is given in the following theorem in which we make explicit the dependence of n and m on the sample size N.

Theorem 10.7.1. For multinomial, or product-multinomial sampling, if the log-linear model $\mu = X\beta$ holds, then as $N \to \infty$,

$$N^{-1/2}(n_N - \hat{m}_N) \xrightarrow{L} N(0, D(I - A))$$

where $D = D(m^*)$, m^* is defined as in Section 3, $A = X(X'DX)^{-1}X'D$, and N is $n_{\cdot}$ for multinomial sampling and $n_{\cdot\cdot}$ for product-multinomial sampling.

Proof. An argument similar to that in the proof of Lemma 12.3.3 gives

$$N^{1/2}[N^{-1}\hat{m}_N - N^{-1}m_N - DAD^{-1}N^{-1}(n_N - m_N)] \xrightarrow{P} 0\,.$$

[This is obtained by doing a Taylor expansion of the function $\hat{m}(\cdot)$ rather than $\hat{\mu}(\cdot)$.] Adding and subtracting $N^{-1/2}n_N$ and multiplying by -1 gives

$$\begin{aligned} N^{1/2}&[N^{-1}(n_N - \hat{m}_N) - N^{-1}(n_N - m_N) + DAD^{-1}N^{-1}(n_N - m_N)] \\ &= [N^{-1/2}(n_N - \hat{m}_N) - (I - DAD^{-1})N^{-1/2}(n_N - m_N)] \xrightarrow{P} 0\,. \end{aligned}$$

It follows that $N^{-1/2}(n_N - \hat{m}_N)$ and $(I - DAD^{-1})N^{-1/2}(n_N - m_N)$ have the same asymptotic distribution. By Theorem 12.3.1,

$$N^{-1/2}(n_N - m_N) \xrightarrow{L} N(0, D - DA_z).$$

Some algebra shows that

$$(I - DAD^{-1})N^{-1/2}(n_N - m_N) \xrightarrow{L} N(0, D(I - A)).$$

□

For large samples, Theorem 10.7.1 gives the approximation

$$n - \hat{m} \sim N(0, D(\hat{m})(I - A(\hat{m})))$$

where $A(\hat{m}) = X(X'D(\hat{m})X)^{-1}X'D(\hat{m})$. We are now in a position to define both standardized residuals and crude standardized residuals. *Crude standardized residuals* are defined by ignoring the fact that m is estimated or, in other words, by ignoring the matrix $A(\hat{m})$. Thus,

$$\tilde{r}_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i}}\,.$$

In discussions of residuals for contingency tables, these values are often called the residuals or the standardized residuals. In previous chapters,

these were referred to as the *Pearson residuals*. As in linear models, the primary advantage of the crude standardized residuals is that they do not require the computation of the diagonal elements of a complicated matrix depending on X. Note also that the sum of the squared crude residuals is precisely the Pearson test statistic for lack of fit.

The *standardized residuals* are defined as

$$r_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i(1 - \hat{a}_{ii})}}$$

where $\hat{a}_{ii}$ is the ith diagonal element of the square matrix $A(\hat{m})$. In some discussions of residuals, $\hat{a}_{ii}$ is defined to be the ith diagonal element of $D(\sqrt{\hat{m}})X(X'D(\hat{m})X)^{-1}X'D(\sqrt{\hat{m}})$. It is easily seen that the diagonal elements of these matrices are the same. These standardized residuals are also known as *adjusted residuals*. The term "adjusted" was introduced to distinguish these from the crude standardized residuals because the crude standardized residuals are often referred to as standardized residuals.

Given the standardized residuals, we can check for normality. Although maximum likelihood estimates and tests based on the likelihood ratio test statistics make sense with any sample size, we have discussed particular confidence intervals and tests that assume the validity of asymptotic distribution theory. We would like to know if this assumption is reasonable. One way to check is to see whether the standardized residuals really seem to be normally distributed. As in regression analysis, we can check this assumption by doing a normal (rankit) plot or a Shapiro-Francia test, cf. Christensen (1996a, Section 2.4). Note that the validity of these procedures depends on having a valid log-linear model.

Another way to check the validity of the asymptotic distributions is by comparing G^2 and X^2. If the asymptotic approximations are good and the model is true, then G^2 and X^2 should be about equal.

If the asymptotic distributions do not seem to be valid, we have a problem. One possible solution is simply to accept the fact that significance levels and confidence coefficients given by asymptotics are very crude. If our assumed sampling schemes are appropriate, the point estimates and test statistics are reasonable, but without valid distributions only crude conclusions can be made. The conditional approaches discussed in Section 3.5 or Bayesian methods similar to Chapter 13 can also be used here. Finally, another possibility is to try to incorporate a more realistic sampling scheme than the simple multinomial and product-multinomial schemes considered here.

The other primary use of standardized residuals is in identifying outliers. Standardized residuals are asymptotically distributed as $N(0, 1)$, so we can test whether residuals really have mean zero. Typically, we would be interested in the standardized residuals with largest absolute values. This is equivalent to testing all residuals, so a multiple comparison method would be appropriate. The Bonferroni method is easy to apply (cf. Christensen,

1996a, Sections 6.2, 7.9 or Christensen, 1996b, Section 5.3). We declare that a case is an outlier if

$$|r_i| > z(1 - \alpha/2q)$$

where $z(\eta)$ is the ηth percentile of a standard normal distribution, α is the size of the test, and q is the number of cells in the table.

An alternative method for identifying outliers is based on *exploratory data analysis*. This method does not rely on asymptotic distributions. Treating the standardized residuals as a sample, compute the quartiles and the interquartile range (IQR). Any case with a residual more than (1.5)IQR from the nearest quartile is considered an outlier.

Rather than using an ad hoc test for outliers based on standardized residuals, we can construct a likelihood ratio test. Suppose our model is

$$\mu = X\beta\,. \tag{4}$$

Without loss of generality, consider testing whether the observation in the qth cell, n_q, is an outlier. An outlier in the qth cell can be modeled by fitting a separate parameter to the cell. Let $v_q = (0, \ldots, 0, 1)'$ and consider the model

$$\mu = X\beta + v_q\gamma\,. \tag{5}$$

The likelihood ratio test of this model against the reduced model (4) is a test of whether the qth cell is an outlier.

Typically, we would want to examine each cell for being an outlier; thus, we need q likelihood ratio test statistics. Again, applying the Bonferroni method for multiple comparisons would be appropriate.

Computing each of the q likelihood ratio test statistics requires an iterative procedure for obtaining estimates in models like model (5). In computing estimates of m, μ, and β in model (5), we can use estimates from model (4) as starting values. To reduce costs, we might stop after just one step of the iterative procedure. For these one-step procedures, closed forms for the likelihood ratio test can be obtained. Unfortunately, there are several possible approaches to deriving one-step approximations and it is not clear which, if any of them, work well. The remainder of this section is devoted to deriving a one-step approximation to Cook's distance.

Derivation of Cook's Distance

Rewrite model (5) as

$$\mu_{[q]} = \log(m_{[q]}) = X\beta + v_q\gamma\,,$$

where $\mu_{[q]}$ and $m_{[q]}$ are used to distinguish the parameters in model (5), where cell q may be an outlier, from the parameters in model (4). The MLE

of $m_{[q]}$ must satisfy

$$\begin{pmatrix} X' \\ v_q' \end{pmatrix} n = \begin{pmatrix} X' \\ v_q' \end{pmatrix} \hat{m}_{[q]}$$

where $\log(\hat{m}_{[q]}) \in C(X, v_q)$. In the discussion below, a subscript (q) indicates that the row corresponding to case q has been deleted from a matrix or vector, so $X = \begin{bmatrix} X_{(q)} \\ x_q' \end{bmatrix}$. Notice that $C(X, v_q) = C\left(\begin{bmatrix} X_{(q)} & 0 \\ 0 & 1 \end{bmatrix}\right)$. Thus, $\hat{m}_{[q]}$ satisfies

$$\begin{bmatrix} X_{(q)}' & 0 \\ 0 & 1 \end{bmatrix} n = \begin{bmatrix} X_{(q)}' & 0 \\ 0 & 1 \end{bmatrix} \hat{m}_{[q]}$$

or, equivalently, writing $\hat{m}_{[q]}' = (\hat{m}_{[q](q)}', \hat{m}_{[q]q})$,

$$X_{(q)}' n_{(q)} = X_{(q)}' \hat{m}_{[q](q)}$$

and

$$n_q = \hat{m}_{[q]q}\,.$$

Thus, $\hat{m}_{[q](q)}$ can be obtained by fitting the model, say $\mu_{(q)} = X_{(q)}\beta_{(q)}$, in which cell q has been deleted and $\beta_{(q)}$ denotes the new parameter vector that applies to this model. In particular, $\hat{\mu}_{[q](q)} = \hat{\mu}_{(q)}$.

A natural version of Cook's distance that is appropriate for log-linear models is

$$C_q(X'D(\hat{m})X, p) = \frac{(\hat{\beta} - \hat{\beta}_{(q)})'X'D(\hat{m})X(\hat{\beta} - \hat{\beta}_{(q)})}{p}.$$

This is the same measure as used in Section 6.7, but is written in a different form. Note that $X'D(\hat{m})X$ is the inverse of the estimated asymptotic covariance matrix for $\hat{\beta}$ under model (4) with Poisson sampling. A one-step version of Cook's distance is

$$C_q^1(X'D(\hat{m})X, p) = \frac{(\hat{\beta} - \hat{\beta}_{(q)}^1)'X'D(\hat{m})X(\hat{\beta} - \hat{\beta}_{(q)}^1)}{p}$$

where $\hat{\beta}_{(q)}^1$ is a one-step approximation to $\hat{\beta}_{(q)}$.

Using the Newton-Raphson method with a starting value of $\hat{\beta}$ and a result similar to Proposition 13.5.1 in Christensen (1996b) on the inverse of a sum of matrices, the one-step estimate is

$$\begin{aligned}
\hat{\beta}_{(q)}^1 &= \hat{\beta} + [X_{(q)}D(\hat{m}_{(q)})X_{(q)}]^{-1}X_{(q)}'[n_{(q)} - \hat{m}_{(q)}] \\
&= \hat{\beta} + [X'D(\hat{m})X - \hat{m}_q x_q x_q']^{-1}[X'(n - \hat{m}) - x_q(n_q - \hat{m}_q)] \\
&= \hat{\beta} + [X'D(\hat{m})X - \hat{m}_q x_q x_q']^{-1}[-x_q(n_q - \hat{m}_q)] \\
&= \hat{\beta} - \left[(X'D(\hat{m})X)^{-1} + \frac{\hat{m}_q}{1 - \hat{a}_{qq}}(X'D(\hat{m})X)^{-1}x_q x_q'(X'D(\hat{m})X)^{-1}\right]
\end{aligned}$$

$$\begin{aligned} & \times [x_q(n_q - m_q)] \\ = \quad & \hat{\beta} - \frac{1}{1-\hat{a}_{qq}}(X'D(\hat{m})X)^{-1}x_q(n_q - \hat{m}_q)\,. \end{aligned}$$

The equation

$$\hat{\beta}^1_{(q)} = \hat{\beta} - \frac{n_q - \hat{m}_q}{1-\hat{a}_{qq}}(X'D(\hat{m})X)^{-1}x_q$$

leads to a computational formula for the one-step version of Cook's distance. The one-step version can be written as

$$\begin{aligned} C^1_q(X'D(\hat{m})X, p) &= \frac{(\hat{\beta} - \hat{\beta}^1_{(q)})'X'D(\hat{m})X(\hat{\beta} - \hat{\beta}^1_{(q)})}{p} \\ &= \frac{1}{p}\frac{\hat{a}_{qq}}{(1-\hat{a}_{qq})^2}\frac{(n_q - \hat{m}_q)^2}{\hat{m}_q} = \frac{1}{p}r^2_q\frac{\hat{a}_{qq}}{1-\hat{a}_{qq}}\,. \end{aligned}$$

As mentioned just prior to Subsection 6.7.1, this definition of Cook's distance has weaknesses. Primarily, it does not take into account the marginal constraints imposed by multinomial or product-multinomial sampling, so it is most appropriate for Poisson sampling. In general, the implications of deleting a cell in a multinomial distribution are hard to grasp. Anderson (1992) has a valuable idea. For multinomial sampling, rather than merely deleting cell i, he proposes looking at the probabilities than an observation occurs in the other cells conditional on the observation not appearing in cell i. He then develops a version of Cook's distance that can be written in terms of the standardized residuals. Unfortunately, Anderson (1992) uses standardized residuals that seem to conflict with Theorem 10.7.1; i.e., he seems to have a different asymptotic variance for $N^{-1/2}(n_i - \hat{m}_i)$. The problem appears to be that he does not use the term A_z in Theorem 10.3.1b. [Of course, the really fun thing about this is that the reader gets to wonder who is making the mistake. Anderson's standardized residuals are based on Rao (1973, p. 394). Theorem 10.7.1 is, I believe, identical to a result in Haberman (1974a).] In any case, the reported results should be used with care. In other work, Thomas and Cook (1989, 1990) discuss influence for generalized linear models (which include log-linear models, cf. Chapter 9).

10.8 Exercises

EXERCISE 10.8.1. Show that for a 3×3 table with $n_{11} = n_{13}$, $n_{31} = n_{33}$, and $n_{\cdot 1} = n_{\cdot 3}$ that the $\hat{m}$'s for the independence model are also the $\hat{m}$'s for the uniform association model.

EXERCISE 10.8.2. Waite (1911) reports data on classifications of general intelligence made for students from a secondary school in London. Classifications were made after two different school terms and each student

was classified by two instructors. Classifications were based on Pearson's criteria as explained in Exercise 2.6.3. The data are given in Table 10.1. The entry for Term 1, row C, column D of 55 indicates that there were 55 people who were classified as C by one instructor and D by the other. We want to develop interesting log-linear models for such data. For the moment, consider only Term 1. A model of interest is that the two teachers have the same marginal distribution of assigning various classifications and that they assign them independently. Write the marginal probabilities for the categories C, D, E, F, and G as p_1, p_2, p_3, p_4, and p_5, respectively. What are the table probabilities p_{ij} in terms of the marginal probabilities? Take logs of the p_{ij}'s to identify a log-linear model. Write the log-linear model $\log(m_{ij}) = \alpha_i + \beta_j$ for these data in matrix form. Incorporate the restrictions $\alpha_i = \beta_i$ to get a model $\log(m) = X\gamma$ and fit the model to the data of Table 10.1. What conclusions can you reach about the data?

TABLE 10.1. Intelligence Classifications

Term 1

	C	D	E	F	G
C	13	55	29	1	0
D		123	326	24	0
E			421	253	17
F				107	31
G					5

Term 2

	C	D	E	F	G
C	17	51	17	1	0
D		129	479	46	5
E			700	343	28
F				109	72
G					21

EXERCISE 10.8.3. Use the saturated model and Theorem 10.2.1 to find the large sample distribution of a multinomial sample in terms of its category probabilities.

EXERCISE 10.8.4. *The Delta Method.*
Let v_N be a sequence of $q \times 1$ random vectors and suppose that

$$\sqrt{N}(v_N - \theta) \xrightarrow{L} N\big(0, \Sigma(\theta)\big).$$

Suppose that $F(\cdot)$ is a differentiable function taking q vectors into r vectors. Let dF be the $r \times q$ matrix of partial derivatives of F. Then

$$\sqrt{N}\big(F(v_N) - F(\theta)\big) \xrightarrow{L} N\big(0, dF\,\Sigma(\theta)\,dF'\big).$$

For technical reasons, it is advantageous to assume that dF and $\Sigma(\theta)$ are also continuous. For mathematical details, see Bishop, Fienberg, and Holland (1975, Section 14.6.3).

Assuming the saturated model for a 2×2 table, use Theorem 10.2.1c and the delta method to find an asymptotic standard error and asymptotic confidence intervals for the odds ratio. Show that the intervals do not change if based on Theorem 10.3.1b. Apply this method to the data of Example 2.1.1 to get a 95% interval. How does this interval compare to the interval given at the end of the subsection on The Odds Ratio in Section 2.1?

EXERCISE 10.8.5. Use the delta method of the previous exercise to show that if

$$\sqrt{N}(v_N - \theta) \xrightarrow{L} N\big(0, \Sigma(\theta)\big),$$

then for any $r \times q$ matrix A,

$$\sqrt{N}(Av_N - A\theta) \xrightarrow{L} N\big(0, A\Sigma(\theta)A'\big).$$

Show that if $\mathrm{Cov}(\sqrt{N}v_N) = \Sigma(\theta)$, then $\mathrm{Cov}(\sqrt{N}Av_N) = A\Sigma(\theta)A'$.

EXERCISE 10.8.6. *Testing Marginal Homogeneity in a Square Table.* For an $I \times I$ table, the hypothesis of marginal homogeneity is

$$H_0 : \; p_{i\cdot} = p_{\cdot i}\,,$$

$i = 1, \ldots, I$. A natural statistic for testing this hypothesis is the vector $d = (n_{1\cdot} - n_{\cdot 1}, \ldots, n_{I\cdot} - n_{\cdot I})'$ Clearly,

$$\mathrm{E}(d_i) = p_{i\cdot} - p_{\cdot i}\,.$$

Use the results of Exercise 1.6.5 to show that

$$\mathrm{Var}(d_i) = N\left[(p_{i\cdot} + p_{\cdot i} - 2p_{ii}) - (p_{i\cdot} - p_{\cdot i})^2\right]$$

and

$$\mathrm{Cov}(d_h, d_i) = N\left[(p_{hi} + p_{ih}) + (p_{h\cdot} - p_{\cdot h})(p_{i\cdot} - p_{\cdot i})\right].$$

Use the previous exercise to find the large sample distribution of d. Show that $\Pr(J'd = 0) = 1$. Show that the asymptotic covariance matrix of d has rank $I - 1$ and thus is not invertible. It is well known that if $Y \sim N(0, V)$ with V an $s \times s$ nonsingular matrix, then $Y'V^{-1}Y \sim \chi^2(s)$. Use this fact along with the asymptotic distribution of d to obtain a test of the hypothesis of marginal homogeneity. Apply the test of marginal homogeneity to the data of Exercise 2.6.10.

11

The Matrix Approach to Logit Models

In this chapter, we again discuss logistic regression and logit models, but here we use the matrix approach of Chapter 10. Section 1 discusses the equivalence of logit models and log-linear models. This equivalence is used to arrive at results on estimation and testing. Because the data in a typical logistic regression correspond to very sparse data in a contingency table, the asymptotic results of Section 10.2 are not appropriate. Section 6 presents results from Haberman (1977) that *are* appropriate for logistic regression models. Section 2 discusses model selection criteria for logistic regression. Direct fitting of logit models is considered in Section 3. The appropriate maximum likelihood equations and Newton-Raphson procedure are given. Section 4 indicates how the weighted least squares model-fitting procedure is applied to logit models. Models appropriate for response variables with more than two categories are examined in Section 5. Finally, Section 7 considers the discrimination problem.

11.1 Estimation and Testing for Logistic Models

In general, if the dependent variable has only two categories, regardless of the number of predictor variables, the table can be considered as a two-dimensional table with two columns (one column for each category of the dependent variable). In this structure, all predictor variables are being pooled into the rows of the table. For t distinct sets of predictor variables,

we can write the $2t \times 1$ vector of observations as

$$n = (n_{11}, n_{21}, \ldots, n_{t1}, n_{12}, \ldots, n_{t2})'$$

with similar notations for p, m, and μ. A logistic model is a linear model for the values $\log(p_{i1}/p_{i2})$. Note that we are modeling the log odds of category 1 compared to category 2. These are the log odds of observing category 1 given that the observation falls in row i. If each row constitutes an independent binomial, then we are simply modeling the log odds for the various binomials.

Logistic models are nothing more than log-linear models. All of the results of Chapter 10 apply to logistic models. We now consider the exact nature of this equivalence for prospective studies.

Let $\eta = \big(\log(p_{11}/p_{12}), \log(p_{21}/p_{22}), \ldots, \log(p_{t1}/p_{t2})\big)'$ be the vector of log odds. A linear logistic model is a model $\eta = X\beta$, where X is a $t \times k$ matrix. Define

$$L' = [I_t, -I_t]$$

and note that for prospective studies, $\log(p_{i1}/p_{i2}) = \log(m_{i1}/m_{i2}) = \log(m_{i1}) - \log(m_{i2})$, so

$$\eta = L'\mu\,.$$

Thus, the logistic model can be written as

$$L'\mu = X\beta\,.$$

Now define a log-linear model

$$\mu = X_*\xi$$

where

$$X_* = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \quad \text{and} \quad \xi = \begin{bmatrix} \gamma \\ \beta \end{bmatrix}.$$

It is easily seen that if $\mu = X_*\xi$, then $L'\mu = X\beta$. It is only moderately more difficult to see (cf. Section 12.4 and Christensen, 1996b, Section 3.3) that

$$\{\mu | L'\mu = X\beta\} = C(X_*)\,.$$

Thus, the restriction on μ imposed by the logistic model $\eta = X\beta$ is precisely the same as the log-linear model $\mu = X_*\xi$. In other words, *the logistic model $\eta = X\beta$ is identical to the log-linear model $\mu = X_*\xi$.*

Unfortunately, there can be problems with the asymptotic results of Chapter 10 when applied to logistic models. The asymptotic results of Section 10.2 are based on the assumption of a fixed number of cells q in the table. This number is $q = 2t$. It is assumed that the sample size in each cell gets large. Often, logistic models are used in situations that are more similar to regression than analysis of variance. In such cases, any additional

observations obtained typically correspond to new rows of the design matrix X. This implies the addition of a new row to the $t \times 2$ table; thus, the assumption of a fixed number of cells is invalidated. These issues are dealt with in Section 6.

In practice, the data are fixed and neither the sample sizes nor the number of cells increases. Both remain constant. If the number of observations in each cell is reasonably large, the usual asymptotic theory should work adequately. If the number of cells is large relative to the number of observations, new asymptotic results are required as a basis for statistical inference.

Frequently, when the dependent variable has two categories, the data are collected so that for each unique set of predictor variables (i.e., for each row of the $t \times 2$ table), the counts are independent and have a binomial distribution. As in Section 10.4, the existence of product-multinomial sampling (product-binomial, in this instance) restricts us to a special form for the design matrix of the log-linear model. In particular, it is required that $C(Z) \subset C(X_*)$ where Z is a matrix of indicators for the rows of the $t \times 2$ table. With $\mu = (\mu_{11}, \ldots, \mu_{t1}, \mu_{12}, \ldots, \mu_{t2})'$,

$$Z = \begin{bmatrix} I_t \\ I_t \end{bmatrix}.$$

Since

$$X_* = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix},$$

it is clearly the case that $C(Z) \subset C(X_*)$. It follows that hypothesizing a logistic model automatically makes the model appropriate for product-binomial sampling.

Moreover, when fitting logistic models, it suffices to imagine fitting a log-linear model to a table with product-binomial sampling. This mental device of imagining product-binomial sampling assures that the structure of the log-linear model implies the existence of a valid logistic model. To see this, note that a quite arbitrary model appropriate for product-binomial sampling can be written with the design matrix

$$\begin{bmatrix} I_t & X_1 \\ I_t & X_2 \end{bmatrix}.$$

However, $C\left(\begin{bmatrix} I_t & X_1 \\ I_t & X_2 \end{bmatrix}\right) = C\left(\begin{bmatrix} I_t & X_1 - X_2 \\ I_t & 0 \end{bmatrix}\right)$ where $\begin{bmatrix} I_t & X_1 - X_2 \\ I_t & 0 \end{bmatrix}$ has the form given earlier for logistic models.

Estimation

We now consider the problem of estimation in a logistic model $\eta = X\beta$. Recall that X is $t \times k$ and that η and β are t and k vectors, respectively.

In particular, consider estimation of a linear function $\rho'\eta$, where ρ is an arbitrary $t \times 1$ vector. Note that $\rho'\eta = \rho' X\beta$, so these functions can be considered as functions of β. If $\text{rank}(X) = k$ and e_i is a t vector of 0s with a 1 in the ith position, then choosing $\rho' = e_i'(X'X)^{-1}X'$ gives $\rho' X\beta = \beta_i$, where $\beta = (\beta_1, \ldots, \beta_k)'$.

Write

$$X_L = \begin{bmatrix} X \\ 0 \end{bmatrix},$$

where X_L is $2t \times k$ and 0 is a $t \times k$ matrix of zeros. As above, $\eta = X\beta$ is equivalent to

$$\begin{aligned} \mu = X_*\xi &= \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} = [Z, X_L] \begin{bmatrix} \gamma \\ \beta \end{bmatrix} \\ &= Z\gamma + X_L\beta\,, \end{aligned}$$

where β is identical in the logistic and log-linear models.

Using invariance of the MLEs, the MLE of $\rho'\eta$ comes directly from the MLE of μ. Because $\eta = L'\mu$,

$$\rho'\hat{\eta} = \rho' L'\hat{\mu}\,.$$

In terms of estimating β, we get

$$\begin{aligned} \rho' X\hat{\beta} &\equiv \rho'\hat{\eta} \\ &= \rho' L'\hat{\mu} \\ &= \rho' L'(Z\hat{\gamma} + X_L\hat{\beta}) \\ &= \rho' L' X_L\hat{\beta} \\ &= \rho' X\hat{\beta}\,. \end{aligned}$$

Thus, $\rho' X\beta$ can be estimated in either the logistic model or the log-linear model and the estimates are identical.

To form tests and confidence intervals for $\rho' X\beta$, we need a distribution for $\rho' X\hat{\beta}$. Asymptotically, for any vector ρ_*,

$$\frac{\rho_*'\hat{\mu} - \rho_*' X_*\xi}{\sqrt{\rho_*'(A - A_z)D^{-1}(m)\rho_*}} \sim N(0,1) \tag{1}$$

and

$$\begin{aligned} A - A_z &= X_*(X_*'DX_*)^{-1}X_*'D - Z(Z'DZ)^{-}Z'D \\ &= X_*(X_*'D(m)X_*)^{-1}X_*'D(m) - Z'(Z'D(m)Z)^{-1}Z'D(m)\,, \end{aligned}$$

so

$$(A - A_z)D^{-1}(m) = X_*(X_*'D(m)X_*)^{-1}X_*' - Z(Z'D(m)Z)^{-1}Z'\,.$$

For estimating $\rho' X\beta$ in a logistic model, let $\rho'_* = \rho' L'$, so $\rho'_* \hat{\mu} = \rho' \hat{\eta}$ and $\rho'_* X_* \xi = \rho' X \beta$. To apply (1), we need to find $\rho'_*(A - A_z)D^{-1}(m)\rho_*$. In the appendix to this section, it is shown that

$$\rho'_*(A - A_z)D^{-1}(m)\rho_* = \rho' X[X'D(b)X]^{-1}X'\rho\,, \tag{2}$$

where

$$b = (b_1, \ldots, b_t)'$$

and

$$b_i = m_{i1}m_{i2}/(m_{i1} + m_{i2})\,.$$

Taking $\hat{b}_i = \hat{m}_{i1}\hat{m}_{i2}/n_{i\cdot}$ and $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_t)'$ gives, asymptotically,

$$\frac{\rho'\hat{\eta} - \rho' X\beta}{\sqrt{\rho' X[X'D(\hat{b})X]^{-1}X'\rho}} \sim N(0,1)\,.$$

Tests and confidence intervals follow in the usual way. In particular, using $\rho' = e'_i = (0, \cdots, 0, 1, 0, \cdots, 0)$, for large samples

$$\frac{\log(\hat{p}_{i1}/\hat{p}_{i2}) - \log(p_{i1}/p_{i2})}{\sqrt{\hat{a}_{ii}/\hat{b}_i}} \sim N(0,1)\,, \tag{3}$$

where $\hat{a}_{ii}$ is the leverage as found in Section 4.3 as modified by Subsection 4.4.1. The asymptotic variance for log odds ratios can also be computed. Except for the difference in the weights b_i, the value

$$\mathrm{Var}(\rho'\hat{\eta}) = \rho' X\left(X'D(b)X\right)^{-1}X'\rho$$

looks like that used in Section 10.2. Methods for evaluating variances are similar.

Using the delta method of Exercise 10.8.4, the logistic transform, (3), and writing $N_i = n_{i1} + n_{i2}$, asymptotic inferences for p_{ij} are based on

$$\frac{\hat{p}_{ij} - p_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})\hat{a}_{ii}/N_i}} \sim N(0,1)\,,$$

cf. Exercise 11.8.5

Testing Hypotheses

Assume a logistic model $\eta = X\beta$ and consider the problem of testing a reduced model $\eta_0 = X_0\beta_0$ against $\eta = X\beta$, where $C(X_0) \subset C(X)$. This test can be performed by testing log-linear models. The full model corresponds to $\mu = X_*\xi$. We can write the reduced model as

$$\mu_0 = X_{*0}\xi_0,$$

where

$$X_{*0} = [Z, X_{L0}]$$

and

$$X_{L0} = \begin{bmatrix} X_0 \\ 0 \end{bmatrix}.$$

The degrees of freedom for the chi-square test is $\text{rank}(X_*) - \text{rank}(X_{*0}) = \text{rank}(X) - \text{rank}(X_0)$. The likelihood ratio test statistic is

$$G^2 = 2 \sum_{i=1}^{t} \sum_{j=1}^{2} \hat{m}_{ij} \log(\hat{m}_{ij}/\hat{m}_{0ij}).$$

Appendix

To simplify notation somewhat, let

$$D_m \equiv D(m).$$

In this appendix, we wish to show equation (2), i.e., for $\rho'_* = \rho' L'$,

$$\rho'_* X_* [X'_* D_m X_*]^{-1} X'_* \rho - \rho'_* Z [Z' D_m Z]^{-1} Z' \rho_* = \rho' X [X' D(b) X]^{-1} X' \rho.$$

The algebra necessary for this demonstration is quite nasty. We break it up into several parts.

Lemma 11.1.1. $\rho'_* Z [Z' D_m Z]^{-1} Z' \rho_* = 0.$

Proof. $\rho'_* Z = \rho' L' Z$ but $L' Z = 0$. □

Now, all we have to deal with is

$$\rho'_* X_* [X'_* D_m X]^{-1} X'_* \rho.$$

We will rewrite this using a perpendicular projection operator (cf. Christensen, 1996b, Appendix B) and then use a property of perpendicular projection operators to derive the result. Let

$$D_m^{1/2} = D(\sqrt{m_{11}}, \ldots, \sqrt{m_{t2}})$$

so that

$$\begin{aligned} & \rho'_* X_* [X'_* D_m X_*]^{-1} X'_* \rho_* \\ &= \rho' L' X_* [X'_* D_m X_*]^{-1} X'_* L \rho \\ &= \rho' L' D_m^{-1/2} \left[D_m^{1/2} X_* [X'_* D_m X_*]^{-1} X'_* D_m^{1/2} \right] D_m^{-1/2} L \rho \qquad (4) \\ &= \rho' L' D_m^{-1/2} P D_m^{-1/2} L \rho, \end{aligned}$$

where $P = D_m^{1/2} X_*[X_*' D_m X_*]^{-1} X_*' D_m^{1/2}$. The matrix P is the perpendicular projection operator onto

$$C(D_m^{1/2} X_*) = C(D_m^{1/2} Z, D_m^{1/2} X_L).$$

We need one property of the perpendicular projection operator (cf. Christensen, 1996b, Section 9.2), namely

$$P = M_1 + M_2,$$

where

$$M_1 = D_m^{1/2} Z[Z' D_m Z]^{-1} Z' D_m^{1/2} \tag{5}$$

and

$$M_2 = (I - M_1) D_m^{1/2} X_L [X_L' D_m^{1/2} (I - M_1) D_m^{1/2} X_L]^{-1} \times X_L' D_m^{1/2} (I - M_1). \tag{6}$$

From (4), we see that

$$\rho_*' X_* [X_*' D_m X_*]^{-1} X_*' \rho = \rho' L' D_m^{-1/2} M_1 D_m^{-1/2} L\rho + \rho' L' D_m^{-1/2} M_2 D_m^{-1/2} L\rho\,.$$

The first term on the right-hand side vanishes.

Lemma 11.1.2. $\rho' L' D_m^{-1/2} M_1 D_m^{-1/2} L\rho = 0.$

Proof. Note that $0 = L'Z = L' D_m^{-1/2} [D_m^{1/2} Z]$. Using formula (5) for M_1, we see that $0 = L' D_m^{-1/2} M_1$. □

By Lemmas 11.1.1 and 11.1.2, we have reduced the demonstration of equation (2) to showing

$$\rho' L' D_m^{-1/2} M_2 D_m^{-1/2} L\rho = \rho' X [X' D(b) X]^{-1} X' \rho\,. \tag{7}$$

Again, we break the demonstration into parts.

Lemma 11.1.3. $\rho' L' D_m^{-1/2} (I - M_1) D_m^{1/2} X_L = \rho' X.$

Proof. As in the proof of Lemma 11.1.2, $L' D_m^{-1/2} M_1 = 0$. Thus,

$$\begin{aligned} \rho' L' D_m^{-1/2} (I - M_1) D_m^{1/2} X_L &= \rho' L' D_m^{-1/2} D_m^{1/2} X_L \\ &= \rho' L' X_L \\ &= \rho' X\,. \end{aligned}$$

□

Define $m_1 = (m_{11}, \ldots, m_{t1})'$ and $m_2 = (m_{12}, \ldots, m_{t2})'$.

Lemma 11.1.4. $X_L' D_m Z = X' D(m_1)$.

Proof.

$$\begin{aligned} X_L' D_m Z &= [X', 0'] \begin{bmatrix} D(m_1) & 0 \\ 0 & D(m_2) \end{bmatrix} \begin{bmatrix} I_t \\ I_t \end{bmatrix} \\ &= X' D(m_1)\,. \end{aligned}$$

□

A similar argument yields

Lemma 11.1.5. $X_L' D_m X_L = X' D(m_1) X$.

We need two additional results.

Lemma 11.1.6. $[Z' D_m Z] = D(m_1 + m_2)$.

Proof.

$$\begin{aligned} Z' D_m Z &= [I_t, I_t] \begin{bmatrix} D(m_1) & 0 \\ 0 & D(m_2) \end{bmatrix} \begin{bmatrix} I_t \\ I_t \end{bmatrix} \\ &= D(m_1) + D(m_2) \\ &= D(m_1 + m_2). \end{aligned}$$

□

Lemma 11.1.7. $X_L' D_m^{1/2} (I - M_1) D_m^{1/2} X_L = X' D(b) X$.

Proof. Using Lemmas 11.1.4, 11.1.5, and 11.1.6 gives

$$\begin{aligned} & X_L' D_m^{1/2} (I - M_1) D_m^{1/2} X_L \\ &= X_L' D_m X_L - X_L' D_m Z [Z' D_m Z]^{-1} Z' D_m X_L \\ &= X' D(m_1) X - X' D(m_1) D^{-1} (m_1 + m_2) D(m_1) X \\ &= X' [D(m_1) - D(m_1) D^{-1} (m_1 + m_2) D(m_1)] X \\ &= X' D(b) X\,. \end{aligned}$$

□

We can now obtain equation (7). Using (6), Lemmas 11.1.3 and 11.1.7 give

$$\rho' L' D_m^{-1/2} M_2 D_m^{-1/2} L \rho = \rho' X [X' D(b) X]^{-1} X' \rho$$

and we are done.

11.2 Model Selection Criteria for Logistic Regression

The purpose of this section is to show that not only do the model selection criteria of Section 3.6 apply to logistic regression, but that they have interpretations similar to those in normal theory regression.

For a logistic regression model $\eta = X\beta$ or, equivalently,

$$\mu = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix},$$

write the likelihood ratio test statistic for testing the model against the saturated model as $G^2(X)$. Let J be a $t \times 1$ matrix of one's.

A natural definition of R^2 for logit models gives precisely the same definition as for log-linear models. In defining R^2 for logistic regression, the smallest interesting model is typically the model that contains only the intercept

$$\eta = J\gamma\,.$$

The corresponding log-linear model is

$$\mu = \begin{bmatrix} I_t & J \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \gamma \end{bmatrix} \tag{1}$$

which is equivalent to

$$\mu = X_0\delta \equiv \begin{bmatrix} I_t & J & 0 \\ I_t & 0 & J \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \end{bmatrix}. \tag{2}$$

These are equivalent because the μ vectors that can be obtained from the models are identical. Any μ from model (1) can be obtained from model (2) by taking $\alpha = \xi$, $\gamma_1 = \gamma$, and $\gamma_2 = 0$. Conversely, any μ from model (2) can be obtained from model (1) by taking $\xi = \alpha + J\gamma_2$ and $\gamma = \gamma_1 - \gamma_2$. This is really just a demonstration that $C(X_0)$ is identical to the column space of the model matrix in (1). If we think of logistic regression as the analysis of a $t \times 2$ table, model (2) is

$$\log(m_{ij}) = \alpha_i + \gamma_j$$

which is the model of independence. So comparing the logistic model $\eta = X\beta$ to the logistic intercept model $\eta = J\gamma$ is the same as comparing the log-linear equivalent of $\eta = X\beta$ to the independence model.

The $G^2(X)$ statistic is the same whether considering logit models or their log-linear equivalents. If $k = \text{rank}(X)$, the degrees of freedom are the number of cells in the table minus the rank of the log-linear model design matrix, $2t - (t + k) = t - k$. Note that this can also be viewed as the number

of cases (number of independent binomials) minus the rank of the logistic model design matrix, just like the degrees of freedom error in a normal theory model. The independence model is equivalent to fitting an intercept in logistic regression, so $G^2(X_0)$ has degrees of freedom $2t - t - 1 = t - 1$. Note that this is the number of cases minus 1 for the intercept, just like the error for fitting only an intercept in normal theory regression.

The log-linear model definition

$$R^2 = \frac{G^2(X_0) - G^2(X)}{G^2(X_0)}$$

(cf. Section 3.6) makes perfect sense when applied to logistic regression models. The definition of Adj R^2 when applied to logit models gives

$$\text{Adj } R^2 = 1 - \frac{G^2(X)/(t-k)}{G^2(X_0)/(t-1)}\,.$$

Finally, Akaike's information criterion suggests picking X to minimize

$$\begin{aligned} A_x &= G^2(X) - \big[(2t) - 2\big(t+k\big)\big] \\ &= G^2(X) + 2k. \end{aligned}$$

Because t is fixed, minimizing A_x is equivalent to minimizing

$$A_x - q \equiv A_x - 2t = G^2(X) - 2[t-k]\,.$$

As illustrated in Section 4.1, it is common to report A^*, the information relative to a full model.

11.3 Likelihood Equations and Newton-Raphson

When dealing with logit models, some simplification occurs in the likelihood equations and the Newton-Raphson algorithm. Write the log-linear model version of the logit model $\eta = X\beta$ as

$$\mu = \begin{bmatrix} I_t & X \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix}, \tag{1}$$

where X is a $t \times k$ matrix. Write $m_j = (m_{1j}, \ldots, m_{tj})'$ and $n_j = (n_{1j}, \ldots, n_{tj})'$ for $j = 1, 2$, so $m' = (m_1', m_2')$ and $n' = (n_1', n_2')$. Also write $N = (N_1, \ldots, N_t)'$, where

$$N_i = n_{i1} + n_{i2}\,.$$

As in Chapter 10, the likelihood equations are

$$\begin{bmatrix} I_t' & I_t' \\ X' & 0 \end{bmatrix} (n-m) = 0\,.$$

Noting that $n_2 = N - n_1$, we get the equations

$$\begin{bmatrix} I_t' & I_t' \\ X' & 0 \end{bmatrix} \begin{bmatrix} n_1 - m_1 \\ N - n_1 - m_2 \end{bmatrix} = 0$$

or

$$\begin{bmatrix} (n_1 - m_1) + (N - n_1 - m_2) \\ X'(n_1 - m_1) \end{bmatrix} = 0\,,$$

which simplifies to

$$\begin{bmatrix} N - (m_1 + m_2) \\ X'(n_1 - m_1) \end{bmatrix} = 0\,. \tag{2}$$

We are seeking values $\hat{\gamma}$ and $\hat{\beta}$ that give solutions to equation (2). We now show that the value of $\hat{\gamma}$ can be determined by the value of $\hat{\beta}$. Write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_t' \end{bmatrix}$$

and $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_t)'$; then,

$$\begin{aligned} \hat{m}_{i1} &= \exp[\hat{\gamma}_i + x_i'\hat{\beta}]\,, \\ \hat{m}_{i2} &= \exp[\hat{\gamma}_i]\,. \end{aligned} \tag{3}$$

Because $\hat{\gamma}$ and $\hat{\beta}$ provide a solution to (2), $N_i = \hat{m}_{1i} + \hat{m}_{2i}$ and

$$\begin{aligned} N_i &= \exp[\hat{\gamma}_i + x_i'\hat{\beta}] + \exp[\hat{\gamma}_i] \\ &= e^{\hat{\gamma}_i}[1 + \exp(x_i'\hat{\beta})]\,, \end{aligned}$$

so

$$e^{\hat{\gamma}_i} = N_i/[1 + \exp(x_i'\hat{\beta})] \tag{4}$$

and

$$\hat{\gamma}_i = \log(N_i/[1 + \exp(x_i'\hat{\beta})])\,.$$

This completes the demonstration.

Because $\hat{\gamma}$ is a function of $\hat{\beta}$, the likelihood equations can be reduced to the bottom half of (2), which is solely a function of β. The top half of (2) is satisfied for any $\hat{\beta}$ by taking $\hat{\gamma}$ as indicated above. To write the bottom

half of (2) as a function of β, we need only write m_1 as a function of β. From equations (3) and (4)

$$\begin{aligned} m_{i1} &= e^{\gamma_i} e^{x_i'\beta} \\ &= (N_i/[1+e^{x_i'\beta}])e^{x_i'\beta} \\ &= N_i \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} . \end{aligned} \tag{5}$$

We can use the Newton-Raphson algorithm to find a solution to the likelihood equations

$$X'(n_1 - m_1) = 0 .$$

The highlights of using Newton-Raphson are given below. More detailed discussions are given in Chapter 10 and Section 12.4. To apply Newton-Raphson, we need the derivative of $X'(n_1 - m_1(\beta))$ with respect to β, i.e., the matrix of partial derivatives with respect to the β_j's. Using the chain rule, this is just $-X'$ times the matrix of partial derivatives of $m_1(\beta)$ with respect to β. Note that

$$\begin{aligned} m_1(\beta) &= [m_{11}(\beta), \ldots, m_{t1}(\beta)]' \\ &= \left[N_1 e^{x_1'\beta}/(1+e^{x_1'\beta}), \ldots, N_t e^{x_t'\beta}/(1+e^{x_t'\beta})\right]' . \end{aligned}$$

The partial derivative of $m_{1i}(\beta)$ with respect to β_j is

$$\begin{aligned} \frac{\partial m_{i1}(\beta)}{\partial \beta_j} &= N_i x_{ij} e^{x_i'\beta}(1+e^{x_i'\beta})^{-1} \\ &\quad + N_i e^{x_i'\beta}(-1)(1+e^{x_i'\beta})^{-2} x_{ij} e^{x_i'\beta} \\ &= N_i x_{ij} \left[\frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})} - \left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)^2\right] . \end{aligned} \tag{6}$$

Recalling equation (5), define

$$p_i \equiv m_{i1}/N_i = e^{x_i'\beta}\Big/(1+e^{x_i'\beta}) . \tag{7}$$

For product-binomial sampling, p_i is the probability of an observation in the ith row occurring in the first column of the table. Substituting p_i into (6), the partial derivative is

$$\begin{aligned} \frac{\partial m_{i1}(\beta)}{\partial \beta_j} &= N_i x_{ij}(p_i - p_i^2) \\ &= N_i p_i (1-p_i) x_{ij} . \end{aligned}$$

The matrix of partial derivatives can be written as

$$\begin{aligned} dm_1(\beta) &= \begin{bmatrix} N_1p_1(1-p_1)x_{11} & \cdots & N_1p_1(1-p_1)x_{1k} \\ \vdots & & \vdots \\ N_tp_t(1-p_t)x_{t1} & \cdots & N_tp_t(1-p_t)x_{tk} \end{bmatrix} \\ &= D(N_ip_i(1-p_i))X\,. \end{aligned}$$

The matrix of partial derivatives for $X'(n - m(\beta))$ is then

$$X'dm_1(\beta) = X'D(N_ip_i(1-p_i))X\,.$$

Note that $N_ip_i(1-p_i) = b_i$ from (11.1.2).

We can now apply the Newton-Raphson algorithm. Given a current estimate β_s, the next estimate is

$$\beta_{s+1} = \beta_s + \delta_s\,,$$

where

$$\delta_s = [X'D(N_ip_i(1-p_i))X]^{-1}X'(n_1 - m_1(\beta_s))$$

and p_i in $N_ip_i(1-p_i)$ is actually $p_i = p_i(\beta_s)$ as defined by equation (7).

As with log-linear models, the estimates can be found by doing a series of weighted regressions. Let $y_i = x_i'\beta_s + [n_{i1} - m_{i1}(\beta_s)]/N_ip_i(1-p_i)$, so that

$$Y = X\beta_s + D(N_ip_i(1-p_i))^{-1}(n_1 - m_1(\beta_s))\,.$$

If the weighted regression model

$$Y = X\beta + e\,,\ \mathrm{E}(e) = 0\,,\ \mathrm{Cov}(e) = D^{-1}(N_ip_i(1-p_i))$$

is fitted, then the estimate of β is

$$\begin{aligned} \beta_{s+1} &= [X'D(N_ip_i(1-p_i))X]^{-1}X'D(N_ip_i(1-p_i))Y \\ &= \beta_s + \delta_s\,. \end{aligned}$$

11.4 Weighted Least Squares for Logit Models

The methods introduced by Grizzle, Starmer, and Koch (1969) are actually quite general and can be applied to logit models as well as log-linear models, cf. Section 10.6. As before, asymptotic properties of the GSK method are often the same as maximum likelihood, but the small sample justification is less compelling. Moreover, for some small samples, the GSK method cannot be used at all unless ad hoc modifications to the data are introduced.

As with log-linear models, the GSK method amounts to performing one step of the Newton-Raphson algorithm. For a logit model

$$\eta = X\beta\,,$$

where

$$\begin{aligned}\eta &= [\mu_{11} - \mu_{21}, \ldots, \mu_{1t} - \mu_{2t}]' \\ &= [\log(p_{11}/p_{21}), \ldots, \log(p_{1t}/p_{2t})]' .\end{aligned}$$

The model

$$Y = X\beta + e\,,\ \mathrm{E}(e) = 0\,,\ \mathrm{Cov}(e) = D^{-1}(N_i\hat{p}_i(1-\hat{p}_i))$$

is fitted where

$$\hat{p}_i \equiv \hat{p}_{i1} = n_{i1}/N_i \tag{1}$$

and

$$Y = [y_1, \ldots, y_t]'$$

with

$$y_i = \log[\hat{p}_i/(1-\hat{p}_i)]\,.$$

The justification for the GSK procedure is asymptotic. The estimates obtained are asymptotically optimal if the N_i's are all large. As the justification for the GSK procedure is based on its large sample optimality, there is no apparent reason to use GSK for small samples. Moreover, it is not clear that the sort of asymptotic arguments which involve large numbers of small samples can be extended to the GSK approach.

Perhaps the most obvious difficulty with the GSK approach is that if $\hat{p}_i$ is either zero or one, y_i is not defined. When $N_i = 1$, y_i is always undefined. Of course, the corresponding weight from the inverse of the covariance matrix is also zero, so one could argue that GSK simply ignores such cases. But this could result in ignoring great quantities of data. For the Chapman data of Section 4.1, all of the data would be ignored. An ad hoc correction for this problem has been proposed, which is simply to substitute for any value $\hat{p}_i$ that is 0 or 1, the values $\hat{p}_i \pm \epsilon_i$, where the substitution forces $\hat{p}_i$ to be between 0 and 1 and ϵ_i is some small number. It is frequently suggested that any values $n_{ij} = 0$ be replaced by $n_{ij} = 0.5$.

It may be noted that the $\hat{p}_i$'s given in (1) are also the natural starting values for the Newton-Raphson algorithm and that small samples also require that $\hat{p}_i$'s of 0 or 1 be adjusted before they can be used. The key difference is that the justification for MLEs does not depend on properties of the starting values. Any starting values that lead to MLEs are perfectly acceptable. The GSK method depends crucially on the initial estimates.

The details of a GSK logit model analysis will not be given because, for uniformly large samples, they are exactly analogous to the log-linear model analysis given in Section 10.6. For large N_i's, the SSE can be used to give a chi-square test for lack of fit. The degrees of freedom for the test are the degrees of freedom for error. Models can be compared by comparing sums of squares for error. Reported standard errors must be corrected for the root mean square error; t statistics must be corrected for the root mean square error and compared to the standard normal distribution. The only difference is that a valid standard error exists for the intercept.

11.5 Multinomial Response Models

An integral point in the definition of logistic regression models is that the response variable has only two categories. The model posits a linear mean structure for $\log(m_{i1}/m_{i2})$. As discussed in the previous chapter, if the response variable has more than two categories, it is by no means clear how to extend the logistic regression model to deal with the additional categories. It may then be somewhat surprising to find that it is clear how to extend the log-linear model version of the logistic regression model to more than two categories. We will discuss the appropriate log-linear model for a three-category response model. Extensions to responses with more than three categories follow the same pattern.

Before beginning with three-category responses, we reconsider the nature of the log-linear model for a two-category response. The model is

$$\mu = \begin{bmatrix} I & X \\ I & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{1}$$

This model lacks symmetry in the categories. The model matrix has an X submatrix corresponding to the first category of each response, but for the second category, the submatrix is zero. A model that treats the first and second categories on the same basis is the model

$$\mu = \begin{bmatrix} I & X & 0 \\ I & 0 & X \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \end{bmatrix}. \tag{2}$$

The important thing to note about model (2) is that it is equivalent to model (1). Any vector μ from model (1) can be obtained from (2) and vice versa. To see this, note that

$$C\left(\begin{bmatrix} I & X \\ I & 0 \end{bmatrix}\right) = C\left(\begin{bmatrix} I & X & 0 \\ I & 0 & X \end{bmatrix}\right).$$

Any two models with the same column space are equivalent models. Model (2) is a reparametrization of (1).

Given model (2), there is an obvious generalization to a three-category response. Simply take

$$\mu = \begin{bmatrix} I & X & 0 & 0 \\ I & 0 & X & 0 \\ I & 0 & 0 & X \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} \tag{3}$$

where $\mu = (\mu_{11}, \ldots, \mu_{t1}, \mu_{12}, \ldots, \mu_{t2}, \mu_{13}, \ldots, \mu_{t3})'$. As model (2) is equivalent to model (1), it is easily seen that model (3) is equivalent to

$$\mu = \begin{bmatrix} I & X & 0 \\ I & 0 & X \\ I & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Writing

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_t' \end{bmatrix}$$

gives model (3) as

$$\log(m_{ij}) = \mu_{ij} = \alpha_i + x_i'\gamma_j \tag{4}$$

where $i = 1, \ldots, t$ and $j = 1, 2, 3$. With this notation, it becomes a simple matter to consider models such as

$$\log(m_{ij}/m_{i\,j+1}) = x_i'(\gamma_j - \gamma_{j+1}), \qquad j = 1, \ldots, J-1,$$

or

$$\log(m_{ij}/m_{iJ}) = x_i'(\gamma_j - \gamma_J), \qquad j = 1, \ldots, J-1,$$

as were discussed in Chapter 4. Note that if X contains a column of ones, model (4) includes a term for column effects; thus, both the row and column margins of the $t \times 3$ table are fixed.

11.6 Asymptotic Results

We now consider asymptotic results for log-linear models contained in Haberman (1977). The results are quite general. In particular, we will show that they give the standard asymptotics for log-linear models and we will show how they apply to logistic regression. In Section 11.1, no explicit discussion was given concerning the nature of the asymptotic theory used to justify the results on estimation and testing. If the expected count in each cell approaches infinity, the usual asymptotic theory applies. Unfortunately, in regression settings, this is often not appropriate. For regression problems, a more reasonable approach is to allow additional observations to have distinct values of the predictor variables rather than having them occur at values of the predictor variables that have already occurred. These additional observations with new predictors constitute new cells in the table, so the table itself is getting larger. We need to think in terms of the convergence of a sequence of models. Haberman's results do this. They apply when fitting log-linear models to many situations in which there are few observations relative to the number of cells in the table. In particular, they satisfy our need for asymptotic results appropriate to logistic regression. Frequently, they do not apply when testing a log-linear model against the saturated model with data from a large sparse multinomial distribution. For asymptotic results that apply to this case, see Koehler (1986). Other results on large sparse multinomials are contained in Koehler and Larntz (1980), Simonoff (1983, 1985, 1986), and Zelterman (1987). The mathematics in this section are the most sophisticated in the book (with the possible exception of Chapter 12).

Before discussing Haberman's asymptotic results, we set notation for a fixed sample size problem. Consider a log-linear model

$$\mu = X\beta \tag{1}$$

where $\mu = \log(m)$. To deal with the sampling scheme, assume that $C(Z) \subset C(X)$. *We are interested in the asymptotic distribution of those estimable functions of β that do not depend on the parameters that are forced into the model to deal with the sampling scheme.* In other words, we are interested in functions $\gamma'\mu = \gamma' X\beta$ for which $\gamma' Z = 0$. Moreover, to avoid trivial cases, we assume that $\gamma' X \neq 0$. The MLE of μ is denoted $\hat{\mu}$. The asymptotic variance of $\gamma'\hat{\mu}$ will be related to the function

$$\sigma^2(\gamma'\hat{\mu}) = \gamma' A(m) D^{-1}(m)\gamma$$

where $A(m) = X(X'D(m)X)^{-1}X'D(m)$. This function can be estimated by

$$\hat{\sigma}^2(\gamma'\hat{\mu}) = \gamma' A(\hat{m}) D^{-1}(\hat{m})\gamma\,.$$

Also of interest are the asymptotic distributions of the Pearson test statistic X^2 and the likelihood ratio test statistic G^2 for testing model (1) against a reduced model

$$\mu = W\delta \tag{2}$$

where $C(Z) \subset C(W) \subset C(X)$. Let $r = \text{rank}(X) - \text{rank}(Z)$ and $s = \text{rank}(W) - \text{rank}(Z)$.

The asymptotic results require one more concept. Let

$$A_z = Z(Z'D(m)Z)^{-1}Z'D(m)$$

and let $\mathcal{N}(A_z)$ be the null space of A_z, i.e.,

$$\mathcal{N}(A_z) = \{x | A_z x = 0\}.$$

If we write a vector in $C(X)$ as $x = (x_1, \ldots, x_q)'$, define d to be

$$d = \sup\left\{ |x_i| / \sqrt{x'D(m)x} \;:\; x \in \mathcal{N}(A_z) \cap C(X) \right\}. \tag{3}$$

To get asymptotic results, consider a sequence of log-linear models indexed by t. Thus, the log-linear models are

$$\mu_t = X_t \beta_t$$

where $\mu_t = \log(m_t)$ and $C(Z_t) \subset C(X_t)$. Our estimable functions of interest are $\gamma_t'\mu_t$, where $\gamma_t' Z_t = 0$. The MLE of $\gamma_t'\mu_t$ is $\gamma_t'\hat{\mu}_t$. Similarly,

$$\sigma^2(\gamma_t'\hat{\mu}_t) = \gamma_t' A_t(m_t) D^{-1}(m_t)\gamma_t$$

and

$$\hat{\sigma}^2(\gamma_t'\hat{\mu}_t) = \gamma_t' A_t(\hat{m}_t) D^{-1}(\hat{m}_t)\gamma_t \,.$$

The reduced model of interest in testing models is

$$\mu_t = W_t\delta_t$$

with $C(Z_t) \subset C(W_t) \subset C(X_t)$. The ranks are $r_t = \text{rank}(X_t) - \text{rank}(Z_t)$ and $s_t = \text{rank}(W_t) - \text{rank}(Z_t)$. Note that r_t is also the rank of $\mathcal{N}(A_{z_t}) \cap C(X_t)$, cf. Proposition 11.6.5. Finally,

$$d_t = \sup\left\{ |x_i| \Big/ \sqrt{x' D(m_t) x} \,:\, x \in \mathcal{N}(A_{z_t}) \cap C(X_t) \right\} .$$

Obviously, the standard asymptotic results will not hold for all sequences of log-linear models. There must be some restrictions on the models. The restriction involves d_t.

Theorem 11.6.1. If $r_t d_t \to 0$ as $t \to \infty$, then

(a) $[\gamma_t'\hat{\mu}_t - \gamma_t'\mu_t] \Big/ \sqrt{\sigma^2(\gamma_t'\hat{\mu}_t)} \xrightarrow{L} N(0,1)$,

(b) $\hat{\sigma}^2(\gamma_t'\hat{\mu}_t)/\sigma^2(\gamma_t'\hat{\mu}_t) \xrightarrow{P} 1$.

Corollary 11.6.2. If $r_t d_t \to 0$ as $t \to \infty$, then

$$[\gamma_t'\hat{\mu}_t - \gamma_t'\mu_t] \Big/ \sqrt{\hat{\sigma}^2(\gamma_t'\hat{\mu}_t)} \xrightarrow{L} N(0,1) \,.$$

Let G^2 and X^2 be the likelihood ratio and Pearson test statistics for testing model (2) against model (1).

Theorem 11.6.3. If $r_t d_t \to 0$ and $r_t - s_t \to f$ as $t \to \infty$ and if $\mu_t \in C(W_t)$ for $t \geq 0$, then

(a) $G_t^2 \xrightarrow{L} \chi^2(f)$,

(b) $X_t^2 \xrightarrow{L} \chi^2(f)$,

(c) $G_t^2 - X_t^2 \xrightarrow{P} 0$.

These results imply the usual asymptotic results, cf. Chapter 12. In the usual results, replace the index t by the sample size N. For all N, we have $Z_N = Z$, $W_N = W$, $X_N = X$, and $\gamma_N = \gamma$. Moreover, $r_N = r$,

$r_N - s_N = r - s$, and $m_N = Nm^*$, where m^* is defined as in Section 10.3. With these adjustments, Theorems 1 and 3 give the standard results. For the theorems to apply, we need $r_N d_N \to 0$. Because r_N is fixed, this is simply the condition that $d_N \to 0$. To see that $d_N \to 0$, use the Cauchy-Schwartz inequality. Let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ where the 1 is in the ith place. Thus, for any vector x, $x_i = e_i' x$.

Proposition 11.6.4. As $N \to \infty$, $d_N \to 0$.

Proof. By Cauchy-Schwartz, for $x \in C(X)$

$$\begin{aligned}(e_i'x)^2 &= ([e_i'D(1/\sqrt{m_N})][D(\sqrt{m_N})x])^2 \\ &\leq (e_i'D(1/m_N)e_i)(x'D(m_N)x)\,.\end{aligned}$$

By the definition of d_N,

$$\begin{aligned}d_N^2 &\leq \sup\{(e_i'x)^2/x'D(m_N)x \,:\, x \in C(X)\} \\ &\leq \max_i e_i'D(1/m_N)e_i \\ &= N^{-1}\max_i e_i'D(1/m^*)e_i\,.\end{aligned}$$

Because $\max_i e_i'D(1/m^*)e_i$ is a fixed positive constant, $d_N^2 \to 0$; thus, $d_N \to 0$. □

A primary use of these theorems is in their application to logistic regression models. In logistic regression, the model is $\eta = X\beta$, where $\eta = (\log(m_{11}/m_{21}), \ldots, \log(m_{1t}/m_{2t}))'$. The equivalent log-linear model is

$$\mu = \begin{bmatrix} I & X \\ I & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \equiv [Z, X_L] \begin{bmatrix} \alpha \\ \beta \end{bmatrix}\,.$$

As discussed in Section 1, rather than letting the m_{ij}'s get large (i.e., taking additional observations in existing cells), it is more realistic to let the number of cells get large while retaining the same basic structure in the logistic regression model. In other words, as the sample size increases, we add new rows to the design matrix while the expected counts in each cell are allowed to remain small.

Setting the notation for this case, μ_t is a $2t \times 1$ matrix,

$$\mu_t = \begin{bmatrix} I_t & X_t \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} \tag{4}$$

where I_t is a $t \times t$ identity matrix, 0 is a $t \times p$ matrix of zeros, and X_t is a $t \times p$ design matrix for the logistic regression model. We assume that for

t large enough, the matrix X_t has $\text{rank}(X_t) = p$. For testing (4) against a reduced model, write the reduced model as

$$\mu_t = \begin{bmatrix} I_t & X_{0t} \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \eta_t \end{bmatrix} \tag{5}$$

where $C(X_{0t}) \subset C(X_t)$ and $\text{rank}(X_{0t}) = p_0$. Note that the full model is a log-linear model where the rank of the design matrix is $t + p$ and $r_t = (t+p) - t = p$. Similarly, for the reduced model, $s_t = (t+p_0) - t = p_0$.

As a practical matter, we never really have a sequence of models. We have one model. Thus, t is fixed (it is the number of binomial populations in the logistic regression). To apply Theorem 11.6.1, we need $r_t d_t \to 0$; thus, if rd is small for our model, we use the approximation

$$\frac{\gamma'\hat{\mu} - \gamma'\mu}{\sqrt{\hat{\sigma}^2(\gamma'\hat{\mu})}} \sim N(0,1)\,.$$

To apply Theorem 11.6.3, we need $r_t d_t \to 0$ and $(r_t - s_t) \to f$. In our current setup, $r_t - s_t = p - p_0$, which is fixed; so, again, if rd is small and the reduced model is adequate, we use the approximations

$$G^2 \sim \chi^2(p - p_0)$$

and

$$X^2 \sim \chi^2(p - p_0)\,.$$

It remains for us to get a handle on what conditions are necessary to have $r_t d_t \to 0$. Before doing that, we comment on why lack of fit tests often work poorly for logistic regression. The standard test for lack of fit of a logistic regression model is to test a model against the saturated log-linear model. To simplify notation, we will use model (5) as the model to be tested, but now, instead of testing model (5) against a model with similar structure like model (4), we test it against the saturated model. The saturated model can be written

$$\mu_t = \begin{bmatrix} I_t & I_t \\ I_t & 0 \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix}\,.$$

The difference between the saturated model and model (4) is that X_t is replaced by I_t. Whereas $\text{rank}(X_t) = p$, we now have $\text{rank}(I_t) = t$. With the saturated model as the full model, we have $r_t = t$. For Theorem 11.6.3 to apply, we need $r_t d_t \to 0$ and, for some value f, $r_t - s_t \to f$. First, $r_t - s_t = t - p_0 \to \infty$, so there is no appropriate number of degrees of freedom for the asymptotic test. More importantly, because the condition $r_t d_t \to 0$ is the condition used to ensure that the model behaves well asymptotically and because the saturated model has $r_t d_t = t d_t$, for the saturated model to behave well asymptotically it must have $d_t \to 0$ very rapidly. Under standard sampling schemes, this does not happen and Theorem 11.6.3 does

not apply. This is not to say that the lack of fit test will always work poorly. If the expected counts in each cell are large, we do not need to appeal to the sequence of models argument and, thus, the usual asymptotic results give an approximate χ^2 distribution for the lack of fit test. However, if expected cell counts are not large, there is no reason to believe that the asymptotic lack of fit test will work well and experience indicates that it does not work well.

A condition under which $r_t d_t \to 0$ for a sequence of logistic regression models is that the logistic regression leverages a_{ii} all approach zero while the b_i's do not, cf. equation (11.1.2). The remainder of this section is devoted to mathematical details associated with this demonstration. It requires a facility with linear models comparable to that developed in Christensen (1996b). The primary result is finding an explicit form for $(A - A_z)D^{-1}(m)$.

With $r_t = p$ for all t, it suffices to show that $d_t \to 0$. To do this, we will characterize d for an arbitrary logistic regression model.

The expected cell counts are $m = (m_{11}, m_{21}, \ldots, m_{t1}, m_{12}, \ldots, m_{t2})'$. Write $m_j = (m_{1j}, \ldots, m_{tj})$ for $j = 1, 2$ so that $m' = (m_1', m_2')$. For a logistic regression model, write

$$W \equiv X_* = \begin{bmatrix} I & X \\ I & 0 \end{bmatrix},$$

so the symbol W is playing the role generally reserved for X in a log-linear model. Write $A = W(W'D(m)W)^{-1}W'D(m)$.

For logistic regression, the value of d is defined as the sup of a function of w over all w's in $\mathcal{N}(A_z) \cap C(W)$. First, we need to characterize $\mathcal{N}(A_z) \cap C(W)$.

Proposition 11.6.5. $\mathcal{N}(A_z) \cap C(W) = C(A - A_z)$.

Proof. A is a projection operator onto $C(W)$. In particular, for $w \in C(W)$, $Aw = w$ and $AA = A$. Similarly, A_z is a projection operator onto $C(Z)$. Moreover, $AA_z = A_z$.

If $w \in \mathcal{N}(A_z) \cap C(W)$, then $(A - A_z)w = Aw - A_z w = Aw = w$; thus, $w \in C(A - A_z)$.

If $w \in C(A - A_z)$, clearly $w \in C(W)$. We need to show that $w \in \mathcal{N}(A_z)$. The matrix $(A - A_z)$ is a projection operator, so $(A - A_z)w = w$ and, thus, $w = (A - A_z)w = Aw - A_z w = w - A_z w$, so $A_z w = 0$. □

We now examine the behavior of d. For $i = 1, \ldots, 2t$, let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ where the 1 is in the ith place. By (3),

$$d = \sup \left\{ \max_i |e_i' x| / \sqrt{x' D(m) x} \ : \ x \in C(A - A_z) \right\}.$$

Because $x \in C(A - A_z)$ can be written as $x = (A - A_z)b$ for some b, write

$$d = \sup \left\{ \max_i |e_i'(A - A_z)b| \Big/ \sqrt{b'(A - A_z)'D(m)(A - A_z)b} \ : \ \text{all } b \right\} .$$

Note that

$$\begin{aligned} & |e_i'(A - A_z)b| \\ & \quad = \quad |e_i'(A - A_z)(A - A_z)b| \\ & \quad = \quad |[e_i'(A - A_z)D^{-1}(\sqrt{m})][D(\sqrt{m})(A - A_z)b]| \\ & \quad \leq \quad \sqrt{e_i'(A - A_z)D^{-1}(m)(A - A_z)'e_i}\sqrt{b'(A - A_z)'D(m)(A - A_z)b} \end{aligned}$$

where the inequality is just the Cauchy-Schwartz inequality. It follows that

$$d \leq \max_i \sqrt{e_i'(A - A_z)D^{-1}(m)(A - A_z)'e_i} .$$

Proposition 11.6.6. $(A - A_z)D^{-1}(m)(A - A_z)' = (A - A_z)D^{-1}(m)$.

Proof. Using the definitions of A and A_z, the fact that $(A - A_z)Z = 0$, and that $AZ = Z$, so $Z'A' = Z'$, we find

$$\begin{aligned} & (A - A_z)D^{-1}(m)(A - A_z)' \\ & \quad = \quad (A - A_z)D^{-1}(m)D(m)[W(W'D(m)W)^{-1}W' - Z(Z'D(m)Z)^{-1}Z'] \\ & \quad = \quad (A - A_z)[W(W'D(m)W)^{-1}W' - Z(Z'D(m)Z)^{-1}Z'] \\ & \quad = \quad (A - A_z)[W(W'D(m)W)^{-1}W'] \\ & \quad = \quad A[W(W'D(m)W)^{-1}W'] - A_z[W(W'D(m)W)^{-1}W'] \\ & \quad = \quad [W(W'D(m)W)^{-1}W'] \\ & \qquad - Z(Z'D(m)Z)^{-1}Z'D(m)[W(W'D(m)W)^{-1}W'] \\ & \quad = \quad AD^{-1}(m) - Z(Z'D(m)Z)^{-1}Z'A' \\ & \quad = \quad AD^{-1}(m) - Z(Z'D(m)Z)^{-1}Z' \\ & \quad = \quad AD^{-1}(m) - A_zD^{-1}(m) \\ & \quad = \quad (A - A_z)D^{-1}(m). \end{aligned}$$

□

Using Proposition 11.6.6, we now have

$$d \leq \max_i \sqrt{e_i'(A - A_z)D^{-1}(m)e_i} . \tag{6}$$

So far, we have not used the logistic regression structure of the log-linear model. We now use the special structure of logistic regression to further characterize inequality (6).

Proposition 11.6.7.

$$A_z = \begin{bmatrix} D(m_1)D^{-1}(m_1+m_2) & D(m_2)D^{-1}(m_1+m_2) \\ D(m_1)D^{-1}(m_1+m_2) & D(m_2)D^{-1}(m_1+m_2) \end{bmatrix}.$$

Proof. This follows immediately from Lemma 11.1.6, the definitions of Z and A_z, and the fact that

$$D(m) = \begin{bmatrix} D(m_1) & 0 \\ 0 & D(m_2) \end{bmatrix}. \qquad \square$$

To simplify notation, for vectors $v = (v_1, \ldots, v_q)'$ and $u = (u_1, \ldots, u_q)'$, let $vu = (v_1u_1, \ldots, v_qu_q)'$ and $v/u = (v_1/u_1, \ldots, v_q/u_q)'$. As in (11.1.2), $b \equiv m_1m_2/(m_1+m_2)$.

Proposition 11.6.8.

$$\begin{aligned} &A - A_z \\ &= \begin{bmatrix} D(m_2/(m_1+m_2))X \\ -D(m_1/(m_1+m_2))X \end{bmatrix} [X'D(b)X]^{-1} [X'D(b)\,,\, -X'D(b)]\,. \end{aligned}$$

Proof. The perpendicular projection operator onto $C(D(\sqrt{m})W)$ is $D(\sqrt{m})AD^{-1}(\sqrt{m}) \equiv M_w$ and the perpendicular projection operator onto $C\left(D(\sqrt{m})Z\right)$ is $D(\sqrt{m})A_zD^{-1}(\sqrt{m}) \equiv M_z$. It follows that

$$\begin{aligned} M_w - M_z &= D(\sqrt{m})AD^{-1}(\sqrt{m}) - D(\sqrt{m})A_zD^{-1}(\sqrt{m}) \\ &= D(\sqrt{m})(A - A_z)D^{-1}(\sqrt{m})\,. \end{aligned}$$

If we can find $M_w - M_z$, then $A - A_z = D^{-1}(\sqrt{m})(M_w - M_z)D(\sqrt{m})$.

The matrix $M_w - M_z$ is the perpendicular projection operator onto

$$C\left((I - M_z)D(\sqrt{m})\begin{bmatrix} X \\ 0 \end{bmatrix}\right),$$

cf. Christensen (1996b, Sections 9.1, 9.2).

$$\begin{aligned} &(I - M_z)D(\sqrt{m})\begin{bmatrix} X \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} D(\sqrt{m_1})X \\ 0 \end{bmatrix} - \begin{bmatrix} D(m_1\sqrt{m_1}/(m_1+m_2))X \\ D(m_1\sqrt{m_2}/(m_1+m_2))X \end{bmatrix} \\ &= \begin{bmatrix} D(\sqrt{m_1}m_2/(m_1+m_2))X \\ -D(\sqrt{m_2}m_1/(m_1+m_2))X \end{bmatrix}. \end{aligned}$$

Multiplying out to get the perpendicular projection operator and simplifying gives the result. □

Finally, the main result is

Proposition 11.6.9.

$$(A - A_z)D^{-1}(m) = \begin{bmatrix} D(m_2/(m_1+m_2))X \\ -D(m_1/(m_1+m_2))X \end{bmatrix} [X'D(b)X]^{-1} \times [X'D(m_2/(m_1+m_2)), -X'D(m_1/(m_1+m_2))]$$

Proof. Multiply $A - A_z$ by $D^{-1}(m)$. □

We can now examine the exact nature of $e_i'(A - A_z)D^{-1}(m)e_i$. Write

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_t' \end{bmatrix};$$

then for $i = 1, \ldots, t$, let $j = i$ and

$$e_i'(A - A_z)D^{-1}(m)e_i = [m_{j2}/(m_{j1} + m_{j2})]^2 x_j'[X'D(b)X]^{-1}x_j,$$

for $i = t+1, \ldots, 2t$, let $j = i - t$ and

$$e_i'(A - A_z)D^{-1}(m)e_i = [m_{j1}/(m_{j1} + m_{j2})]^2 x_j'[X'D(b)X]^{-1}x_j.$$

If these terms approach zero for a sequence of logistic regression models, then inequality (6) implies that Theorems 11.6.1 and 11.6.3 hold. In practice, if these terms are small for all i, then Theorems 11.6.1 and 11.6.3 should provide reasonable approximate distributions. The key (cf. Exercise 11.1) is that the X matrix needs to have the property that $x_i'[X'D(b)X]^{-1}x_i$ is small for all i.

This condition can also be related to linear model theory. If the elements of m_1 and m_2 are bounded above zero, then the terms

$$x_i'[X'D(b)X]^{-1}x_i$$

will converge to zero if and only if the terms $x_i'[X'X]^{-1}x_i$ converge to zero. The condition that the terms $x_i'[X'X]^{-1}x_i$ all converge to zero is known as Huber's condition. This is the condition assumed by Arnold (1981) to show that the usual distributions hold asymptotically for linear models with non-normal independent errors. Similarly, if Huber's condition holds

for a logistic regression model, then the usual asymptotic results for logistic regression hold. Note that Huber's condition is sufficient to imply that asymptotic results hold; it is not a necessary condition.

EXERCISE 11.1. Show for logistic regression that $d_t \to 0$ as the maximum leverage goes to zero.

11.7 Discrimination, Allocation, and Retrospective Data

The Chapman data of Section 4.1 are the result of a *prospective* study; a large number of people were sampled and they were classified by whether they had experienced a coronary incident and by their values on the variables age, systolic blood pressure, diastolic blood pressure, cholesterol, height, and weight. Only 26 of the 200 people had coronary incidents, so most of the information in the data is about people who did not have coronaries.

Retrospective studies are commonly used to examine events that are relatively rare, like coronary incidents. They address the problem of having a sample that contains few observations on the rare event. Consider a response variable with three levels: no incident, mild coronary incident, and severe coronary incident. One might take a sample of 125 people with no incidents, a sample of 40 people with mild incidents, and a sample of 35 people with severe coronary incidents. Thus, the sample sizes in the rare event categories are fixed by design. Once again, the case variables age, systolic blood pressure, diastolic blood pressure, cholesterol, height, and weight can be measured for each of the 200 individuals. When the response categories are also the populations sampled, it is easier to get substantial numbers of observations in each response category. Prospective and retrospective studies were discussed earlier at the beginning of Chapter 4.

While all of the case variables discussed above are really continuous, there are only a finite number of values that one could actually measure for any of the variables. For example, one often measures age in integer values of years and height in integer values of inches. Moreover, there are upper bounds on these values. Similar limitations based on the accuracy of the measuring instruments exist for all continuous variables. Thus, there are only a finite number of combinations of the case variables that can be considered. Call this very large but finite number S. The retrospective study described above yields a $3 \times S$ table in which each of the three rows is an independent multinomial sample. We wish to model these multinomials so that we can explain the data and, perhaps more importantly, predict the population into which a new case would fall when only the information on the case variables is available. The modeling problem can be thought of

as *discriminating* among the three populations. The prediction problem is one of *allocating* new cases to the appropriate population.

Consider the allocation problem in more detail. Write the case variables as a vector $x_i = (Ag_i, S_i, D_i, Ch_i, H_i, W_i)'$, $i = 1, \ldots, S$. The value p_{hi} is the probability under population h of being in the category determined by x_i. Here, $h = 1, 2, 3$ and $i = 1, \ldots, S$. Write

$$f(x_i|h) = p_{hi}\,.$$

The function $f(x_i|h)$ is just the discrete density (probability mass function) of population h. The value of h is a parameter. Given a new case with known case variables x in one of the S observable categories, we can view $f(x|h)$ as a likelihood function. The maximum likelihood allocation rule assigns the new case x to the population h that has the largest value for the likelihood.

The only problem with this procedure is that the probabilities $f(x|h)$ are not known. They have to be estimated from the data. S is typically extremely large, so there are typically many more parameters (probabilities) to be estimated than observations with which to estimate them. Some sort of additional assumptions must be made in order to proceed.

One way to proceed is to abandon the fact that the observed data are discrete and assume a continuous density $f(x|h)$ as a model for the observations. If the underlying but unobservable case variables are continuous, this is extremely reasonable. In fact, it is so reasonable that people often overlook the fact that it is an approximation to what is properly a discrete distribution for the observations. The problem is not in approximating a discrete distribution with a continuous distribution but in finding a continuous distribution that provides an appropriate model. Traditionally, the case variables have been modeled with a multivariate normal distribution. For multivariate normals, estimating the mean vector and the covariance matrix for each population leads to natural estimates for the $f(x|h)$'s. This approach to discrimination and allocation is originally due to R. A. Fisher (1936). More recently, nonparametric density estimation has been used to model the distributions, cf. Seber (1984, Section 6.5).

Rather than invoking continuity, another way to proceed is to cut down the number of categories to a manageable size. Rather than using all S categories, one can restrict attention to the x values that were actually observed. In our hypothetical example, if all the x_i's are distinct, this restriction yields a 3×200 table. The x_i's are frequently distinct when any of the case variables are continuous, but if they are not distinct, it simply reduces the number of columns in the table. We will assume that the x_i's are distinct.

The original sampling scheme for the $3 \times S$ table was product-multinomial in the rows and the standard method of analysis, as illustrated in Section 4.7, also treats the 3×200 table as product-multinomial in the rows. Alas, restricting the table invalidates this product-multinomial sampling scheme for the reduced table. For example, under product-multinomial sampling,

there is a positive probability of getting zeros for all three of the counts in a given column. We are using only the x_i's that are actually observed, so every column must have at least one observation in it.

There are two ways to analyze the data in the 3×200 table. The standard method illustrated in Section 4.7 is both a *partial likelihood* analysis and an *extended likelihood* analysis. Alternatively, one can perform a *conditional likelihood* analysis to obtain information on some parameters.

Partial likelihood analysis depends on having a likelihood that can be factored into the product of two terms. One term, the partial likelihood, must involve all of the parameters of interest and only those parameters. The second term involves only nuisance parameters. Without loss of generality, assume that the actual observations occur in the first 200 categories. For each population h, let $p_h = (p_{h1}, \ldots, p_{hS})'$ be the probability vector and let $N_h \equiv n_{h\cdot}$ be the sample size. The likelihood is

$$\begin{aligned} L(p_1, p_2, p_3) &= \prod_{h=1}^{3} \prod_{i=1}^{S} p_{hi}^{n_{hi}} \\ &= \left\{ \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \right\} \left\{ \prod_{h=1}^{3} \prod_{i=201}^{S} p_{hi}^{n_{hi}} \right\} \\ &= \left\{ \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \right\} \end{aligned}$$

where the last equality holds because for $i > 200$, $n_{hi} = 0$ and any log-linear model has $p_{hi} > 0$ for all h and i. With all the zero counts, the likelihood cannot be maximized subject to the condition of positive cell probabilities. However, this function is also the partial likelihood involving only the parameters p_{hi}, $h = 1, 2, 3,\ i = 1, \ldots, 200$. As such, it can be maximized.

Technically, write

$$L(p_1, p_2, p_3) = \left\{ \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \right\} \cdot \Gamma(p_{hi} : h = 1, 2, 3;\ i = 201, \ldots, S)$$

where

$$\Gamma(p_{hi} : h = 1, 2, 3;\ i = 201, \ldots, S) \equiv 1.$$

We have factorized the likelihood appropriately, so the partial likelihood for $p_{hi} : h = 1, 2, 3,\ i = 1, \ldots, 200$ is

$$\prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}}.$$

Obviously, the log of the partial likelihood is

$$\sum_{h=1}^{3}\sum_{i=1}^{200} n_{hi}\log(p_{hi})\,.$$

We now incorporate a log-linear model into the analysis. Assume a typical multinomial response model for the parameters

$$m_{hi} = N_h p_{hi}$$

that consists of

$$\log(m_{hi}) = \alpha_i + x_i'\gamma_h \tag{1}$$

for all h and i, cf. model (11.5.4). Dropping a constant that depends only on the N_h's, the log-likelihood becomes

$$\begin{aligned}\ell(\gamma_1,\gamma_2,\gamma_3,\alpha_i, i=1,\ldots,S) &= \sum_{h=1}^{3}\sum_{i=1}^{S} n_{hi}\log(m_{hi})\\ &= \sum_{h=1}^{3}\sum_{i=1}^{S} n_{hi}\left[\alpha_i + x_i'\gamma_h\right]\\ &= \sum_{h=1}^{3}\sum_{i=1}^{200} n_{hi}\left[\alpha_i + x_i'\gamma_h\right]\end{aligned}$$

where, again, the last equality follows from the fact that $n_{hi} = 0$ for $i = 201,\ldots,S$. In the $3 \times S$ table, the row totals are fixed by the product-multinomial sampling scheme. The standard analysis of the 3×200 table also treats the rows as product-multinomials, so the row totals are fixed. Fixing the row totals requires the inclusion of main effects for rows in the log-linear model. These can be incorporated into the $x_i'\gamma_h$ terms. Write $x_i' = (1, x_{i2},\ldots,x_{ip})$, where the case variables are $x_{i2},\ldots,x_{ip}$. Then with $\gamma_h = (\gamma_{h1},\ldots,\gamma_{hp})'$, the intercept parameters γ_{h1} are the row main effects.

For a partial likelihood analysis, observe that the log-likelihood is the sum of two terms, one of which depends on the parameters of interest $\gamma_1,\gamma_2,\gamma_3,\alpha_i,\ i = 1,\ldots,200$, and another, which in this case is identically equal to zero, that depends only on $\alpha_i, i = 201,\ldots,S$. (The second function is identically zero, so it depends on anything we want it to depend on.) A partial likelihood analysis then obtains estimates of $\gamma_1,\gamma_2,\gamma_3,\alpha_i, i = 1,\ldots,200$, by maximizing the term that involves only those parameters. Of course, the only difference between the log partial likelihood and the log-likelihood is that the log partial likelihood is considered as a function of fewer parameters. In particular, the log partial likelihood is exactly the same as the log-likelihood for the 3×200 table under product-multinomial sampling. Thus, the MLEs from the standard analysis are maximum partial likelihood estimates for the full $3 \times S$ table.

Another productive way to use the log-likelihood is to consider *extended maximum likelihood estimates*, cf. Haberman (1974a, p. 402). An estimate $\hat{m}$ is an extended maximum likelihood estimate if the log-likelihood $\ell(m)$ converges to its supremum as m converges to $\hat{m}$. In this setup, the usual MLEs from the 3×200 table are extended MLEs. The log-likelihood function only depends on $\gamma_1, \gamma_2, \gamma_3, \alpha_i$, $i = 1, \ldots, 200$, so the reduced table MLEs together with $\hat{p}_{hi} = 0$ for $h = 1, 2, 3$, $i = 201, \ldots, S$ maximize the full table log-likelihood function subject to the constraints $\hat{p}_{h\cdot} = 1$. The only problem is that log-linear models do not allow $\hat{p}_{hi} = 0$ for any h, i. Allowing extended MLEs removes the problem.

Whether the justification is partial likelihood or extended likelihood, we arrive at an analysis based on the MLEs for the 3×200 table with product-multinomial sampling in the rows. Details of the analysis are given in the next subsection.

The *conditional likelihood analysis* simply defines the likelihood in terms of the conditional distribution of the 3×200 table given that these were the only 200 columns of the $3 \times S$ table that were observed. The conditional likelihood of the 3×200 table is

$$\prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{n_{hi}} \Big/ \sum_{r} \prod_{h=1}^{3} \prod_{i=1}^{200} p_{hi}^{r_{hi}} \tag{2}$$

where the sum is over all $3 \times S$ tables of counts $r = (r_{11}, \ldots, r_{3S})'$ with

$$r_{h\cdot} = n_{h\cdot} = N_h, \qquad h = 1, 2, 3\,,$$

$$r_{\cdot i} = n_{\cdot i} = 1, \qquad i = 1, \ldots, 200\,,$$

$$r_{\cdot i} = 0, \qquad i = 201, \ldots, S\,.$$

It is not difficult to see that the conditional likelihood does not depend on the α_i's or the intercept terms γ_{h1}, cf. Exercise 11.8.4. Thus, any inferences that require estimates of these quantities cannot be made using the conditional likelihood approach. In particular, it will be seen in the next subsection that allocation of observations depends on the vectors γ_h, including the components γ_{h1}. Thus, *the conditional likelihood approach cannot be used for allocation.*

The key early paper on logistic discrimination was written by Anderson (1972). Anderson and Blair (1982) clarified several aspects of the theory and introduced another basis for analysis: penalized maximum likelihood. Some other relevant works are Farewell (1979), Prentice and Pyke (1979), and Breslow and Day (1980).

The Partial Likelihood Analysis

We have a vector of allocation variables $x_i' = (x_{i1}, \ldots, x_{ip})$ that are observed on each of t individuals. Thus, far in the section, we have always

used $t = 200$, but the conclusions hold for any value of t. In addition to observing the x_i's, we know to which of the three populations each individual belongs. Our goal is to use the information on these t individuals in order to allocate future individuals into an appropriate population.

To do this, we set up a model similar to the standard multinomial response model. Let

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_t' \end{bmatrix}.$$

Our data are

$$n = (n_{11}, \ldots, n_{1t}, n_{21}, \ldots, n_{2t}, n_{31}, \ldots, n_{3t})'$$

where $n_{hi} = 1$ if the ith case belongs to population h and $n_{hi} = 0$ otherwise. For a prospective multinomial response model, the $3 \times t$ table either is or can be considered to be the result of taking t independent trinomial samples and can be analyzed by standard logistic regression. With the retrospective sampling appropriate for discrimination problems, the sampling scheme is the result of taking three independent multinomial samples each with S categories where S is large and unknown. As discussed above, for the purpose of estimation the samples can be treated as three independent multinomials with t categories. *The standard prospective approach assumes that column totals of the $3 \times t$ table are fixed by the sampling scheme, whereas the retrospective (discrimination) approach assumes that the row totals are fixed by the sampling scheme.*

With a $3 \times t$ table in which the row totals are fixed, indicator variables must be included in the model to ensure that the estimated row totals equal the observed row totals. This is accomplished by requiring that the X matrix include a column of 1s (or its equivalent). In other words, for logistic models, the sampling scheme requires that models for discrimination data include intercepts. It is a common practice to include an intercept in multinomial response models, so the fact that an intercept is *required* for retrospective data is easily overlooked.

The log-linear model $\log(m_{hi}) = \alpha_i + x_i'\gamma_h$ for the $3 \times t$ table is written in matrix form as

$$\log(m) = \mu = \begin{bmatrix} I_t & X & 0 & 0 \\ I_t & 0 & X & 0 \\ I_t & 0 & 0 & X \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}.$$

This model is of exactly the same form as a standard multinomial response model and is fit in exactly the same way. The difference is in the interpretation of the underlying probabilities. In prospective sampling, the multinomial expected cell counts are $m_{hi} = n_{\cdot i} p_{hi}$ where $p_{\cdot i} = 1$. For retrospective

sampling,

$$m_{hi} = n_{h\cdot} p_{hi} = N_h p_{hi}$$

where

$$p_{h\cdot} = 1\,.$$

In particular,

$$\log(p_{hi}) = \alpha_i + x_i'\gamma_h - \log(n_{h\cdot})\,.$$

The MLE of $\log(p_{hi})$, under the device of treating the sampling as product-multinomial in the rows of the $3 \times t$ table, is both the maximum partial likelihood estimate and an extended maximum likelihood estimate of $\log(p_{hi})$ in the full $3 \times S$ table.

The maximum likelihood allocation method applied to the observed data assigns case i to the population h with $\log(p_{hi}) = \max_k\{\log(p_{ki})\}$. The estimated maximum likelihood allocation method assigns i to the population h with

$$\log(\hat{p}_{hi}) = \max_k\{\log(\hat{p}_{ki})\}. \tag{3}$$

Note that equation (3) is equivalent to

$$\hat{\alpha}_i + x_i'\hat{\gamma}_h - \log(n_{h\cdot}) = \max_k \{\hat{\alpha}_i + x_i'\hat{\gamma}_k - \log(n_{k\cdot})\}$$

which is equivalent to

$$x_i'\hat{\gamma}_h - \log(n_{h\cdot}) = \max_k \{x_i'\hat{\gamma}_k - \log(n_{k\cdot})\}\ . \tag{4}$$

Equation (4) does not depend on the $\hat{\alpha}_i$'s; thus, the allocation procedure depends on the individual only through the value of x_i.

Equation (4) can also be used as the basis for classifying new cases from an unknown population into one of the three possible populations. If the new case has observation vector x, the estimated maximum likelihood allocation rule is to classify the new case into population h if

$$x'\hat{\gamma}_h - \log(n_{h\cdot}) = \max_k \{x'\hat{\gamma}_k - \log(n_{k\cdot})\}\ .$$

This result depends on the fact that the allocation rule is really a function of the likelihood ratios of the various populations. The likelihood ratios depend only on x, the γ's, and the $\log(n_{h\cdot})$'s. All of these are either known or can be estimated. For a given value of x, the corresponding value of α does not enter into the analysis. Moreover, as illustrated in Section 4.7, if the likelihood ratios can be estimated, the posterior probabilities can also be estimated.

To evaluate how well the model discriminates between populations, check to see how often the cases in the data are allocated to the correct population. In other words, when case i is really in population h, see how often

the probability that case i comes from population h is larger than the probabilities that case i comes from any of the other populations. Because the evaluation is carried out on the same data that generated the estimates of the p_{hi}'s, the results of the evaluation will be biased in favor of the discrimination method; i.e., the method will look better than it really is. See Section 4.7 for more discussion of this problem.

11.8 Exercises

EXERCISE 11.8.1. Analyze the data of Exercise 8.4.1 as a logistic regression with nodal involvement as the response. Include the investigation of higher-order interactions in your analysis. The original investigator was particularly interested in whether acid was a valuable predictor of nodal involvement.

EXERCISE 11.8.2. *Asymptotic Inference for the* $LD(50)$.
In Exercise 4.8.9, models and methods for estimating the $LD(50)$ were discussed. Use the delta method of Exercise 10.8.4 to obtain an asymptotic standard error for the $LD(50)$. Using the data of Exercise 4.8.11, give a 99% confidence interval for the $LD(50)$.

EXERCISE 11.8.3. *Fieller's Method for the* $LD(50)$.
Fieller's method is an alternative to the delta method for obtaining an asymptotic confidence interval for the $LD(50)$, cf. Exercise 11.8.2. Fieller's method is thought to be less sensitive to the high correlation that is typically present between $\hat{\alpha}$ and $\hat{\beta}$. From standard results, one can obtain the estimated asymptotic variance and covariance for $\hat{\alpha}$ and $\hat{\beta}$; thus, for any fixed but unknown value w, an asymptotic standard error for $\hat{\alpha} + \hat{\beta}w$ is readily available as a function of w. Denote this standard error by $\hat{\sigma}(w)$. If $\alpha + \beta w$ is some known value Q, a 99% confidence region for w can be obtained from

$$\begin{aligned} .99 &= \Pr\left(-2.5758 \leq \frac{(\hat{\alpha} + \hat{\beta}w) - Q}{\hat{\sigma}(w)} \leq 2.5758\right) \\ &= \Pr\left((\hat{\alpha} + \hat{\beta}w - Q)^2 - 2.5758^2\,\hat{\sigma}^2(w) \leq 0\right). \end{aligned}$$

The 99% confidence region consists of all values of w that satisfy

$$(\hat{\alpha} + \hat{\beta}w - Q)^2 - 2.5758^2\hat{\sigma}^2(w) \leq 0.$$

Show how to use the quadratic formula to find the end points of the region. Under what conditions is the confidence region an interval? What other possibilities exist? How does this result apply to estimating the $LD(50)$?

Using the data of Exercise 4.6.11, give a 99% confidence interval for the $LD(50)$.

EXERCISE 11.8.4. Show that the conditional likelihood given by equation (11.7.1) and display (11.7.2) does not depend on the α_i's or the intercept terms γ_{h1}. Here, $x_i' = (1, x_{i2}, \ldots, x_{ik})$ and $\gamma = (\gamma_{h1}, \ldots, \gamma_{hk})'$.

EXERCISE 11.8.5. Use the delta method of Exercise 10.8.4, the logistic transform, and (11.1.3) to show that

$$\frac{\hat{p}_{ij} - p_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})\hat{a}_{ii}/N_i}} \sim N(0, 1)\,,$$

where $N_i = n_{i1} + n_{i2}$.

12

Maximum Likelihood Theory for Log-Linear Models

This chapter presents the basic theoretical results of fitting log-linear models by maximum likelihood. The level of mathematical sophistication is considerably higher than in the rest of the book. The presentation assumes knowledge of advanced calculus, mathematical statistics, and large sample theory. Although the results in this chapter are proven in a different manner than for regular linear models, the results themselves are quite similar in nature. The common linear structure of the two techniques leads to the well-known analogies between them. A familiarity with log-linear models at the level of, say Fienberg (1980), is assumed.

In order to simplify proofs, the $o(\cdot)$, $O(\cdot)$, $o_p(\cdot)$, $O_p(\cdot)$ notations have been used extensively. See Bishop, Fienberg, and Holland (1975) for a detailed discussion of these.

Section 1 introduces notation and recalls some analytic results. Section 2 discusses finite sample properties of maximum likelihood estimators. Section 3 treats asymptotic results. Section 4 examines how the theory applies to weighted least squares, obtaining variance estimates, and logit and multinomial response models. Section 5 contains two proofs that are more involved than the rest of the chapter.

12.1 Notation

All vectors are considered as column vectors unless otherwise stated.

We will frequently apply the same real valued function to each element of a vector. Let x be a vector in $\mathbf{R}^s$ and f a function from $\mathbf{R}$ to $\mathbf{R}$, then the function from $\mathbf{R}^s$ to $\mathbf{R}^s$ that maps x elementwise is denoted

$$f(x) = \big(f(x_1), \ldots, f(x_s)\big)'.$$

The most common choice of f will be the log function; thus, $\log(x) = (\log x_1, \ldots, \log x_s)'$.

A diagonal matrix with the values of the vector x on the diagonal will be written $D(x)$. As usual, an $s \times 1$ vector of ones is written J_s with the subscript dropped when the dimension is clear.

Proofs using the $o_p(\cdot)$, $O_p(\cdot)$ notation are usually divided into an analytic argument and a stochastic argument. To reduce the length of the discussion, these arguments have frequently been run together. Therefore, if $f(x) = o(g(x))$ and X_N is a sequence of random variables, we write $f(X_N) = o(g(X_N))$. Two frequently used properties are (a) $o(O_p(N^{-\alpha})) = o_p(N^{-\alpha})$ for any $\alpha > 0$ and (b) $o(o_p(1)) = o_p(1)$. For example, if $g(X_N) = o_p(1)$, we have $f(X_N) = o(g(X_N)) = o(o_p(1)) = o_p(1)$.

If F is a function from $\mathbf{R}^s$ into $\mathbf{R}^t$ with $F(x) = (f_1(x), \ldots, f_t(x))'$, then the derivative of F at c is the $t \times s$ matrix of partial derivatives,

$$dF(c) = [\partial f_i / \partial x_j|_{x=c}].$$

If g maps $\mathbf{R}^s$ into $\mathbf{R}$, then $dg(c)$ is a $1 \times s$ *row* vector. The second derivative matrix of g at c is

$$d^2g(c) = d[(dg(x))'|_{x=c}] = \big[\partial^2 g / \partial x_i \partial x_j|_{x=c}\big],$$

which is an $s \times s$ matrix. Taylor's theorem can be written

$$g(x) = g(c) + dg(c)(x-c) + (x-c)'\big[d^2g(c)\big](x-c)/2 + o\big(\|x-c\|^2\big),$$

where $\|x-c\|^2 = (x-c)'(x-c)$. Critical points are points c, where $dg(c) = 0$. The chain rule can be written as a matrix product: If $f : \mathbf{R}^s \to \mathbf{R}^t$ and $g : \mathbf{R}^t \to \mathbf{R}^u$, then $d(g \circ f)(c) = dg(f(c))df(c)$.

12.2 Fixed Sample Size Properties

Consider a table of counts with q cells. The observations are denoted $n = (n_1, \ldots, n_q)'$. The n_i's are assumed to be the result of Poisson, multinomial, or product-multinomial sampling. Let $\mathrm{E}(n) = m$ and let X be a known $q \times p$ matrix. The log-linear model is $\log(m) \equiv \mu = Xb$ for some vector b. The log-linear model is simply the requirement that $\mu \in C(X)$. Unless otherwise indicated, X will be assumed to have rank p.

If the observations n_i in the cells are independent Poisson(m_i) random variables, the likelihood function is

$$\prod_{i=1}^{q} \left[e^{-m_i} m_i^{n_i} / n_i! \right]. \tag{1}$$

The log-likelihood is

$$\begin{aligned} \ell^{\mathbf{P}}(n, \mu) &= \sum_{i=1}^{q} [-m_i + n_i \log m_i - \log(n_i!)] \\ &= \sum_{i=1}^{q} [-e^{\mu_i} + n_i \mu_i - \log(n_i!)] \\ &= \sum_{i=1}^{q} -e^{\mu_i} + n'\mu - \sum_{i=1}^{q} \log(n_i!). \end{aligned} \tag{2}$$

If the log-linear model holds, $\mu = Xb$, so

$$\ell^{\mathbf{P}}(n, \mu) = n'Xb - \sum_{i=1}^{q} e^{\mu_i} - \sum_{i=1}^{q} \log(n_i!).$$

Since the distribution of n is in the exponential family, $X'n$ is a complete sufficient statistic.

If the observations come from a product-multinomial sampling scheme, certain of the margins are fixed. Assume that there are r independent multinomials. (If $r = 1$, the sampling scheme is a simple multinomial.) Partition $\{1, \ldots, q\}$ into r sets $Q_1, \ldots, Q_r$, each set containing the indices for one of the multinomials. For $i = 1, \ldots, q$, $j = 1, \ldots, r$, let x_j be a vector with ith row, $x_{ij} = \delta_{Q_j}(i)$ where $\delta_{Q_j}(i)$ is one if $i \in Q_j$ and zero otherwise. Thus, x_j is a column of dummy variables indicating the jth population. By the sampling scheme, $n'x_j$ is fixed for $j = 1, \ldots, r$. In particular, $n'x_j = m'x_j = N_j$, the sample size for the jth multinomial. It will be convenient to combine the vectors x_j into a matrix, say $X_0 = [x_1, \ldots, x_r]$.

With product-multinomial sampling, there are two restrictions on the parameters: (a) $\mu \in C(X)$, and (b) $n'X_0 = m'X_0$. Estimates of the parameters also need to satisfy these conditions. If we assume that $C(X_0) \subset C(X)$, we will see that the MLE of m, based only on condition (a), will automatically satisfy condition (b).

We will assume throughout that $C(X_0) \subset C(X)$. For Poisson sampling, X_0 can be taken as a matrix of zeros. We also need the assumption that $J_q \in C(X)$. For product-multinomial sampling, $J_q \in C(X_0)$, so this is not a new restriction. For Poisson sampling, we are requiring that an overall mean (or its equivalent) be fitted.

Recall that the probability of an occurrence in the ith cell under product-multinomial sampling is m_i/N_j, where $i \in Q_j$, so the likelihood function

is

$$\prod_{k=1}^{r}\left[\left(N_k!\Big/\prod_{i\in Q_k} n_i!\right)\prod_{i\in Q_k}\left(\frac{m_i}{N_k}\right)^{n_i}\right]. \tag{3}$$

Let $\ell^{\mathbf{m}}(n,\mu)$ be the log of (3). For $\ell^{\mathbf{m}}(n,\mu)$ to be a log-likelihood, μ must have the property that $n'X_0 = m'X_0$, where $m = e^{\mu}$. In fact, $\ell^{\mathbf{m}}(n,\mu)$ is only defined for such μ. In particular, the maximum likelihood estimate of μ, under product-multinomial sampling, must be a value of μ for which $\ell^{\mathbf{m}}(n,\mu)$ is defined. We will expand the domain of $\ell^{\mathbf{m}}(n,\mu)$ to include all real vectors μ. For a log-linear model $\mu \in C(X)$ with $C(X_0) \subset C(X)$, we will find the maximum of $\ell^{\mathbf{m}}(n,\mu)$ without reference to the condition $n'X_0 = m'X_0$. We will then observe that the value of μ that maximizes $\ell^{\mathbf{m}}(n,\mu)$ also satisfies the condition $n'X_0 = m'X_0$. This value of μ must be the maximum likelihood estimate.

We now proceed to expand the domain of $\ell^{\mathbf{m}}(n,\mu)$. Since $\sum_{i\in Q_k} n_i = \sum_{i\in Q_k} m_i = N_k$, (3) can be rewritten as

$$\prod_{k=1}^{r}\left[N_k! e^{N_k} N_k^{-N_k}\prod_{i\in Q_k}\left(m_i^{n_i} e^{-m_i}/n_i!\right)\right]$$

or

$$\left[\prod_{k=1}^{r} N_k! e^{N_k} N_k^{-N_k}\right]\left[\prod_{i=1}^{q} e^{-m_i} m_i^{n_i}/n_i!\right]. \tag{4}$$

The second term in (4) is exactly (1). The first term depends only on the N_k's. If we write $a(N_1,\ldots,N_r)$ as the log of the first term we can write the log-likelihood for product-multinomial sampling as

$$\ell^{\mathbf{m}}(n,\mu) = a(N_1,\ldots,N_r) + \ell^{\mathbf{p}}(n,\mu).$$

Since $\ell^{\mathbf{m}}(n,\mu)$ is defined only for values of μ satisfying $n'X_0 = m'X_0$, this relationship holds only for such values of μ. However, the relationship can be used to define the function $\ell^{\mathbf{m}}(n,\mu)$ for all values of μ, because $\ell^{\mathbf{p}}(n,\mu)$ is defined for all values of μ. Since the difference of $\ell^{\mathbf{p}}(n,\mu)$ and $\ell^{\mathbf{m}}(n,\mu)$ does not depend on μ, the maximums of the two functions, with respect to μ, will occur at the same place.

Rather than using either $\ell^{\mathbf{p}}(n,\mu)$ or $\ell^{\mathbf{m}}(n,\mu)$, remove the term in (2) that does not depend on μ to define

$$\ell(n,\mu) = n'\mu - \sum_{i=1}^{q} e^{\mu_i} = n'\mu - J'm. \tag{5}$$

For any of the sampling schemes considered, it will be enough to find MLEs by maximizing $\ell(n,\mu)$.

If the log-linear model $\mu \in C(X)$ holds, we need to maximize $\ell(n, \mu)$ subject to the condition that $\mu \in C(X)$. Since X is of full rank, μ can be written uniquely as $\mu = Xb$. We need to find the unconstrained maximum of

$$f_n(b) \equiv \ell(n, \mu).$$

Taking derivatives with respect to b (and μ)

$$df_n(b) = [d\ell(n, \mu)]\, d\mu(b),$$

$$d\ell(n, \mu) = d\left(n'\mu - \sum_{i=1}^{q} e^{\mu_i} \right) = n' - (e^{\mu_1}, \ldots, e^{\mu_q}) = n' - m',$$

and

$$d\mu(b) = d(Xb) = X.$$

Substituting, we get

$$df_n(b) = (n - m)'X. \tag{6}$$

It should be recalled that m is a function of b ($m = m(b)$).

Critical points are found by setting the partial derivative matrix, $df_n(b)$, equal to zero. As will be seen below, $\hat{b}$, the MLE of b, must occur at a critical point, so $\hat{b}$ must satisfy

$$X'm(\hat{b}) = X'n. \tag{7}$$

In particular, since $C(X_0) \subset C(X)$, we have $X_0'm(\hat{b}) = X_0'n$. As indicated above, if $C(X_0) \subset C(X)$ the additional restriction on the MLEs from product-multinomial sampling is automatically satisfied.

By considering $d^2f_n(b)$, we can investigate the nature of the critical points.

$$d^2f_n(b) = d(df_n(b)') = d(X'n - X'm(b)) = -X'dm(b). \tag{8}$$

Write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_q' \end{bmatrix} = [x_{ij}],$$

then

$$m(b) = (m_1, \ldots, m_q)' = \left(e^{x_1'b}, \ldots, e^{x_q'b} \right)'.$$

Therefore,

$$dm(b) = \begin{bmatrix} x_{11}e^{x_1'b} & \cdots & x_{1p}e^{x_1'b} \\ \vdots & & \vdots \\ x_{q1}e^{x_q'b} & \cdots & x_{qp}e^{x_q'b} \end{bmatrix}$$

$$\begin{aligned} &= \begin{bmatrix} x_1' e^{x_1'b} \\ \vdots \\ x_q' e^{x_q'b} \end{bmatrix} = \begin{bmatrix} x_1' m_1 \\ \vdots \\ x_q' m_q \end{bmatrix} \\ &= D(m)X, \end{aligned} \tag{9}$$

where $m = m(b)$. Substitution into (8) gives

$$d^2 f_n(b) = -X' D(m(b)) X. \tag{10}$$

In a log-linear model, m_i is always positive because $m_i = e^{\mu_i}$; therefore, $d^2 f_n(b)$ is negative definite and $f_n(b)$ is strictly convex. If $f_n(b)$ takes on its maximum, it must be at a critical point and the maximum is unique. The maximum is the MLE $\hat{b} = \hat{b}(n)$. $\hat{b}$ uniquely determines MLEs $\hat{\mu} = \hat{\mu}(n) = X\hat{b}$ and $\hat{m} = \hat{m}(n) = m(\hat{\mu}) = m(\hat{b})$.

A simple condition exists that ensures that $f_n(b)$ takes on its maximum.

Theorem 12.2.1. If there exists $\xi \perp C(X)$ such that $n_i + \xi_i > 0$ for $i = 1, \ldots, q$, then $\ell(n, \mu) = f_n(b)$ attains its maximum.

Proof.

$$\ell(n, \mu) = n'\mu - \sum_{i=1}^{q} e^{\mu_i}.$$

Let $\xi \perp C(X)$, then for $\mu \in C(X)$, $\ell(n, \mu) = g(\mu)$ where

$$\begin{aligned} g(\mu) &= (n + \xi)'\mu - \sum_{i=1}^{q} e^{\mu_i} \\ &= \sum_{i=1}^{q} (n_i + \xi_i)\mu_i - e^{\mu_i}. \end{aligned}$$

With all the $n_i + \xi_i$'s positive, as any $\mu_i \to -\infty$, $g(\mu) \to -\infty$. Similarly, if any $\mu_i \to \infty$, $g(\mu) \to -\infty$. This establishes that the convex function $g(\mu)$ must take on its maximum. Moreover, $g(\mu)$ must take on its maximum for $\mu \in C(X)$ because $C(X)$ is a closed set. □

In particular, if all the n_i's are positive, the MLE of μ exists. Henceforth, we assume that the (unique) MLE of μ exists.

In finding MLEs of μ and m, we relied on a particular parameterization $\mu = Xb$. Any alternative parameterization $\mu = Z\gamma$ where $C(Z) = C(X)$ is equally valid. Regardless of the parameterization, we are maximizing $\ell(n, \mu)$ subject to the constraint that $\mu \in C(X)$. Since we found a unique MLE $\hat{\mu}$, it must be valid for any parameterization. Similarly, $\hat{m}$ does not depend on the parameterization. In particular, Z need not be of full column rank.

When Z is not of full rank, γ is not estimable. The problem of estimability can be handled as it is for standard linear models.

For testing hypotheses, the likelihood ratio test statistic (LRTS) is often used. To test

$$H_0\colon \mu = \mu_0 \text{ versus } H_A\colon \mu \neq \mu_0 \qquad \text{where } \mu_0 \in C(X),$$

the LRTS is $-2[\ell(n, \mu_0) - \ell(n, \hat{\mu})]$. For testing

$$H_0\colon \mu \in C(X_1) \text{ versus } H_A\colon \mu \notin C(X_1) \qquad \text{where } C(X_1) \subset C(X), \quad (11)$$

the LRTS is $-2[\ell(n, \hat{\mu}_1) - \ell(n, \hat{\mu})]$, where $\hat{\mu}_1$ is the MLE of μ for the log-linear model $\mu \in C(X_1)$. [The reduced model is also assumed to have $C(X_0) \subset C(X_1)$ and $J \in C(X_1)$.]

In both cases, the LRTS simplifies considerably. We investigate the LRTS for hypothesis (11). From (5), $\ell(n, \hat{\mu}_1) - \ell(n, \hat{\mu}) = n'(\hat{\mu}_1 - \hat{\mu}) - J'(\hat{m}_1 - \hat{m})$. Since $J \in C(X_1) \subset C(X)$, from (7) we get $J'(\hat{m}_1 - \hat{m}) = J'n - J'n = 0$. Substitution gives

$$\ell(n, \hat{\mu}_1) - \ell(n, \hat{\mu}) = n'(\hat{\mu}_1 - \hat{\mu}).$$

Since $\hat{\mu}_1, \hat{\mu} \in C(X)$, (7) also gives

$$n'(\hat{\mu}_1 - \hat{\mu}) = n'M(\hat{\mu}_1 - \hat{\mu}) = \hat{m}'M(\hat{\mu}_1 - \hat{\mu}) = \hat{m}'(\hat{\mu}_1 - \hat{\mu}),$$

where $M = X(X'X)^{-1}X'$ is the perpendicular projection matrix onto $C(X)$. Thus, the LRTS is

$$-2[\hat{m}'(\hat{\mu}_1 - \hat{\mu})] = 2\sum_{i=1}^{q} \hat{m}_i \log(\hat{m}_i/\hat{m}_{1i}). \quad (12)$$

Contingency tables with structural zeros ($m_i = 0$) frequently occur. Under the sampling schemes considered here, $m_i = 0$ implies that $\Pr(n_i = 0) = 1$; therefore, without loss of generality, such cells can simply be dropped from the model and the theory does not change.

12.3 Asymptotic Properties

In establishing asymptotic properties, we let N go to infinity, where N is the total sample size in product-multinomial sampling and the sum of the expected values for the q cells in Poisson sampling. For product-multinomial sampling, we assume that the probabilities in each cell remain constant and that the individual sample sizes for each population remain in a fixed proportion; i.e., N_j/N remains constant. For Poisson sampling, we assume that the ratio of any pair of expected values remains constant.

Terminology appropriate for product-multinomial sampling will be used, but the results also hold for Poisson sampling. N will be referred to as the sample size. n_N is a sample of size N with $\mathrm{E}(n_N) = m_N$. By our assumptions, $m_N = Nm^*$ for some vector m^*. For multinomial sampling, m^* is the vector of probabilities. For product-multinomial sampling, m^* is the vector of normalized probabilities. The probabilities are normalized by the relative sizes of the populations so that $J'm^* = \sum_{i=1}^{q} m_i^* = 1$. In particular, if i is a cell in the jth multinomial population, m_i^* is $p_i(N_j/N)$ where p_i is the probability of getting an observation from the jth population in the ith cell. For Poisson sampling, m^* is the vector of expected values divided by N, thus $m_N = Nm^*$. Note that for all sampling schemes, $J'm_N = N$, hence $J'm^* = 1$.

For a log-linear model to be valid, we need additional restrictions on m_N. In particular, writing $\mu_N = \log(m_N)$, we need $\mu_N \in C(X)$. Since $m_N = Nm^*$, we have $\mu_N = \log(m_N) = (\log N)J + \log(m^*)$. Since J is assumed to be in $C(X)$, it is sufficient to require $\log(m^*) \in C(X)$. Henceforth, we make the assumption that for some b^*, $\log(m^*) = Xb^*$. Define $\mu^* = Xb^*$; then, $\mu_N = \mu^* + (\log N)J$.

Some special matrices will be used frequently in the sequel. Let $D = D(m^*)$ and $A = X(X'DX)^{-1}X'D$. Let A_0 and A_1 be defined as A, with X_0 and X_1 replacing X. Note that A is a projection matrix onto $C(X)$. (In fact, A is the perpendicular projection matrix for the inner product determined by D.) A has the same form as a matrix giving best linear unbiased estimates in a regular linear model with covariance matrix D^{-1}. Taking $D^{1/2} = D(\sqrt{m^*})$ we see that $P = D^{1/2}AD^{-1/2}$ is the perpendicular projection matrix onto $C(D^{1/2}X)$ (with the usual inner product).

We first consider properties of n_N for large samples. These results are well known but necessary for the rest of the development.

Theorem 12.3.1.

(1) $N^{-1}n_N \to m^*$ a.s.,

(2) $N^{-1}n_N \xrightarrow{P} m^*$,

(3) $N^{-1/2}(n_N - m_N) \xrightarrow{L} N(0, D[I - A_0])$, and

(4) $N^{-1}(n_N - m_N) = O_p\big(N^{-1/2}\big)$.

Justification. For product-multinomial sampling, (1) is a direct application of the strong law of large numbers applied to the elements of n_N. For Poisson sampling, (1) follows from Chebyshev's inequality and the Borel-Cantelli Lemma. Part (2) follows from (1). Part (4) follows from (3). It remains only to show (3).

(a) *Poisson sampling.* For Poisson sampling, $X_0 = 0$, so $A_0 = 0$. Since the n_{Ni}'s are independent, it suffices to show the $N^{-1/2}(n_{Ni} - m_{Ni})$ is asymptotically $N(0, m_i^*)$. We use a moment generating function argument. The moment generating function of a random variable w is $\varphi_w(u) = \mathrm{E}(e^{uw})$. These behave similarly to characteristic functions, but moment generating functions do not exist for some random variables. The moment generating function we need is

$$\varphi_{N^{-1/2}(n_{Ni}-m_{Ni})}(u) = \varphi_{n_{Ni}}(N^{-1/2}u)\exp\Big(-N^{-1/2}m_{Ni}u\Big),$$

where

$$\varphi_{n_{Ni}}(u) = \exp[-m_{Ni}(1 - e^u)].$$

Using a Taylor's expansion,

$$e^{au} = 1 + au + a^2u^2/2 + a^3\tilde{u}^3/6,$$

for some $\tilde{u} \in [0, u]$, we can write

$$\begin{aligned}
&\log \varphi_{N^{-1/2}(n_{Ni}-m_{Ni})}(u) \\
&\quad = -m_{Ni}\left[1 - e^{N^{-1/2}u}\right] - N^{-1/2}m_{Ni}u \\
&\quad = -m_{Ni}\left[1 - \left(1 + N^{-1/2}u + N^{-1}u^2/2 + N^{-3/2}\tilde{u}^3/6\right)\right] - N^{-1/2}m_{Ni}u \\
&\quad = N^{-1}m_{Ni}u^2/2 + N^{-3/2}m_{Ni}\tilde{u}^3/6 \\
&\quad = m_i^*u^2/2 + N^{-1/2}m_i^*\tilde{u}^3/6.
\end{aligned}$$

As $N \to \infty$,

$$\log \varphi_{N^{-1/2}(n_{Ni}-m_{Ni})}(u) \to m_i^*u^2/2,$$

so

$$N^{-1/2}(n_{Ni} - m_{Ni}) \xrightarrow{L} N(0, m_i^*).$$

(b) *Multinomial Sampling.* For multinomial sampling, $X_0 = J$, and $A_0 = JJ'D$, so $D[I - A_0] = D - DJJ'D$. Part (3) is then the standard large sample result for multinomials, which is an immediate consequence of the multivariate central limit theorem.

(c) *Product-Multinomial Sampling.* The different multinomial populations are asymptotically normal and they are independent by assumption. It remains only to establish that $D[I - A_0]$ is the correct block diagonal covariance matrix. Write n_N so that all observations in the first multinomial population are listed first, the second population is listed second, etc. Considering the implied structure of X_0, it is easily seen that $(X_0'DX_0)^{-1} = ND^*$, where $D^* = \mathrm{Diag}\big(N_1^{-1}, \ldots, N_r^{-1}\big)$, and that $D[I - A_0] = D - NDX_0D^*X_0'D$. It is easily seen that $D - NDX_0D^*X_0'D$ is precisely the block diagonal matrix needed. □

The main result used in finding asymptotic properties of maximum likelihood estimates is a relationship between the MLE $\hat{\mu}_N$ and the observations n_N.

Lemma 12.3.2. $N^{1/2}(\hat{\mu}_N - \mu_N) - \big(AD^{-1}\big)\, N^{-1/2}(n_N - m_N) \xrightarrow{P} 0.$

Proof. See Section 5. □

Since the asymptotic distribution of $N^{-1/2}(n_N - m_N)$ is known, the lemma gives the asymptotic distribution of $N^{1/2}(\hat{\mu}_N - \mu_N)$ and, by a Taylor's expansion, the asymptotic distribution of $N^{-1/2}(\hat{m}_N - m_N)$.

Theorem 12.3.3.

(1) $N^{1/2}(\hat{\mu}_N - \mu_N) \xrightarrow{L} N\big(0, [A - A_0]\, D^{-1}\big)$.

(2) $\hat{\mu}_N - \mu_N \xrightarrow{P} 0$.

(3) $N^{-1/2}(\hat{m}_N - m_N) \xrightarrow{L} N(0, D\,[A - A_0])$.

(4) $N^{-1}\hat{m}_N \xrightarrow{P} m^*$.

Proof.
(1) Theorem 12.3.1 and Lemma 12.3.2 imply that

$$N^{1/2}(\hat{\mu}_N - \mu_N) \xrightarrow{L} N\big(0, AD^{-1}[D(I - A_0)]D^{-1}A'\big).$$

It is easily seen that

$$\begin{aligned} AD^{-1}[D(I - A_0)]D^{-1}A' &= AD^{-1}A' - AA_0D^{-1}A' \\ &= AD^{-1} - A_0D^{-1} \\ &= (A - A_0)D^{-1}. \end{aligned}$$

(2) From (1), $N^{1/2}(\hat{\mu}_N - \mu_N) = O_p(1)$, so $\hat{\mu}_N - \mu_N = O_p\big(N^{-1/2}\big) = o_p(1)$.

(3) Recall that $\exp(y) = (e^{y_1}, \ldots, e^{y_q})'$. Taylor's theorem gives

$$\begin{aligned} \exp(y) &= \exp(x) + [d\exp(x)]\,(y - x) + o(\|y - x\|) \\ &= \exp(x) + D\big(\exp(x)\big)(y - x) + o(\|y - x\|). \end{aligned} \tag{1}$$

Let $y = \hat{\mu}_N - \big(\frac{1}{2}\log\, N\big)\, J$ and $x = \mu_N - \big(\frac{1}{2}\log\, N\big)\, J$. Since $\hat{\mu}_N - \mu_N = o_p(1)$, rearranging equation (1) and evaluating $\exp(y)$ and $\exp(x)$ gives

$$N^{-1/2}\hat{m}_N - N^{-1/2}m_N - \Big[D\Big(N^{-1/2}m_N\Big)\Big](\hat{\mu}_N - \mu_N) = o\big(o_p(1)\big) = o_p(1).$$

Since $m_N = Nm^*$, we have

$$N^{-1/2}(\hat{m}_N - m_N) - [D]\, N^{1/2}(\hat{\mu}_N - \mu_N) = o_p(1).$$

Applying part (1) of the theorem, we see that

$$N^{-1/2}(\hat{m}_N - m_N) \xrightarrow{L} N\big(0, D(A - A_0)D^{-1}D\big)\,,$$

but $D(A - A_0)D^{-1}D = D(A - A_0)$.

(4) $N^{-1/2}(\hat{m}_N - m_N) = O_p(1)$ by part (3), so $N^{-1}(\hat{m}_N - m_N) = O_p\big(N^{-1/2}\big) = o_p(1)$. Now part (4) follows by observing that $N^{-1}m_N = m^*$. □

It is of interest to note that Theorem 12.3.3 implies $\hat{\mu}_N - (\log N)J \xrightarrow{P} \mu^*$, so $\hat{\mu}_N$ is not consistent for μ^*.

The following corollary will be useful in examining the likelihood ratio test.

Corollary 12.3.4. $\hat{b}_N - b_N = O_p\big(N^{-1/2}\big)$, therefore $\hat{b}_N - b_N \xrightarrow{P} 0$.

Proof. From Theorem 12.3.3, $X\hat{b}_N - Xb_N = \hat{\mu}_N - \mu_N = O_p\big(N^{-1/2}\big)$. It follows that $\hat{b}_N - b_N = (X'X)^{-1}X'\Big[X\hat{b}_N - Xb_N\Big] = (X'X)^{-1}X'O_p\big(N^{-1/2}\big) = O_p\big(N^{-1/2}\big)$. □

We now consider the problem of testing

$$H_0\colon \mu_N = \mu_{0N} \quad \text{versus} \quad H_A\colon \mu_N \neq \mu_{0N}. \tag{2}$$

As the sample size N gets larger, m_N gets larger and so does μ_N. Using a fixed value for the hypothesis, say $H_0\colon \mu_N = \mu_0$, is not appropriate.

μ_{0N} must be in $C(X)$ and compatible with having a sample size of N, i.e., $J'm_{0N} = N$. In accordance with our other assumptions, we consider only the case where $\mu_{0N} = (\log N)J + \mu_0^*$ with $\mu_0^* \in C(X)$, and $J'm_0^* = 1$. m_{0N}, b_{0N}, m_0^*, and b_0^* are defined in the usual way. Note that (2) is equivalent to

$$H_0\colon m^* = m_0^* \quad \text{versus} \quad H_A\colon m^* \neq m_0^*,$$

so one can think of the hypothesis as being on the (normalized) vector of probabilities.

The asymptotic distribution theory for the more interesting hypothesis

$$H_0\colon \mu_N \in C(X_1) \quad \text{versus} \quad H_A\colon \mu_N \notin C(X_1), \tag{3}$$

where $C(X_0) \subset C(X_1) \subset C(X)$, can be handled quite simply after dealing with (2).

We want to show that the likelihood ratio test statistic (LRTS), $-2[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)]$, has an asymptotic χ^2 distribution.

Theorem 12.3.5. If $\mu_N = \mu_{0N}$, then $-2[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{L} \chi^2(p-r)$.

Proof. The proof is in four parts. The first three find statistics asymptotically equivalent to the LRTS. The last one establishes the distribution of the LRTS.

(a) Let $f_n(b) = \ell(n, \mu(b))$. A Taylor's expansion of $f_n(b)$ about $\tilde{b}$ gives

$$\begin{aligned} f_n(b) &= f_n(\tilde{b}) + [df_n(\tilde{b})](b-\tilde{b}) + \frac{1}{2}(b-\tilde{b})' \left[d^2 f_n(\tilde{b})\right](b-\tilde{b}) \\ &\quad + o\Big(\|b-\tilde{b}\|^2\Big). \end{aligned}$$

Substituting for $df_n(\tilde{b})$ and $d^2 f_n(\tilde{b})$ as found in (12.2.6) and (12.2.10), we get

$$\begin{aligned} f_n(b) - f_n(\tilde{b}) - \Big[n - m(\tilde{b})\Big]' X(b-\tilde{b}) \\ + \frac{1}{2}(b-\tilde{b})' \Big[X' D\big(m(\tilde{b})\big) X\Big](b-\tilde{b}) = o\Big(\|b-\tilde{b}\|^2\Big). \quad (4) \end{aligned}$$

Apply equation (4) with $n = n_N$, $\tilde{b} = \hat{b}_N$, and $b = b_N$. From Corollary 12.3.4, $\hat{b}_N - b_N = o_p(1)$, so

$$\begin{aligned} \ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N) - (n - \hat{m}_N)' X(b_N - \hat{b}_N) \\ + \frac{1}{2}(b_N - \hat{b}_N)' \left[X' D(\hat{m}_N) X\right](b_N - \hat{b}_N) = o\big(o_p(1)\big). \quad (5) \end{aligned}$$

By (12.2.7), $(n_N - \hat{m}_N)' X = 0$. After multiplying by -2, (5) becomes

$$-2[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)] - (\mu_N - \hat{\mu}_N)' D(\hat{m}_N)(\mu_N - \hat{\mu}_N) = o_p(1). \quad (6)$$

The quadratic form in (6) can be rewritten as

$$N^{1/2}(\mu_N - \hat{\mu}_N)' D\big(N^{-1}\hat{m}_N\big)\, N^{1/2}(\mu_N - \hat{\mu}_N). \quad (7)$$

(b) For random variables Y_N and Z_N, it is well known that if $Y_N \xrightarrow{L} Y$ and $Z_N \xrightarrow{P} 0$, then $Y_N Z_N \xrightarrow{P} 0$. Repeated application of this gives the result: if $Y_N \xrightarrow{L} Y$ and $Z_N \xrightarrow{P} Z$, then $Y_N' Z_N Y_N - Y_N' Z Y_N \xrightarrow{P} 0$ where Y_N is a q vector and Z_N is a $q \times q$ matrix. Let $Z_N = D\big(N^{-1}\hat{m}_N\big)$ in (7). Since $N^{-1}\hat{m}_N \xrightarrow{P} m^*$,

$$\begin{aligned} N^{1/2}(\mu_N - \hat{\mu}_N)' D\big(N^{-1}\hat{m}_N\big)\, N^{1/2}(\mu_N - \hat{\mu}_N) \\ - N^{1/2}(\mu_N - \hat{\mu}_N)' D N^{1/2}(\mu_N - \hat{\mu}_N) \xrightarrow{P} 0. \end{aligned}$$

(c) Applying Lemma 12.3.2 gives

$$N^{1/2}(\mu_N - \hat{\mu}_N)'DN^{1/2}(\mu_N - \hat{\mu}_N) - N^{-1/2}(n_N - m_N)'D^{-1}A'DAD^{-1}N^{-1/2}(n_N - m_N) \xrightarrow{P} 0. \quad (8)$$

It is easily seen that

$$D^{-1}A'DAD^{-1} = AD^{-1} = D^{-1/2}PD^{-1/2}.$$

Recall that, $D^{-1/2} = D\left(1/\sqrt{m^*}\right)$ and P is the perpendicular projection matrix onto $C(D^{-1/2}X)$. Rewrite the second quadratic form in (8) as

$$\left[D^{-1/2}N^{-1/2}(n_N - m_N)\right]' P \left[D^{-1/2}N^{-1/2}(n_N - m_N)\right]. \quad (9)$$

(d) From Theorem 12.3.1, $D^{-1/2}N^{-1/2}(n_N - m_N) \xrightarrow{L} Y$, where $Y \sim N\left(0, D^{1/2}[I - A_0]D^{-1/2}\right)$. As in part (c), it is easy to see that $D^{1/2}[I - A_0]D^{-1/2} = I - P_0$, where P_0 is the perpendicular projection matrix onto $C(D^{-1/2}X_0)$. The quadratic form (9) converges in distribution to $Y'PY$. By Theorem 1.3.6 in Christensen (1996b), $Y'PY$ will have a χ^2 distribution with $tr[P(I - P_0)]$ degrees of freedom if

$$(I - P_0)P(I - P_0)P(I - P_0) = (I - P_0)P(I - P_0). \quad (10)$$

Since $C(P_0) \subset C(P)$, we have $PP_0 = P_0$. Simplifying gives both sides of (10) as $(P - P_0)$.

The theorem follows by observing that under H_0, $\mu_N = \mu_{0N}$, so the asymptotic distribution of (9) is the same as that of the LRTS, and that $tr[P(I - P_0)] = tr[P - P_0] = tr(P) - tr(P_0) = p - r$. □

We would like to show that the likelihood ratio test is consistent. That is, if $\mu_{0N} \neq \mu_N$, then $-2[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} \infty$. Consistency is an immediate result of the following theorem.

Theorem 12.3.6.

$$-2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} -2[\ell(m^*, \mu_0^*) - \ell(m^*, \mu^*)].$$

If $\mu_{0N} - \mu_N \equiv \mu_0^* - \mu^* \neq 0$, the right-hand side is strictly positive.

Proof.

$$-2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \hat{\mu}_N)] \quad (11)$$
$$= -2N^{-1}[\ell(n_N, \mu_{0N}) - \ell(n_N, \mu_N)] - 2N^{-1}[\ell(n_N, \mu_N) - \ell(n_N, \hat{\mu}_N)].$$

Consider the second term of (11). From (6),

$$-2N^{-1}[\ell(n_N,\mu_N)-\ell(n_N,\hat{\mu}_N)] \\ -N^{-1}(\mu_N-\hat{\mu}_N)'D(\hat{m}_N)(\mu_N-\hat{\mu}_N)=o_p\big(N^{-1}\big)=o_p(1).$$

Since $(\mu_N-\hat{\mu}_N)\xrightarrow{P}0$ and $N^{-1}\hat{m}_N\xrightarrow{P}m^*$, we have $N^{-1}(\mu_N-\hat{\mu}_N)'D(\hat{m}_N)(\mu_N-\hat{\mu}_N)\xrightarrow{P}0$. Therefore, $-2N^{-1}[\ell(n_N,\mu_N)-\ell(n_N,\hat{\mu}_N)]\xrightarrow{P}0$.

Now consider the first term on the right of (11). By definition,

$$\begin{aligned}&-2N^{-1}[\ell(n_N,\mu_{0N})-\ell(n_N,\mu_N)]\\ &=\ -2N^{-1}[n_N'\mu_{0N}-J'm_{0N}]+2N^{-1}[n_N'\mu_N-J'm_N]\\ &=\ -2N^{-1}[n_N'(\mu_{0N}-\mu_N)-J'(m_{0N}-m_N)]\,.\end{aligned}$$

Since $\mu_{0N}-\mu_N=\mu_0^*-\mu^*$, $N^{-1}m_{0N}=m_0^*$, $N^{-1}m_N=m^*$, and $N^{-1}n_N\xrightarrow{P}m^*$, we have

$$\begin{aligned}-2N^{-1}[\ell(n_N,\mu_{0N})-\ell(n_N,\mu_N)]\ &\xrightarrow{P}\ -2[m^*(\mu_0^*-\mu^*)-J'(m_0^*-m^*)]\\ &=\ -2[\ell(m^*,\mu_0^*)-\ell(m^*,\mu^*)]\,.\end{aligned}$$

As discussed in Lemma 12.5.3, μ^* is the unique MLE of μ when the data are m^* [i.e. $\mu^*=\hat{\mu}(m^*)$], so $\ell(m^*,\mu^*)$ is the unique maximum of $\ell(m^*,\mu)$. If $\mu_0^*\neq\mu^*$, $-2[\ell(m^*,\mu_0^*)-\ell(m^*,\mu^*)]>0$. □

We now consider the problem of testing (3).

Theorem 12.3.7. If $\mu_N\in C(X_1)$, then $-2[\ell(n_N,\hat{\mu}_{1N})-\ell(n_N,\hat{\mu}_N)]\xrightarrow{L}\chi^2(p-p_1)$, where $\hat{\mu}_{1N}$ is the MLE of μ_N under the model $\mu_N\in C(X_1)$ and $r(X_1)=p_1$.

Proof.

$$-2[\ell(n_N,\hat{\mu}_{1N})-\ell(n_N,\hat{\mu}_N)] \\ =-2[\ell(n_N,\mu_N)-\ell(n_N,\hat{\mu}_N)]+2[\ell(n_N,\mu_N)-\ell(n_N,\hat{\mu}_{1N})]\,.$$

Since $C(X_1)\subset C(X)$, the proof of Theorem 12.3.5 applies to both terms on the right-hand side. In particular, (9) can be applied as is and also with X_1 substituted for X. If P_1 is the perpendicular projection matrix onto $C(D^{-1/2}X_1)$, then $P_1=PP_1=P_1P$, so the LRTS is asymptotically equivalent to $\big[D^{-1/2}N^{-1/2}(n_N-m_N)\big]'(P-P_1)\big[D^{-1/2}N^{-1/2}(n_N-m_N)\big]$. Since $(P-P_1)(I-P_0)(P-P_1)=(P-P_1)$, verification of the conditions of Theorem 1.3.6 is trivial. The LRTS has a χ^2 distribution with $r(P-P_1)=p-p_1$ degrees of freedom. □

Theorem 12.3.8 establishes the consistency of the likelihood ratio test for hypothesis (3).

Theorem 12.3.8.

$$-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} -2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)]\,.$$

$\mu_N \notin C(X_1)$ if and only if the right-hand side is positive.

Proof. The proof involves several arguments from the proof of Lemma 12.3.2, so it is deferred until Section 5. □

Finally, we establish the asymptotic equivalence under H_0 of the LRTS and the Pearson test statistic (PTS). The Pearson test statistic is

$$(\hat{m}_N - \hat{m}_{1N})' D^{-1}(\hat{m}_{1N})(\hat{m}_N - \hat{m}_{1N}) \tag{12}$$

for the hypothesis (3).

Theorem 12.3.9. If $\mu_N \in C(X_1)$, then

$$-2[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] - (\hat{m}_N - \hat{m}_{1N})' D^{-1}(\hat{m}_{1N})(\hat{m}_N - \hat{m}_{1N}) \xrightarrow{P} 0.$$

Proof. The PTS can be written elementwise as

$$\sum_{i=1}^{q} (\hat{m}_{Ni} - \hat{m}_{1Ni})^2 / \hat{m}_{1Ni}. \tag{13}$$

The elementwise form for the LRTS is found in (12.2.12). Note that (13) is equivalent to

$$N \sum_{i=1}^{q} \left[N^{-1}(\hat{m}_{Ni} - \hat{m}_{1Ni})\right]^2 / N^{-1}\hat{m}_{1Ni}$$

and the LRTS is

$$2N \sum_{i=1}^{q} \left(N^{-1}\hat{m}_{Ni}\right)\left[\log\left(N^{-1}\hat{m}_{Ni}\right) - \log\left(N^{-1}\hat{m}_{1Ni}\right)\right]. \tag{14}$$

To simplify notation, let $(x, y) = (N^{-1}\hat{m}_N, N^{-1}\hat{m}_{1N})$. Taking a second-order expansion of (14) about (m^*, m^*) gives

$$2N \sum_{i=1}^{q} x_i \left[\log x_i - \log y_i\right]$$

$$
\begin{aligned}
= \quad & 2N \sum_{i=1}^{q} m_i^* \left[\log m_i^* - \log m_i^*\right] \\
& + 2N \sum_{i=1}^{q} \left[\log x_i - \log y_i + x_i(1/x_i)\right] |_{(x,y)=(m^*,m^*)} (x_i - m_i^*) \\
& + 2N \sum_{i=1}^{q} \left[-x_i(1/y_i)\right] |_{(x,y)=(m^*,m^*)} (y_i - m_i^*) \\
& + N \sum_{i=1}^{q} \left[x_i^{-1}\right] |_{x=m^*} (x_i - m_i^*)^2 \\
& + 2N \sum_{i=1}^{q} \left[-y_i^{-1}\right] |_{y=m^*} (x_i - m_i^*)(y_i - m_i^*) \\
& + N \sum_{i=1}^{q} \left[x_i y_i^{-2}\right] |_{(x,y)=(m^*,m^*)} (y_i - m_i^*)^2 \\
& + No\big(\|x - m^*\|^2 + \|y - m^*\|^2\big) .
\end{aligned}
$$

This easily simplifies to

$$
\begin{aligned}
& 2N \sum_{i=1}^{q} x_i \left[\log x_i - \log y_i\right] \\
= \quad & 2N \sum_{i=1}^{q} (x_i - y_i) \\
& + N \sum_{i=1}^{q} (1/m_i^*) \left[(x_i - m_i^*)^2 - 2(x_i - m_i^*)(y_i - m_i^*) + (y_i - m_i^*)^2\right] \\
& + No\big(\|x - m^*\|^2 + \|y - m^*\|^2\big) \\
= \quad & N \sum_{i=1}^{q} (x_i - y_i)^2 / m_i^* + No\big(\|x - m^*\|^2 + \|y - m^*\|^2\big) .
\end{aligned}
$$

The last equality is because $J \in C(X_1) \subset C(X)$, so that by (12.2.7), $NJ'x = J'n_N = NJ'y$ and $J'(x - y) = 0$.

In our regular notation,

$$
\begin{aligned}
2 \sum_{i=1}^{q} \hat{m}_{Ni} \left[\log\big(\hat{m}_{Ni}/\hat{m}_{1Ni}\big)\right] &= \sum_{i=1}^{q} (\hat{m}_{Ni} - \hat{m}_{1Ni})^2 / m_{Ni} \\
& \quad + No\big(\|N^{-1}(\hat{m}_N - m_N)\|^2 + \|N^{-1}(\hat{m}_{1N} - m_N)\|^2\big) .
\end{aligned}
$$

Investigating the argument of $o(\cdot)$,

$$
\begin{aligned}
& \|N^{-1}(\hat{m}_N - m_N)\|^2 + \|N^{-1}(\hat{m}_{1N} - m_N)\|^2 \\
& \qquad = \left[O_p\Big(N^{-1/2}\Big)\right]^2 + \left[O_p\Big(N^{-1/2}\Big)\right]^2 = O_p\big(N^{-1}\big) .
\end{aligned}
$$

Since $o(O_p(N^{-1})) = o_p(N^{-1})$ and $No_p(N^{-1}) = o_p(1)$, we have the LRTS asymptotically equivalent to

$$\begin{aligned}\sum_{i=1}^{q} (\hat{m}_{Ni} - \hat{m}_{1Ni})^2 / m_{Ni} \\ &= (\hat{m}_N - \hat{m}_{1N})' D^{-1}(m_N)(\hat{m}_N - \hat{m}_{1N}) \qquad (15) \\ &= N^{-1/2}(\hat{m}_N - \hat{m}_{1N})' D^{-1}(m^*) N^{-1/2}(\hat{m}_N - \hat{m}_{1N}).\end{aligned}$$

Under H_0, $D^{-1}\left(N^{-1}\hat{m}_{1N}\right) \xrightarrow{P} D^{-1}(m^*)$, so (15) is asymptotically equivalent to (12) the PTS. □

A similar argument holds to show that the LRTS and the PTS are asymptotically equivalent under H_0 for testing the hypothesis (2). Results similar to Theorems 12.3.6 and 12.3.8 also exist for the PTS. For a more extensive discussion of the asymptotic properties of log-linear models, see Haberman (1974a).

12.4 Applications

a) *Weighted Least Squares.* We now consider two methods of obtaining estimates for log-linear models. The first is the Newton-Raphson technique, which is an iterative method for obtaining the maximum likelihood estimates. The Newton-Raphson method can be performed by doing a series of weighted least squares regression analyses. The second method is an approximate technique based on the asymptotic results that we have derived. It is a noniterative weighted least squares regression approach.

The Newton-Raphson technique is an iterative procedure for finding where a function equals the zero vector. Let g be a function mapping $\mathbf{R}^p$ into $\mathbf{R}^p$. We wish to find b_* such that $g(b_*) = 0$. Let b_0 be an initial guess for b_*. Newton-Raphson defines (recursively) a sequence b_t that converges to b_*. By Taylor's theorem, if b_{t+1} is near b_t, we have the approximate equality

$$g(b_{t+1}) = g(b_t) + [dg(b_t)]\,\delta_t\,,$$

where $\delta_t = b_{t+1} - b_t$. The Newton-Raphson technique assumes that b_t is known, sets $g(b_{t+1}) = 0$, and seeks to find b_{t+1}, i.e.,

$$0 = g(b_t) + [dg(b_t)]\,\delta_t\,,$$

so

$$\delta_t = -\,[dg(b_t)]^{-1} g(b_t) \qquad (1)$$

and

$$b_{t+1} = b_t + \delta_t.$$

For finding maximum likelihood estimates, we wish to find where the derivative of $f_n(b) \equiv \ell(n, Xb)$ is zero. With $g(b) \equiv [df_n(b)]'$ and substituting (12.2.6) and (12.2.10) into (1), we get

$$\delta_t = \left[X'D\big(m(b_t)\big)X\right]^{-1} X'\big(n - m(b_t)\big).$$

The sequence b_t converges to a critical point of $\ell(n, Xb)$ which, as we have seen, must be the MLE of b under fairly weak conditions.

A weighted least squares computer program can be used to execute the Newton-Raphson procedure. Fit the model $Y = Xb+e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = D\big(m(b_t)\big)^{-1}$ where Y is taken as

$$\begin{aligned} Y &= Xb_t + D\big(m(b_t)\big)^{-1}\big(n - m(b_t)\big) \\ &= \log(m_t) + D(m_t)^{-1}(n - m_t). \end{aligned}$$

Let b_{t+1} denote the estimate of b obtained from this procedure. Clearly,

$$\begin{aligned} b_{t+1} &= \left[X'D\big(m(b_t)\big)X\right]^{-1} X'D(m_t)Y \\ &= b_t + \left[X'D\big(m(b_t)\big)X\right]^{-1} X'\big(n - m(b_t)\big), \end{aligned}$$

which is the Newton-Raphson value for b_{t+1}.

The second method is based on the asymptotic results of Theorem 12.3.3 and the fact, shown in Theorem 10.1.3 in Christensen (1996b), that best linear unbiased estimates (BLUEs) for the linear model $Y = X\beta+e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = V$, are the same as those for the model $Y = X\beta + e$, $\mathrm{E}(e) = 0$, $\mathrm{Cov}(e) = V + XUX'$, where U is any non-negative definite matrix.

The saturated log-linear model, $\mu \in \mathbf{R}^q$ always fits the data. For the saturated model $\hat{\mu} = \log(n)$ and $A = I$. If N is large, Theorem 12.3.3 gives the asymptotic relation

$$\log(n) - \mu \sim N\big(0, (I - A_0)D^{-1}(m)\big). \tag{2}$$

It will be convenient to rewrite the term $A_0D^{-1}(m)$. It is easily seen that

$$A_0D^{-1}(m) = X_0\left(X_0'D(m)X_0\right)^{-1}X_0'.$$

Now consider the term $X_0'D(m)X_0$. The matrix X_0 has the same structure as the design matrix for a one-way ANOVA model, say

$$y_{ij} = \mu_i + e_{ij}.$$

Exploiting the simple form of X_0 and the fact that $X_0'n = X_0'm = (N_1, \ldots, N_r)'$, it is easily seen that

$$X_0'D(m)X_0 = D(X_0'n).$$

We now have

$$A_0 D^{-1}(m) = X_0 D^{-1}(X_0' n) X_0',$$

and the asymptotic distribution of $\log(n)$ is

$$\log(n) - \mu \sim N\big(0, D^{-1}(m) - X_0 D^{-1}(X_0' n) X_0'\big). \tag{3}$$

Imposing the linear constraint $\mu = Xb$, the asymptotic distribution (3) leads to fitting the linear model

$$\log(n) = Xb + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = D^{-1}(m) - X_0 D^{-1}(X_0' n) X_0'. \tag{4}$$

By Theorem 10.1.3, the BLUEs in model (4) are the same as those in

$$\log(n) = Xb + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = D^{-1}(m). \tag{5}$$

Of course, m is unknown, so (5) cannot be used directly. Estimating m with n gives the model

$$\log(n) = Xb + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = D^{-1}(n). \tag{6}$$

Note that n is the MLE of m under the saturated model. One virtue of model (6) is that it can be fit with any regression program that does weighted regression.

Besides the rationale just given, there are two other justifications for using this approximate procedure. First, if we take $m_0 = n$ in the Newton-Raphson algorithm, then the first step of the algorithm is precisely fitting model (6). Second, for product-multinomial data, fitting model (6) is the same procedure as that proposed by Grizzle, Starmer, and Koch (1969). In their paper, they consider modeling the vector of probabilities $p = (p_1, \ldots, p_q)'$. Their method specifies a generalized linear model $F(p) = Xb$, where F is a quite general function from $\mathbf{R}^q$ into $\mathbf{R}^q$. In particular, one can choose $F(p) = \log(m)$. With this choice of F, their estimation procedure amounts to a weighted least squares analysis with the covariance matrix

$$\mathrm{Cov}(e) = D^{-1}(n) - X_0 D^{-1}(X_0' n) X_0'.$$

Because $C(X_0) \subset C(X)$, the best linear unbiased estimates under this covariance matrix are the same as those using model (6).

b) *Asymptotic Variances Under Saturated Models.* The relation (2) is the basis for a number of asymptotic variance formulas commonly used with saturated models. For a saturated model, $\mathrm{Cov}(\hat{\mu}) = D^{-1}(m) - A_0 D^{-1}(m)$.

Suppose that we write a saturated model $\mu = Xb$, where $X = [X_0, X_1]$ and $b' = [b_0', b_1']$. The parameter b_0 is forced into the model to deal with the product-multinomial sampling. [Recall that to deal with product-multinomial sampling, we always assume that $C(X_0) \subset C(X)$.] For a linear

function $\rho'\mu$ where $\rho \perp X_0$, one gets $\rho'\mu = \rho'X_0b_0 + \rho'X_1b_1 = \rho'X_1b_1$ and $\text{Var}(\rho'\hat{\mu}) = \rho'D^{-1}(m)\rho$ because $\rho'A_0D^{-1}(m)\rho = 0$. The maximum likelihood estimate of $\rho'D^{-1}(m)\rho$ is $\rho'D^{-1}(n)\rho$. In other words, for estimable functions of the parameters that are not forced into the model (i.e., functions involving only b_1), the estimate of the variance of $\rho'\hat{\mu} = \rho'X_1\hat{b}_1$ is $\rho'D^{-1}(n)\rho$.

EXAMPLE 12.4.1. Consider now a 2×3 table. One log-odds ratio is $\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$, with estimate $\log(n_{11}n_{22}/n_{12}n_{21})$. The estimated variance is then $n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1}$.

c) *Logit and Multinomial Response Models.* Suppose that the sampling scheme is product-multinomial where each multinomial has exactly two categories. Without loss of generality, we can write $n = (n_{11}, \ldots, n_{r1}, n_{12}, \ldots, n_{r2})'$ where the pairs (n_{i1}, n_{i2}) are the multinomial outcomes. (This two-subscript notation will be used for all vectors discussed.)

Logits are defined by $\log(m_{i1}/m_{i2}) = \mu_{i1} - \mu_{i2}$. Let $\eta = (\mu_{11} - \mu_{12}, \ldots, \mu_{r1} - \mu_{r2})'$ be the vector of logits. A logit model is a model $\eta = Z\beta_*$ for some β_*, where Z is an $r \times p$ matrix.

We wish to show that the logit model defines a log-linear model for μ. Let I_r be an $r \times r$ identity matrix, and let $L' = [I_r, -I_r]$, so that $\eta = L'\mu$. Then the restriction placed on μ is that $\mu \in \mathcal{M}$, where $\mathcal{M} = \{\mu | L'\mu = Z\beta_* \text{ for some } \beta_*\}$. Because of the product-multinomial sampling, we have $X_0 = [I_r, I_r]'$. Now let

$$X_* = \begin{bmatrix} Z \\ 0_{rp} \end{bmatrix},$$

where 0_{rp} is an $r \times p$ matrix of zeros. It is easily seen that $\mathcal{M} = \{\mu | L'\mu = L'X_*\beta_* \text{ for some } \beta_*\}$. Arguing as in Section 3.3 of Christensen (1996b), $\mathcal{M}$ is a vector space, so the logit model has defined a log-linear model.

EXAMPLE 12.4.2. For a 3×2 table, a linear logit model is defined by $\mu_{i1} - \mu_{i2} = \gamma_0 + \gamma_1 t_i$, $i = 1, 2, 3$. The equation $L'\mu = L'X_*\beta_*$ becomes

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{31} \\ \mu_{12} \\ \mu_{22} \\ \mu_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}.$$

Continuing as in Section 3.3, $\mathcal{M}$ can be rewritten as $\mathcal{M} = \{\mu | \mu = \mu_0 + \mu_1, \text{ where } \mu_0 \perp C(L) \text{ and } \mu_1 \in C(X_*)\}$. Thus, $\mathcal{M}$ is the space spanned by the columns of X_* and any spanning set for the orthogonal complement of $C(L)$. In particular, X_0 is a matrix with $C(X_0) = C(L)^{\perp}$. We can write the log-linear model as $\mu = X_0\beta_0 + X_*\beta_*$.

It was assumed at the beginning of the discussion that the sampling was product-multinomial with two categories in each multinomial. Normally, we consider only log-linear models $\mu = Xb$ such that $C(X_0) \subset C(X)$. In the current case, we have defined the logit model, i.e., $\mu \in \mathcal{M}$, $\mathcal{M} = \{\mu | L'\mu = L'X_*\beta_* \text{ for some } \beta_*\}$. We have then shown that $\mathcal{M} = C(X_0, X_*)$, so that a logit model must satisfy the condition $C(X_0) \subset \mathcal{M}$. It is interesting to note that even if the two-category product-multinomial sampling scheme had not been assumed, the logit model would still correspond to a log-linear model consistent with that sampling scheme.

EXAMPLE 12.4.3. Write the log-linear version of the logit model as $\mu = X\beta$, where $X = [X_0, X_*]$ and $\beta' = [\beta_0', \beta_*']$. Because the MLEs satisfy $X'n = X'\hat{m}$, we have $X_0'n = X_0'\hat{m}$, i.e., $n_{i\cdot} = \hat{m}_{i\cdot}$ for $i = 1, \ldots, r$. The log-linear model must be of the form

$$\log(m_{ij}) = u_{1(i)} + \cdots.$$

The notation we have used is quite general, but it lends itself best to two-dimensional tables. Consider a four-dimensional log-linear model with a logit model in the last variable. If the observations are n_{hijk}, we have done nothing but substitute the three subscripts hij for the one subscript k. The argument presented here implies that the MLEs must satisfy $n_{hij\cdot} = \hat{m}_{hij\cdot}$ and the log-linear model must be of the form

$$\log(m_{hijk}) = u_{123(hij)} + \cdots.$$

Consider now the problem of estimating $\rho_1'\eta$; we can write $\rho = L\rho_1$ so that $\rho_1'\eta = \rho_1'L'\mu = \rho'\mu$. Because of the particular structures of L and X_* and the fact that $C(L) \perp C(X_0)$, the estimate of $\rho_1'\eta$ is

$$\rho_1'\hat{\eta} = \rho'\hat{\mu} = \rho_1'L'(X_0\hat{\beta}_0 + X_*\hat{\beta}_*) = \rho_1'L'X_*\hat{\beta}_* = \rho_1'Z\hat{\beta}_*.$$

The estimates in the logit model come directly from the log-linear model and all the asymptotic distribution results continue to apply. In particular, the estimate of $\eta = Z\beta_*$ is $\hat{\eta} = Z\hat{\beta}_*$, where $\hat{\beta}_*$ is estimated from the log-linear model.

Consider a logit model $\eta = Z\beta_*$ and a corresponding log-linear model $\mu = X\beta$, where $X = [X_0, X_*]$ and $\beta' = [\beta_0', \beta_*']$. We wish to be able to test the adequacy of a reduced logit model, say $\eta = Z_1\gamma_*$, where $C(Z_1) \subset C(Z)$. If the log-linear model corresponding to $\eta = Z_1\gamma_*$, say $\mu = X_1\gamma$, has $C(X_1) \subset C(X)$, then the test can proceed immediately from log-linear model theory. If Z_1 is a $r \times p_1$ matrix, we can write $X_{1*}' = [Z_1', 0_{rp_1}']$ and $X_1 = [X_0, X_{1*}]$. Clearly, if $C(Z_1) \subset C(Z)$, we have $C(X_{1*}) \subset C(X_*)$ and $C(X_1) \subset C(X)$, so the test can proceed.

The hypothesis that a logit model $\eta = Z_1\gamma_*$ fits the data relative to a general log-linear model $\mu = X\beta$ is equivalent to hypothesizing,

for X_{1*} with $C(X_{1*}) \subset C(X)$, that $\mu \in \mathcal{M}$, where $\mathcal{M} = \{\mu | \mu \in C(X)$ and $L'\mu = L'X_{1*}\gamma$ for some $\gamma\}$. We can rewrite $\mathcal{M}$ as $\mathcal{M} = \{\mu | \mu = \mu_0 + \mu_1,$ where $\mu_1 \in C(X_{1*}),\ \mu_0 \in C(X)$ and $\mu_0 \perp C(L)\}$. Thus, $\mathcal{M}$ is the space spanned by the columns of X_{1*} and any spanning set for the subspace of $C(X)$ orthogonal to $C(L)$. The usual test for lack of fit of a logit model is $H_0 \colon \mu \in \mathcal{M}$ versus $H_A \colon \mu \in \mathbf{R}^q$, i.e., $C(X) = \mathbf{R}^q$.

Many types of multinomial response models can be written as log-linear models using the method outlined here. An exception are continuation ratio models. They do not correspond to a single log-linear model.

d) *Estimation of Parameters.* Estimation of parameters in log-linear models is very similar to that in standard linear models. A standard linear model

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0$$

implies that

$$\mathrm{E}(Y) = X\beta.$$

The least squares estimate of $X\beta$ is $\hat{Y} = MY$. The least squares estimate of $\rho' X\beta$ is $\rho' M\hat{Y} = \rho' MY$.

Similarly, in a log-linear model we have

$$\log(m) \equiv \mu = Xb.$$

Computer programs often give the MLE of m, i.e., $\hat{m}$. From this, one can obtain $\hat{\mu} = \log(\hat{m})$. Because $\hat{\mu} \in C(X)$, the MLE of $\rho' Xb$ is $\rho'\hat{\mu} = \rho' M\hat{\mu}$.

The key to finding the estimate of an estimable function $\lambda'\beta$ or $\lambda' b$ is in obtaining $M\rho$ so that $\lambda' = \rho' X = \rho' MX$. Given $M\rho$, estimates in the standard linear model can be obtained from Y and estimates in a log-linear model can be obtained from $\hat{\mu}$. Finding such a vector $M\rho$ depends only on λ and X. It does not depend on whether a linear or a log-linear model is being fitted. Christensen (1996b) discusses, in great detail, how to find estimates of estimable functions for standard linear models. The procedure amounts to finding $M\rho$. Precisely the same vectors $M\rho$ work for log-linear models. In other words, if one knows how to use Y to estimate something in a standard linear model, exactly the same technique applied to $\hat{\mu}$ will give the estimate in a log-linear model.

EXAMPLE 12.4.4. Consider a two-dimensional table with parameterization

$$\mu_{ij} = \gamma + \alpha_i + \beta_j + (\alpha\beta)_{ij}\,.$$

In discussions of log-linear models, this model would commonly be written as

$$\mu_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}\,,$$

but it is the same model with either parameterization. Estimates follow just as in a two-way ANOVA. To simplify this as much as possible, let $z_{ij} = \hat{\mu}_{ij}$ and assume the "usual" side conditions, then

$$\begin{aligned}
\hat{\gamma} &= \bar{z}_{\cdot\cdot}, \\
\hat{\alpha}_i &= \bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot}, \\
\hat{\beta}_i &= \bar{z}_{\cdot j} - \bar{z}_{\cdot\cdot}, \\
\widehat{(\alpha\beta)}_{ij} &= z_{ij} - \bar{z}_{i\cdot} - \bar{z}_{\cdot j} + \bar{z}_{\cdot\cdot}.
\end{aligned}$$

It seems very reasonable (to me at any rate) to restrict estimation to estimable functions of b. In that case, the choice of side conditions is of no importance.

Tests and confidence intervals for $\rho' X b$ can be based on Theorem 12.3.3. A large sample approximation is

$$\frac{\rho'\hat{\mu} - \rho' X b}{\sqrt{\rho'(A - A_0)D^{-1}(m)\rho}} \sim N(0,1).$$

Of course, $AD^{-1}(m)$ has to be estimated in order to get a standard error. [$A_0 D^{-1}(m)$ does not depend on unknown parameters.] As indicated in application (b), variances are easy to find in the saturated model; unfortunately, the estimable functions of b are generally quite complicated in the saturated model. If one is willing to use side conditions, the side conditions can sometimes give the illusion that the estimable functions are not complicated.

12.5 Proofs of Lemma 12.3.2 and Theorem 12.3.8

Two results from advanced calculus are needed. Recall that if $F : \mathbf{R}^q \times \mathbf{R}^p \to \mathbf{R}^p$, then $dF(x,y)$ is a p by $q+p$ matrix. Partition $dF(x,y)$ into a $p \times q$ matrix, say $d_x F = [\partial F_i / \partial x_j]$, and a $p \times p$ matrix, $d_y F = [\partial F_i / \partial y_j]$.

The Implicit Function Theorem. If $F : \mathbf{R}^q \times \mathbf{R}^p \to \mathbf{R}^p$, $F(a,c) = 0$, $F(a,y)$ is differentiable, and $d_y F$ is nonsingular at $y = c$, then $F(x,y) = 0$ determines y uniquely as a function of x in a neighborhood of (a,c).

This unique function, say $\xi(x)$, is differentiable and satisfies $\xi(a) = c$ and $F(x, \xi(x)) = 0$ for x in a neighborhood of a.

Proof. See Bartle (1964). □

Corollary 12.5.1. $d\xi(x) = -[d_y F]^{-1}[d_x F]$ where $y = \xi(x)$.

Proof. See Bartle (1964). □

Lemma 12.5.2. If a is a scalar and n is a q vector of counts, then

(1) $\hat{m}(an) = a\hat{m}(n)$

(2) $\hat{\mu}(an) = [\log(a)]J + \hat{\mu}(n)$.

Proof. $\hat{m}(an)$ is the unique solution of $[an - m]'X = 0$ with $\log(\hat{m}(an)) \in C(X)$. We will show that $a\hat{m}(n)$ is also a solution with $\log(a\hat{m}(n)) \in C(X)$, so $\hat{m}(an) = a\hat{m}(n)$. $\hat{m}(n)$ is the unique solution of $[n - m]'X = 0$ with $\log(\hat{m}(n)) \in C(X)$. Clearly, if $[n - \hat{m}(n)]'X = 0$, then $[an - a\hat{m}(n)]'X = 0$, but $\log(a\hat{m}(n)) = [\log(a)]J + \log(\hat{m}(n)) \in C(X)$ because both J and $\log(\hat{m}(n))$ are in $C(X)$.

Taking logs gives $\hat{\mu}(an) = [\log(a)]J + \hat{\mu}(n)$. □

Lemma 12.5.3. $\hat{\mu}(m^*) = \mu^*$ and $\hat{m}(m^*) = m^*$.

Proof. By definition, $m^* = m(b^*)$, so b^* is a solution of $[m^* - m(b)]'X = 0$. Since $\hat{\mu}(m^*)$ is unique, we must have $\hat{\mu}(m^*) = Xb^* = \mu^*$.

$$\hat{m}(m^*) = \exp[\hat{\mu}(m^*)] = \exp[\mu^*] = m^*. \qquad \square$$

Lemma 12.3.2 $N^{1/2}(\hat{\mu}_N - \mu_N) - \left(AD^{-1}\right) N^{-1/2}(n_N - m_N) \xrightarrow{P} 0.$

Proof. The MLE $\hat{\mu}_N$ is defined by $\hat{\mu}_N = X\hat{b}_N$, where $\hat{b}_N$ is a function of n_N which is defined implicitly as the solution to $df_{n_N}(b) = [n_N - m(b)]'X = 0$.

The proof follows from investigating the properties of the Taylor's expansion

$$\hat{\mu}(n) = \hat{\mu}(n_0) + d\hat{\mu}(n_0)(n - n_0) + o(\|n - n_0\|). \tag{1}$$

The expansion is applied with $n = N^{-1}n_N$ and $n_0 = N^{-1}m_N = m^*$. Rewriting (1) gives

$$\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*) - d\hat{\mu}(m^*)(N^{-1}n_N - m^*) = o(\|N^{-1}n_N - m^*\|). \quad (2)$$

We examine the terms $\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*)$ and $d\hat{\mu}(m^*)$ separately.

(a) We show that for any observations vector n_N,

$$\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*) = \hat{\mu}(n_N) - \mu_N.$$

By Lemmas 12.5.2 and 12.5.3,

$$\hat{\mu}(N^{-1}n_N) - \hat{\mu}(m^*) = [\log\, N^{-1}]J + \hat{\mu}(n_N) - \mu^*.$$

Since $\mu_N = [\log\, N]J + \mu^*$, we have the result.

(b) We characterize the $q \times q$ matrix $d\hat{\mu}(m^*)$. $\hat{\mu}(n) = X\hat{b}(n)$, so $d\hat{\mu}(n) = X[d\hat{b}(n)]$, with $\hat{b}(n)$ defined implicitly as a zero of $F(n, b) = X'[n - m(b)]$. For any fixed vector b_0, let $n_0 = m(b_0)$. Then $F(n_0, b_0) = 0$, so by the Implicit Function Theorem, there exists $\hat{b}(n)$ such that if n is close to n_0, $F(n, \hat{b}(n)) = 0$ and (from Corollary 12.5.1) $d\hat{b}(n) = -[d_{\hat{b}}F]^{-1}[d_nF]$. To find $d\hat{b}(n)$, we need $dF(n, b) = [X', -X'dm(b)]$. From (12.2.9), $dm(b) = D(m(b))X$, so $dF(n, b) = [X', -X'D(m(b))X]$,

$$d\hat{b}(n) = [X'D(\hat{m})X]^{-1}X',$$

and $d\hat{\mu}(n) = X[X'D(m(b))X]^{-1}X'$. In particular, $d\hat{\mu}(n_0)$ is always defined.

We need $d\hat{\mu}(m^*)$. From Lemma 12.5.3, we have that $F(m^*, \hat{b}(m^*)) = 0$, so $d\hat{\mu}(m^*)$ is defined and $d\hat{\mu}(m^*) = X[X'D(\hat{m}(m^*))X]^{-1}X'$. Again, from Lemma 12.5.3, $\hat{m}(m^*) = m^*$, so $D(\hat{m}(m^*)) = D(m^*) = D$ and $d\hat{\mu}(m^*) = X[X'DX]^{-1}X' = AD^{-1}$.

(c) Using $\|N^{-1}n_N - m^*\| = O_p\left(N^{-1/2}\right)$ and the results of (a) and (b) in (2) gives

$$\hat{\mu}(n_N) - \mu_N - \left(AD^{-1}\right) N^{-1}(n_N - m_N) = o\left(O_p\left(N^{-1/2}\right)\right) = o_p\left(N^{-1/2}\right).$$

Multiplying through by $N^{1/2}$ gives

$$N^{1/2}(\hat{\mu}_N - \mu_N) - \left(AD^{-1}\right) N^{-1/2}(n_N - m_N) = o_p(1). \qquad \square$$

Theorem 12.3.8.

$$-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \xrightarrow{P} -2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)]\,.$$

$\mu_N \notin C(X_1)$ if and only if the right-hand side is positive.

Proof.

$$\begin{aligned}
&-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \hat{\mu}_N)] \\
&\quad = -2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \mu_N)] + 2N^{-1}[\ell(n_N, \hat{\mu}_N) - \ell(n_N, \mu_N)]\,.
\end{aligned}$$

As in Theorem 12.3.6,

$$2N^{-1}[\ell(n_N, \hat{\mu}_N) - \ell(n_N, \mu_N)] \xrightarrow{P} 0,$$

so we need only investigate the behavior of

$$\begin{aligned}
&-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \mu_N)] \\
&\qquad\qquad = -2N^{-1}[n_N'(\hat{\mu}_{1N} - \mu_N) - J'(\hat{m}_{1N} - m_N)]\,.
\end{aligned}$$

From Theorem 12.3.1, $N^{-1}n_N \xrightarrow{P} m^*$. As in the proof of Lemma 12.3.2, $\hat{\mu}_{1N} - \mu_N = \hat{\mu}_1\big(N^{-1}n_N\big) - \mu^*$ and $N^{-1}(\hat{m}_{1N} - m_N) = \hat{m}_1\big(N^{-1}n_N\big) - m^*$. By the continuity of $\hat{m}_1(\cdot)$ and $\hat{\mu}_1(\cdot)$ (ensured by the Implicit Function Theorem), $\hat{m}_1\big(N^{-1}n_N\big) \xrightarrow{P} \hat{m}_1(m^*)$ and $\hat{\mu}_1\big(N^{-1}n_N\big) \xrightarrow{P} \hat{\mu}_1(m^*)$, so

$$\begin{aligned}
&-2N^{-1}[\ell(n_N, \hat{\mu}_{1N}) - \ell(n_N, \mu_N)] \\
&\xrightarrow{P} \quad -2[m^{*\prime}(\hat{\mu}_1(m^*) - \mu^*) - J'(\hat{m}_1(m^*) - m^*)] \\
&= \quad -2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)]\,.
\end{aligned}$$

Since $\hat{\mu}(m^*) = \mu^*$, $\ell(m^*, \mu^*)$ is the unique maximum of $\ell(m^*, \mu)$ for $\mu \in C(X)$. Since $\hat{\mu}_1(m^*)$ is in $C(X)$, if $\hat{\mu}_1(m^*) \neq \mu^*$,

$$-2[\ell(m^*, \hat{\mu}_1(m^*)) - \ell(m^*, \mu^*)] > 0.$$

This occurs whenever $\mu^* \notin C(X_1)$ because $\hat{\mu}_1(m^*) \in C(X_1)$. Finally, $\mu^* \notin C(X_1)$ if and only if $\mu_N \notin C(X_1)$. □

13
Bayesian Binomial Regression

Standard methods for analyzing binomial regression data rely on asymptotic inferences. Bayesian methods performed using simple computations apply for any sample size. We discuss Bayesian inferences for binomial regression with an emphasis on inferences for the probability of "success." Furthermore, we illustrate diagnostic tools, perform model selection among non-nested models, and examine the sensitivity of the Bayesian methods. This chapter is closely related to Bedrick, Christensen, and Johnson (1997) and to earlier drafts of that article.

Section 1 introduces Bayesian binomial regression. Section 2 discusses standard Bayesian inference procedures with an emphasis on the predictive distribution. Section 3 presents Bayesian diagnostics including influence measures, global model checking methods, and a procedure for selection of the appropriate link function. Section 4 discusses computations.

13.1 Introduction

The purpose of this chapter is to illustrate the simplicity of a fully Bayesian approach to binomial regression models. Historically, it has been difficult to specify realistic prior information on regression coefficients in nonlinear models, and computations for inference and diagnostics were difficult due to intractable integrations. These difficulties no longer exist. We illustrate a fairly complete analysis for two data sets using methods that are simple and easy to apply. In particular, we discuss a method for specifying the

prior distribution that focuses on binomial probabilities, rather than esoteric regression coefficients. For computations, we focus on Monte Carlo methods because of their flexibility and their ease of implementation. We show how Monte Carlo sampling is used for prediction, making inferences on regression coefficients and probabilities, diagnostics, model checking, link selection, and sensitivity analysis of the prior.

Leonard (1972) first discussed Bayesian hierarchical models for binomial data. Zellner and Rossi (1984) gave a overview of Bayesian methods for binomial regression models. Johnson and Geisser (1985) and Johnson (1985) introduced general Bayesian predictive and estimative case deletion diagnostics that apply to binomial regression. We integrate these ideas along with Box's (1980) work on model checking to provide a variety of tools appropriate for analyzing binomial response data.

Consider regression data (y_i, x_i'), $i = 1, \ldots, n$, where the x_i's are known k vectors of covariates and the y_i's are independent binomial random variables with N_i trials. The probability of success p for any *single* trial y with covariate x is $F(x'\beta)$, i.e., $F(x'\beta) \equiv p \equiv \Pr(y = 1|x, \beta)$. Here, the vector β is an unknown k vector of regression coefficients. Although the function $F(\cdot)$ could be an arbitrary cdf, we consider logistic, probit, and complementary log-log regression models in which $F(x'\beta)$ is modeled as one of

$$F(x'\beta) = \begin{cases} e^{x'\beta} \Big/ \left[1 + e^{x'\beta}\right] & \text{Logistic} \\ \Phi(x'\beta) & \text{Probit} \\ 1 - \exp\left[-e^{x'\beta}\right] & \text{Complementary log-log} . \end{cases}$$

Here, $\Phi(u)$ is the cdf of a standard normal distribution. The success probability p is related to β through $F^{-1}(p) = x'\beta$, which is the link function from Chapter 9. For the logistic, probit, and complementary log-log models, $F^{-1}(p) = \log\{p/(1-p)\}$, $\Phi^{-1}(p)$, and $\log\{-\log(1-p)\}$, respectively. The likelihood function for the complete data $Y = (y_1, \ldots, y_n)'$ is

$$L(\beta|Y) \equiv \prod_{i=1}^{n} L(\beta|y_i) \equiv \prod_{i=1}^{n} \binom{N_i}{y_i} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{N_i - y_i}. \quad (1)$$

For a prior distribution on β, say $\pi(\beta)$, obtaining posterior and predictive distributions requires computing the posterior of β,

$$\pi(\beta|Y) = \frac{L(\beta|Y)\pi(\beta)}{\int L(\beta|Y)\pi(\beta)d\beta}.$$

Most interesting aspects of a Bayesian analysis can be obtained from various integrals involving this posterior density. Integrals involving $\pi(\beta|Y)$ are intractable, so we must use approximations.

Monte Carlo methods yield a discrete approximation to the posterior distribution that takes values β^r with probability $\tilde{q}_r$, $r = 1, \ldots, t$. Methods

for obtaining a discrete approximation are discussed in Section 4. Given a function $h(\beta)$, the posterior expectation $\mathrm{E}\{h(\beta) \mid Y\}$ is approximated by

$$\int h(\beta)\pi(\beta|Y)d\beta \doteq \sum_{r=1}^{t} h(\beta^r)\tilde{q}_r. \tag{2}$$

Typically, the Strong Law of Large Numbers implies that the error in the approximation converges almost surely to zero as the simulation sample size t increases.

13.2 Bayesian Inference

13.2.1 Specifying the Prior and Approximating the Posterior

Bayesian inference requires the specification of a prior distribution $\pi(\beta)$. In the past, several methods of specifying priors for binomial regression problems have been used. The standard approach has been to assume either a normal distribution for β or the "noninformative" diffuse prior $\pi(\beta) = 1$. These are convenient in large sample situations where the posterior on β is approximately normal. See Zellner and Rossi (1984) for relevant discussion. Another type of prior focuses on the assessment of "success" probabilities for various choices of covariate values, rather than on the assessment of regression coefficients.

EXAMPLE 13.2.1. Consider a simple situation with $k = 2$. Imagine that we are recruiting statistics students into a graduate program. We will attempt to recruit from two populations: domestic students ($i = 1$) and international students ($i = 2$). If N_1 domestic students apply and N_2 international students apply, assuming independence of students we successfully recruit $y_1 \sim \mathrm{Bin}(N_1, p_1)$ domestic students and $y_2 \sim \mathrm{Bin}(N_2, p_2)$ international students. We can write a one-way ANOVA logit model

$$\log\{p_i/(1-p_i)\} = \mu + \alpha_i,$$

$i = 1, 2$. This model is overparameterized, so we impose the side condition $\alpha_1 = 0$ to make the model a logistic regression. We now have

$$\log\{p_1/(1-p_1)\} = \mu, \qquad \log\{p_2/(1-p_2)\} = \mu + \alpha_2.$$

The graduate advisor has specified prior distributions $p_1 \sim \mathrm{Beta}(4, 4)$ and $p_2 \sim \mathrm{Beta}(4, 1)$, reflecting (in part) the beliefs that about $80\% = \mathrm{E}(p_2) = 4/(4+1)$ of the international students and half, $[4/(4+4)]$, of the domestic students will be successfully recruited. The prior specification includes the assumption that p_1 and p_2 are independent. Having placed a joint distribution on p_1 and p_2, it is a calculus problem to determine the corresponding

distribution on the "regression" parameters μ and α_2. We discuss the exact procedure later. While we assumed that the distributions of p_1 and p_2 were independent, the approach can, in theory, be carried out with any joint distribution for p_1 and p_2. The problem is not in doing the calculus, but in specifying a realistic joint distribution when the independence assumption is not appropriate.

The independence assumption is a key part of the procedure. With p_1 and p_2 independent, if we were told the value of p_1, we should not be inclined to revise our thinking about p_2. That certainly seems reasonable if we are told that p_1 is something near its expected value .5. It seems less reasonable if we are told, say, that $p_1 \geq .95$. Knowing that $p_1 \geq .95$ would probably make us want to revise our distribution of p_2 to make larger values more probable. However, .95 is 2.7 prior standard deviations above the prior mean for p_1, so this event is extremely unlikely under the prior specification. If $p_1 \geq .95$ is more likely than the original prior specification allows, the entire prior should be recalibrated, at which point the independence assumption may be called in question. However, if, after reflection, those situations that might cause concern about the independence assumption are thought unlikely, then we believe the independence assumption is reasonable.

Lack of independence can also occur if the international students were thought to be very similar to the domestic students regardless of the behavior of the domestic students. In this case, knowing $\tilde{p}_1$ is highly informative about $\tilde{p}_2$ and our prior is not appropriate.

The main idea in Example 13.2.1 was to specify prior distributions for p_1 and p_2 rather than on the regression parameters μ and α_2. We do this because p_1 and p_2 have natural interpretations. In a simple logistic regression,

$$\log\{p/(1-p)\} = \beta_0 + \beta_1 \tau\,,$$

there are again only two regression parameters (β_0 and β_1), but there is no obvious choice for probabilities p_1 and p_2 at which to specify the prior. In such cases, we must pick two values, say $\tilde{\tau}_1$ and $\tilde{\tau}_2$, and specify prior distributions for $\tilde{p}_1$, the probability of success when $\tau = \tilde{\tau}_1$, and $\tilde{p}_2$, the probability of success when $\tau = \tilde{\tau}_2$.

EXAMPLE 13.2.2. *O-Ring Data.*
Consider fitting a simple regression model on temperature to the data in Table 2.1. Let p_i be the probability that any O-ring fails in case i and model this as $F^{-1}(p_i) = \beta_0 + \beta_1 \tau_i = x_i'\beta$, where τ_i is the temperature. Our prior is defined by giving independent distributions to the probabilities of O-ring failure at temperatures $\tilde{\tau}_1 = 55$ and $\tilde{\tau}_2 = 75$ degrees Fahrenheit. Write

$$\beta_0 + \beta_1 \tilde{\tau}_i = [1, \tilde{\tau}_i] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \tilde{x}_i'\beta$$

and define $\tilde{p}_1$ and $\tilde{p}_2$ by $\tilde{p}_i = F(\tilde{x}_i'\beta)$. The $\tilde{\tau}_i$'s should be chosen in the expected range of the observed temperatures but far enough apart so that information about the corresponding probabilities can be reasonably assumed independent. The selected temperatures should also be amenable to expert opinion. Our priors on $\tilde{p}_1$ and $\tilde{p}_2$ are Beta(1, .577) and Beta(.577, 1) respectively. The prior on $\tilde{p}_1$ was chosen because it has a "J" shape and gives $\Pr[\tilde{p}_1 > 1/2] = 2/3$. The prior on $\tilde{p}_2$ has a "J" shape and gives $\Pr[\tilde{p}_2 < 1/2] = 2/3$.

The prior on $\beta = (\beta_0,\ \beta_1)'$ is determined using the change-of-variables method. Under the logistic model, the prior on β is a data augmentation prior (DAP) in the sense that it has the same functional form as the likelihood function, i.e.,

$$\pi(\beta) \propto \prod_{i=1}^{2} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i},$$

where $\tilde{N}_1 = \tilde{N}_2 = 1.577$, $\tilde{y}_1 = 1$, and $\tilde{y}_2 = .577$. With this DAP, the prior on $\tilde{p}_1$ can be thought of as one prior O-ring failure out of 1.577 trials at $\tilde{\tau}_1 = 55$, and for $\tilde{p}_2$, it can be thought of as .577 prior O-ring failures out of 1.577 trials at $\tilde{\tau}_2 = 75$. The weight attached to the prior is equivalent to $\tilde{N}_1 + \tilde{N}_2$ "prior" observations, about 3. The posterior density for β also has the same functional form as the likelihood, i.e.,

$$\pi(\beta|Y) \propto \prod_{i=1}^{n} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{N_i - y_i} \prod_{i=1}^{2} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i}.$$

Many standard computer programs, e.g., GLIM and SPLUS, can be used to find the posterior mode β_M and an asymptotic dispersion matrix $\Sigma(\beta_M)$ for the posterior. To compute the mode, simply augment the observed data with a prior "binomial" observation at 55 degrees consisting of 1.577 trials and 1 observed O-ring failure and include a prior observation at 75 degrees with 1.577 trials and .577 O-ring failures. The posterior mode of β is the maximum likelihood estimate (MLE) from the augmented data. The asymptotic covariance matrix computed from the augmented data is the asymptotic dispersion matrix for the posterior. These quantities are of interest in themselves and can also be used to create a good discrete approximation to the posterior.

Figures 13.1 and 13.2 give contour plots of the prior and posterior distributions on β, respectively. Note the high correlation between β_0 and β_1 in both the prior and the posterior. The posterior exhibits appreciable skewness, with longer tails in the direction of small slopes and large intercepts. The high correlation between β_0 and β_1 is largely eliminated if we standardize the temperature to have mean zero, i.e., if we change the model to $\text{logit}(p_i) = \beta_0 + \beta_1(\tau_i - \bar{\tau}.)$. For some problems, this may be preferable to ease the computational burden. As there were no computational difficulties

with these data, and as prediction and model validation are independent of the regression parameterization, we consider only the original version of the model.

In general, we derive the prior on β for a model with k regression parameters from a prior elicited on success probabilities $\tilde{p}_i$ at k suitably selected predictor vectors $\tilde{x}_i$. We place independent Beta$(\tilde{y}_i, \tilde{N}_i - \tilde{y}_i)$ priors on the $\tilde{p}_i$, regardless of the choice of the link function. For an arbitrary link function, the induced prior on β has the form

$$\pi(\beta) \propto \prod_{i=1}^{k} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i - 1} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i - 1} f(\tilde{x}_i'\beta),$$

where $f(\cdot)$ is the first derivative of the function $F(\cdot)$. In the case of logistic regression,

$$\pi(\beta) \propto \prod_{i=1}^{k} [F(\tilde{x}_i'\beta)]^{\tilde{y}_i} [1 - F(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{y}_i}, \tag{1}$$

which has the same form as the likelihood function. Therefore, (1) is a data augmentation prior (DAP), so named because the likelihood times the prior has the form of a likelihood with additional "prior" data $(\tilde{y}_i, \tilde{N}_i)$, $i = 1, \ldots, k$. In other words, for the logistic model we can think of the parameters of the prior distribution as a prior sample size $\tilde{N}_i$ and a prior number of successes $\tilde{y}_i$ corresponding to the vector of predictors $\tilde{x}_i$.

Incidentally, this procedure can also be executed with priors for the $\tilde{p}_i$'s other than betas. In fact, with different link functions, different distributions on the $\tilde{p}_i$'s lead to different DAPs. (Note that the likelihood depends on the link function, so DAPs depend on the link function.)

We now consider our primary example.

EXAMPLE 13.2.3. *Trauma Data.*
We analyze data on a randomly selected subset of 300 patients admitted to the University of New Mexico Trauma Center between the years 1991 and 1994. For each patient, we have their injury severity score (ISS), their revised trauma score (RTS), their AGE, the type of injuries (TI), that is, whether they were blunt ($TI = 0$), e.g., the result of a car crash, or penetrating ($TI = 1$), e.g., gunshot wounds, and the dependent variable, whether the patient eventually survived the injuries. The ISS is an overall index of a patient's injuries based on the approximately 1300 injuries catalogued in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with severe injuries in three or more body areas. The RTS is an index of physiologic injury and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale. The RTS takes on values from 0 for a patient with no vital signs to 7.84 for a patient with

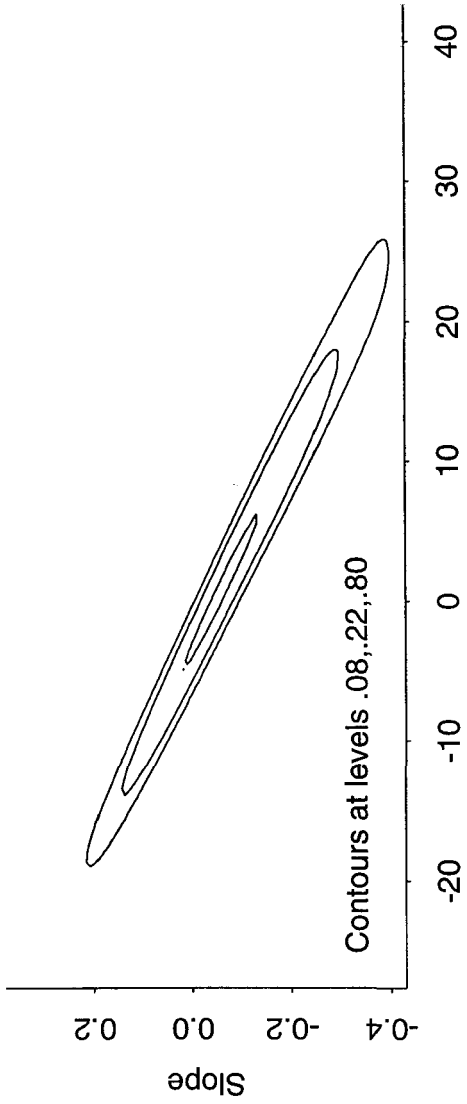


FIGURE 13.1. O-Ring Data: Prior on β

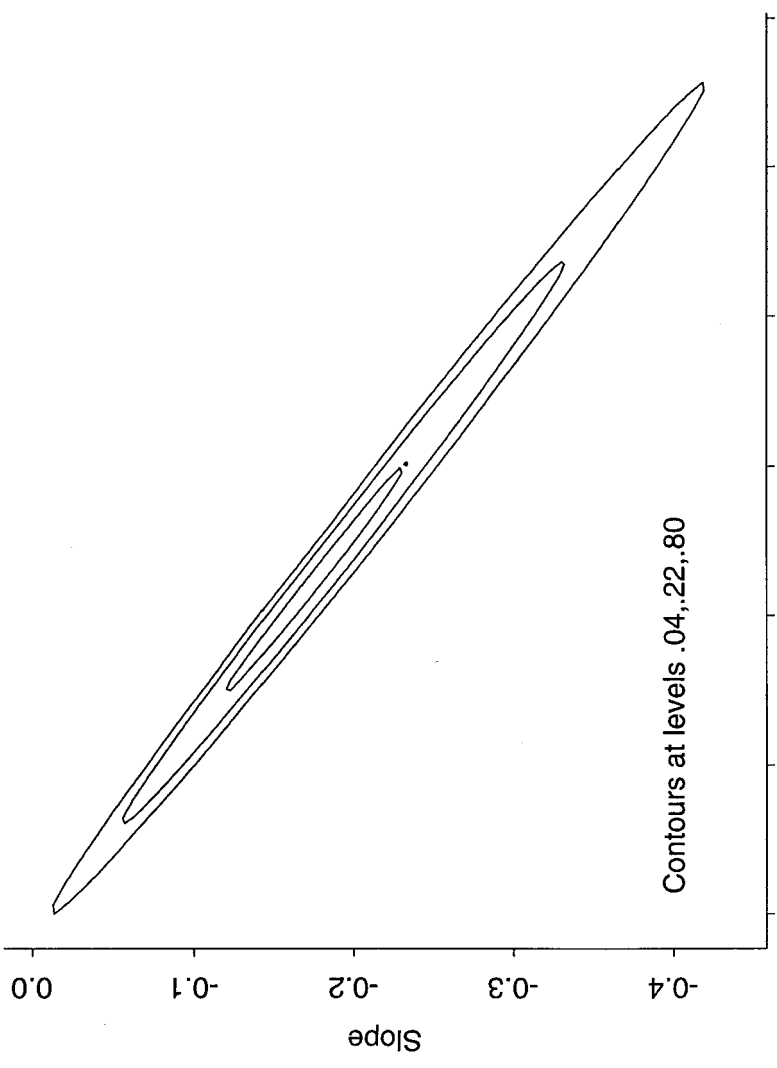


FIGURE 13.2. O-Ring Data: Posterior on β

normal vital signs. The data are available electronically from STATLIB as well as from my web homepage:

`http://stat.unm.edu/~fletcher`

Additional information is given in the Preface.

Figure 13.3 gives side-by-side boxplots comparing the 278 survivors and 22 fatalities on RTS, ISS, and AGE. Seventeen of the 225 patients with blunt injuries died. Five of the 75 patients with penetrating injuries died.

The data were provided by Dr. Turner Osler, a trauma surgeon at the University of Vermont and former head of the Burn Unit at the University of New Mexico Trauma Center. Dr. Osler proposed a logistic regression model to estimate the probability of a patient's death using an intercept and predictors ISS, RTS, patient's AGE (used as a surrogate for physiologic reserve), TI, and an interaction between AGE and TI. Similar logistic models are used by trauma centers throughout the United States. Dr. Osler's expert opinions formed the basis for our prior distribution.

To induce a proper prior distribution on the $k = 6$ dimensional vector β, we require a joint distribution on death probabilities for 6 sets of conditions $\tilde{x}_i' = (1, ISS_i, RTS_i, AGE_i, TI_i, AGE_i \times TI_i)$. Based on discussions with our expert and two-dimensional plots of the data, we defined a 2^4 factorial design having ISS at levels 25 and 41, RTS at levels 3.34 and 7.84, AGE at levels 10 and 60, and TI at levels 0 and 1. The idea was to pick values of the variables that were relatively extreme within the data but still had substantial probabilities for both success and failure. The prior conditions were chosen as a 1/4 replicate of this 2^4 with two center points. However, the center points were taken to be values that could actually exist — none of ISS, RTS, and TI are truly continuous variables. In fact, TI is a binary variable, so one "center point" was taken with $TI = 0$ and the other with $TI = 1$. Bedrick, Christensen, and Johnson (1996) (henceforth referred to as BCJ) recommend calculating the condition number of the matrix $\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_k)'$ to ascertain that the chosen $\tilde{x}_i$'s are not too close or too far apart. See Belsley (1991) for discussion of condition numbers. Beta priors were found to be suitable for the $\tilde{p}_i$'s with parameters given in Table 13.1. Figure 13.4 gives plots of the priors on the $\tilde{p}_i$'s as well as the posteriors. The priors are generally consistent with the posteriors. Relative to the amount of data, the priors are not overwhelming, being the equivalent of $57.5 = \sum_{i=1}^{6} \tilde{N}_i$ observations compared to 300 data points. (The posterior densities were obtained by sampling from the discrete approximate posterior and smoothing the samples.)

Our initial discussion with Dr. Osler involved eliciting 1st, 50th, and 99th percentiles for each $\tilde{p}_i$. These actually overspecify a beta distribution. We wrote a computer program to find the beta distributions that most nearly satisfied the specifications, plotted these distributions, and validated them with our expert.

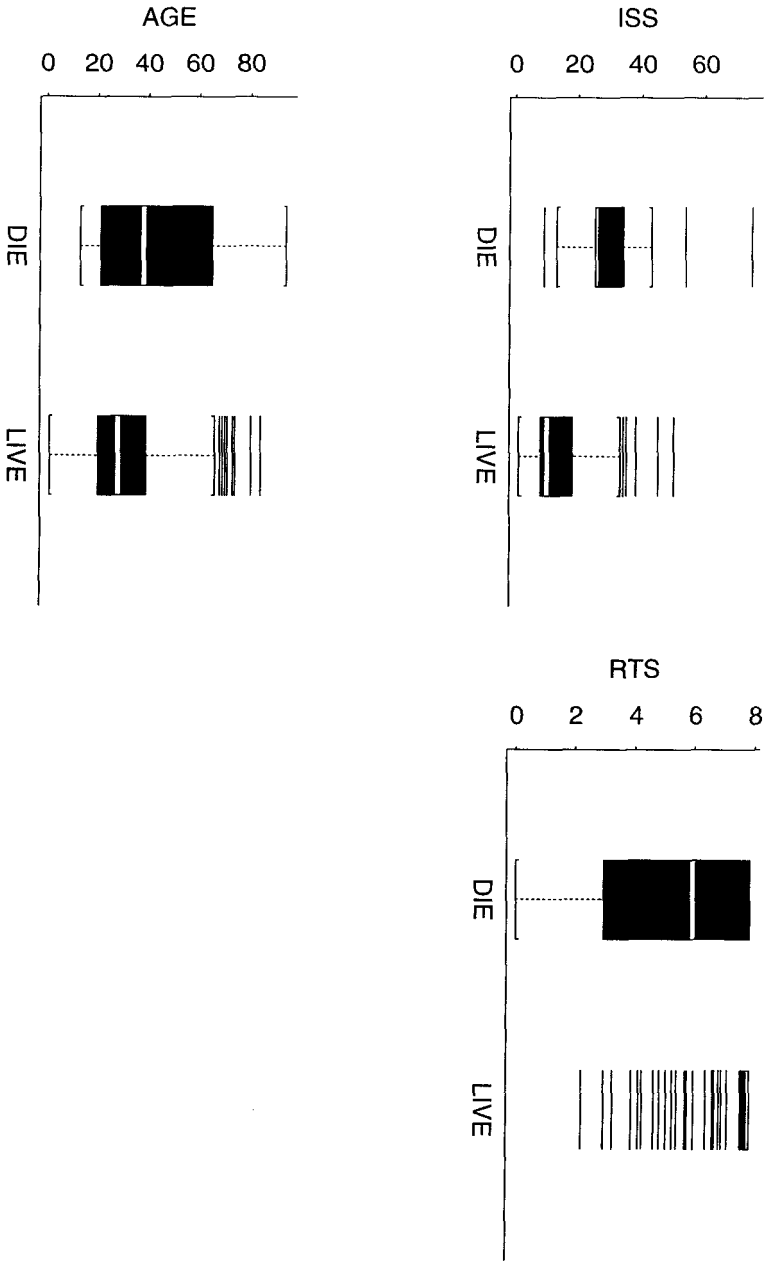


FIGURE 13.3. Trauma Data: Box Plots

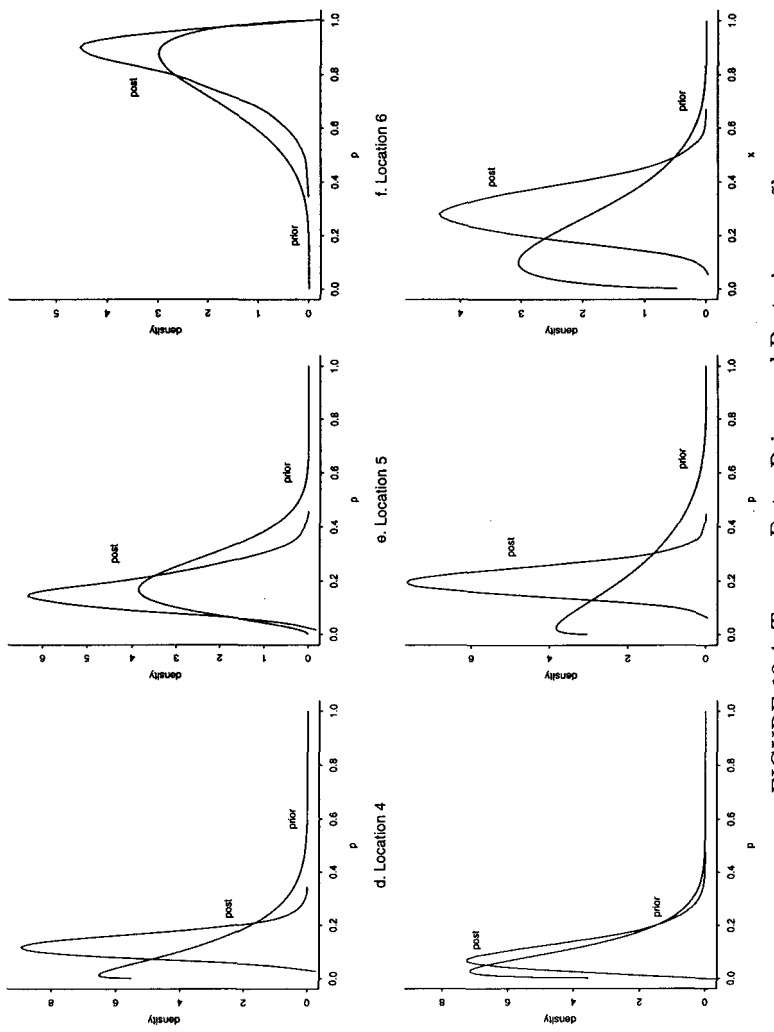


FIGURE 13.4. Trauma Data: Priors and Posteriors on $\tilde{p}$'s

TABLE 13.1. Trauma Data: Prior Specification

	Design for Prior						Beta $(\tilde{y}_i, \tilde{N}_i - \tilde{y}_i)$	
i	$\tilde{x}_i'$						$\tilde{y}_i$	$\tilde{N}_i - \tilde{y}_i$
1	1	25	7.84	60	0	0	1.1	8.5
2	1	25	3.34	10	0	0	3.0	11.0
3	1	41	3.34	60	1	60	5.9	1.7
4	1	41	7.84	10	1	10	1.3	12.0
5	1	33	5.74	35	0	0	1.1	4.9
6	1	33	5.74	35	1	35	1.5	5.5

The first probability $\tilde{p}_1$ corresponds to an individual that "has good physiology, is 'not bad hurt,' does not have a lot of reserve," and for whom there is "added uncertainty due to age." The Beta(1.1, 8.5) suitably reflects Dr. Osler's uncertainty about $\tilde{p}_1$. The median of his prior is around .09. The second type of individual "has bad physiology, is very ill, but is young and resilient and is not so bad hurt." The prior for $\tilde{p}_2$ is Beta(3, 11) with median around .20. Incidentally, "bad physiology" and "very ill" apparently refer to bad RTS scores, while how badly hurt one is relates to ISS. The third individual has "bad physiology, a pretty bad injury, and there is much more uncertainty here due to the age factor." The prior is Beta(5.9, 1.7) with median around .8. Prior individual four "is young, resilient, and has a big injury." The prior is Beta(1.3, 12) with a median of around .07.

Dr. Osler had more difficulty with the 5th and 6th types of individuals because their conditions were both less extreme and more related than those already considered. The priors for $\tilde{p}_5$ and $\tilde{p}_6$ are Beta(1.1,4.9) with approximate median .15, and Beta(1.5, 5.5) with approximate median .19, respectively.

The assumption of independence seems reasonable with the possible exception of $\tilde{p}_5$ and $\tilde{p}_6$. If our expert were told that $\tilde{p}_5 = .3$, he would definitely want to revise his probability for $\tilde{p}_6$ upward. This is because he is fairly confident that the difference between these two probabilities, $\tilde{p}_6 - \tilde{p}_5$, is positive but reasonably small, while he is less certain about the magnitude of the probabilities themselves. Having $\tilde{p}_6 - \tilde{p}_5$ small but positive is reflecting his perception that penetrating injuries are worse than blunt ones but not a lot worse.

Because of our concern about possible lack of independence for the two values $\tilde{p}_6$ and $\tilde{p}_5$ only, we also considered a prior in which the information about $\tilde{p}_6$ was left out of the specification. This results in a partially informative prior (see BCJ, Sec. 4.1, for a full discussion) which is an improper DAP using five prior observations instead of the six required for a proper DAP. We found that all statistical inferences were essentially the same for the two priors, so we have presented results only for the full prior.

Finally, it should be pointed out that the process of coming up with a prior is very much a collaboration between the expert and the statisticians.

The judgement and expertise of both are needed. It is also quite a bit of work for everyone involved.

Bedrick, Christensen, and Johnson (1996, 1997) give further details on this approach to specifying priors for regression problems, including discussions of priors with order restrictions on the $\tilde{p}_i$'s and partial prior information. As mentioned above, a particularly useful form of partial prior information is specifying $k' < k$ values $\tilde{p}_i$.

The genesis of this approach lies with Tsutakawa (1975), Tsutakawa and Lin (1986), and Grieve (1988) who considered independent prior distributions on two probabilities of "success" in simple linear binomial regression problems. Tsutakawa and Lin (1986) argued that eliciting information about success probabilities should be much easier than eliciting information about regression coefficients, a position with which we heartily agree. This is clearly true if one entertains the possibility of two or more models, such as logistic regression versus probit regression. The regression coefficients for these two models require separate elicitations, whereas if one has elicited a prior for probabilities, it is straightforward to induce the requisite prior on β for either model.

BCJ extended the Tsutakawa approach to generalized linear models (GLMs) with multiple covariates. For a hypothetical observation $\tilde{y}_i$ with covariate vector $\tilde{x}_i$, BCJ specify a prior on the mean value $\mathrm{E}(\tilde{y}_i|\tilde{x}_i)$. This is done for k locations $\tilde{x}_i$, $i = 1, \ldots, k$, where k is the common dimension of the $\tilde{x}_i$'s. The prior on the regression coefficient vector β is induced by transforming the distribution on the $\mathrm{E}(\tilde{y}_i|\tilde{x}_i)$'s into a distribution on β. BCJ call such priors conditional means priors (CMPs) and elaborate on the approach in considerable detail. The conditional means provide parameters that are more intuitive than regression coefficients and thus easier to specify prior information for. BCJ also make connections between CMPs and DAPs. (Note that to make the GLM approach apply to binomial regression, just as in Chapter 9, the y_i's have to be defined as binomial proportions rather than our usual binomial counts.)

A key feature in this approach is assuming prior independence of the $\mathrm{E}(\tilde{y}_i|\tilde{x}_i)$'s. This assumption might be unreasonable if the $\tilde{x}_i$'s are "too close" together (cf. Grieve, 1988). There are also technical difficulties if they are "too far apart." BCJ (Sec. 5) examined these issues in detail.

13.2.2 Predictive Probabilities

The predictive probability of success in one new trial y with covariate x is

$$\Pr(y = 1|Y, x) = \mathrm{E}[F(x'\beta)|Y, x] = \int F(x'\beta)\pi(\beta|Y)\, d\beta. \qquad (2)$$

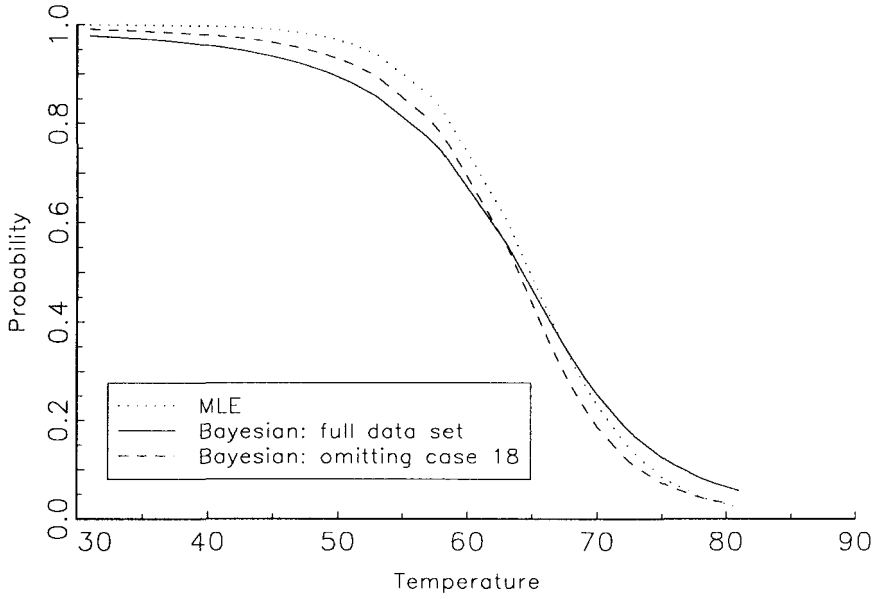


FIGURE 13.5. O-Ring Data: Predictive Probabilities and MLEs

Using the discrete approximation to the posterior gives

$$\Pr(y = 1|Y, x) \doteq \sum_{r=1}^{t} F(x'\beta^r)\tilde{q}_r.$$

Unless specifically stated otherwise, the examples henceforth use logistic models.

EXAMPLE 13.2.2 CONTINUED. Figure 13.5 presents the Bayesian predictive probability of O-ring failure, $\Pr(y = 1|Y, x)$, and the MLE of the probability of an O-ring failure, $F(x'\hat{\beta}_{ml})$, as temperature varies from 30 degrees to 80 degrees. Predictive probabilities are less than the MLEs for temperatures below about 67 and are greater for larger temperatures.

The predictive probability of "success" can be interpreted in two ways. It is the subjective probability of at least one O-ring failure on the next flight at the given temperature. It is also the Bayes estimate of the proportion of flights at the given temperature in which there would be at least one O-ring failure. With the second interpretation, one may be interested in interval estimates. Geisser (1982) established that posterior interval estimates for probabilities can be viewed as (asymptotic) prediction intervals for the proportion of successes from a large number of future trials.

Figure 13.6 contains a plot of the predictive probabilities of O-ring failure as x varies from 30 to 80 degrees along with 90% interval estimates. In other words, it gives $\mathrm{E}[F(x'\beta)|Y, x]$ and a Bayesian posterior interval estimate based on our subjective prior. The Bayesian interval is obtained

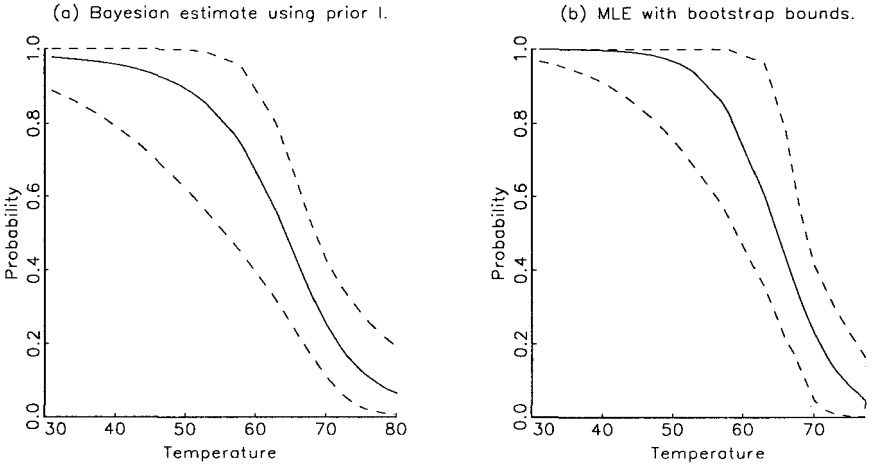


FIGURE 13.6. O-Ring Data: Predictive Probabilities and 90% Intervals

by determining the 5th and 95th percentiles of the approximate posterior distribution for $F(x'\beta)$, i.e., the distribution that takes values $F(x'\beta^r)$ with probability $\tilde{q}_r$. For low temperatures, the posterior distribution of $F(x'\beta)$ is highly skewed to the left; thus, the mean $\mathrm{E}[F(x'\beta)|Y, x]$ is lower than the median.

EXAMPLE 13.2.3 CONTINUED. Figure 13.7 presents predictive probabilities of death as a function of ISS for blunt and penetrating traumas. These are given for various values of RTS and AGE. Note that for 60-year-olds, there is essentially no difference in the probability of death due to blunt or penetrating injury. However for 10-year-olds, the probability of death is higher for a penetrating injury.

13.2.3 Inference for Regression Coefficients

The posterior mean $\mathrm{E}(\beta|Y)$ and covariance matrix

$$\mathrm{Cov}(\beta|Y) = \mathrm{E}(\beta\beta'|Y) - \mathrm{E}(\beta|Y)\mathrm{E}(\beta|Y)'$$

of the regression coefficients are approximated by $\hat{\beta} = \sum_{r=1}^{t} \beta^r \tilde{q}_r$ and

$$\widehat{\mathrm{Cov}}(\beta|Y) = \left[\sum_{r=1}^{t} \beta^r \beta^{k\prime} \tilde{q}_r\right] - \hat{\beta}\hat{\beta}',$$

respectively. The cdf of β_j $(j = 1, \ldots, k)$ and histograms for approximating the marginal posterior density can be obtained from probabilities of the

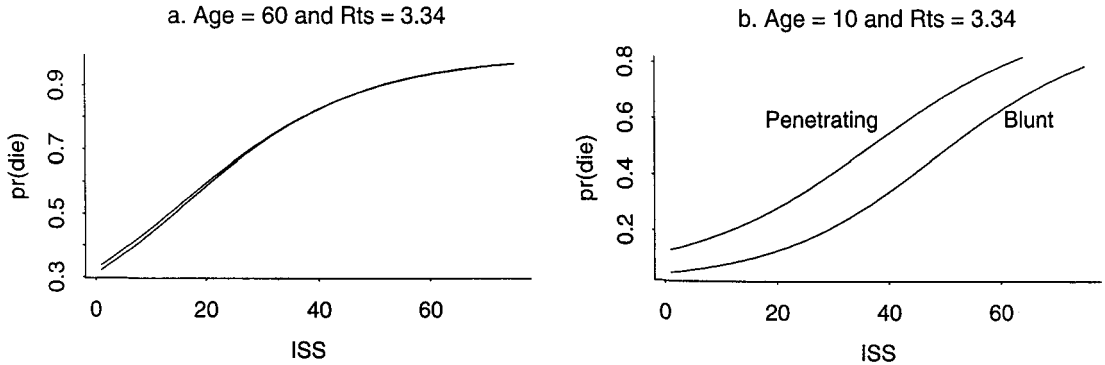


FIGURE 13.7. Trauma Data: Predictive Probabilities

form

$$\Pr(a < \beta_j \le b|Y) = \int I_{(a,b]}(\beta_j)\pi(\beta|Y)d\beta \doteq \sum_{r=1}^{t} I_{(a,b]}(\beta_j^r)\tilde{q}_r$$

where $I_{(a,b]}(\beta_j)$ is 1 if $a < \beta_j \le b$ and 0 otherwise.

EXAMPLE 13.2.2 CONTINUED. Table 13.2 presents posterior means, standard deviations, and percentiles of β_0 and β_1 for the O-ring data. Using our prior, $\Pr(\beta_1 < 0|Y) > .99$, which suggests the slope is not zero. Figure 13.8 gives the Bayesian marginal posterior density for β_1 in the O-ring data. As before, this was actually generated by smoothing 5000 samples from the approximate posterior distribution, i.e., using Rubin's (1987) SIR algorithm.

TABLE 13.2. Posterior Marginal Distribution: O-Rings

	Full Data		Case 18 Deleted	
	β_0	β_1	β_0	β_1
$\hat{\beta}_i = \mathrm{E}(\beta_i\|Y)$	12.97	−.2018	16.92	−.2648
Std. Dev.$(\beta_i\|Y)$	5.75	.0847	7.09	.1056
5%	4.56	−.355	6.85	−.459
25%	9.04	−.251	11.98	−.324
50%	12.44	−.194	16.13	−.252
75%	16.20	−.144	20.86	−.191
95%	23.38	−.077	29.96	−.114

EXAMPLE 13.2.3 CONTINUED. Table 13.3 presents posterior means, standard deviations, and percentiles for the β_j's from the trauma data along with the maximum likelihood estimates, and asymptotic standard errors as well as posterior summaries obtained from the diffuse prior $\pi(\beta) = 1$. In addition, the informative prior gives $\Pr(\beta_1 > 0|Y) > .99$, which suggests that the coefficient of ISS is not zero. Recall that low values of RTS are bad for the patient, so the tendency of the RTS coefficients to be negative is reasonable. Central 90% posterior intervals for the β_j's are about 3/4's as wide using the informative prior as with the diffuse prior.

13.2.4 Inference for LD_α

With the O-ring data, it is of interest to estimate the temperature at which the chance of O-ring failure is, say 50%, or some other prespecified amount α. This percentile is often called the LD_α in bioassay problems (LD for "lethal dose"), and satisfies $LD_\alpha = \{F^{-1}(\alpha) - \beta_0\}/\beta_1$. The LD_α is a function of the vector β, so its approximate posterior distribution is easily

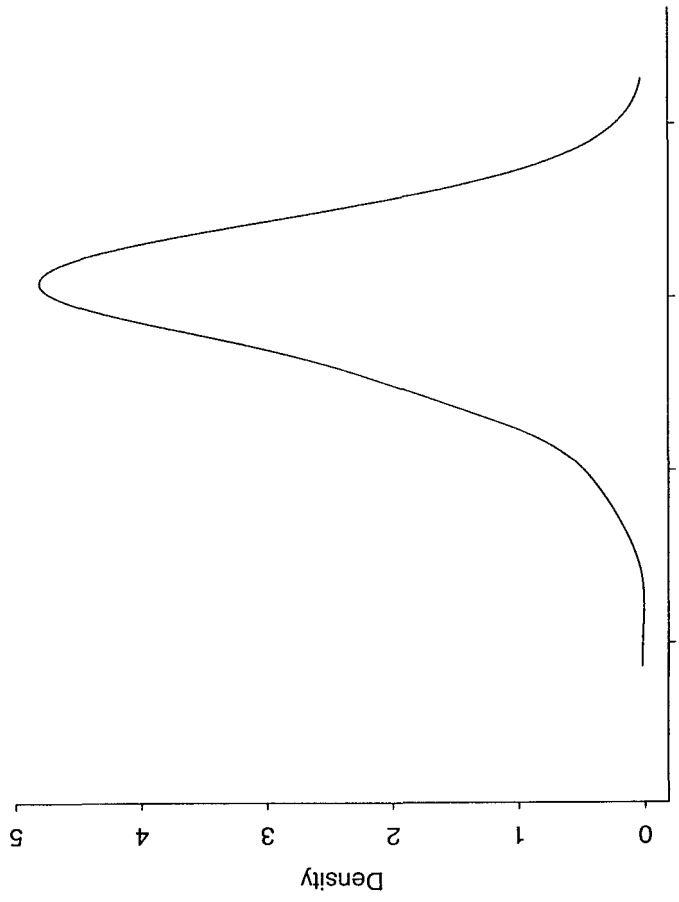


FIGURE 13.8. O-Ring Data: Marginal Density for β_1

TABLE 13.3. Fitted Trauma Model

	Informative Posterior Summaries Based on informative prior				Maximum Likelihood	
Variable	Estimate	Std. Error	.05%	.95%	Estimate	Std. Error
Intercept	−1.79	1.10	−3.54	.02	−2.73	1.62
ISS	.07	.02	.03	.10	.08	.03
RTS	−.60	.14	−.82	−.37	−.55	.17
AGE	.05	.01	.03	.07	.05	.01
TI	1.10	1.06	−.66	2.87	1.34	1.33
AGE × TI	−.02	.03	−.06	.03	−.01	.03
	Posterior Summaries Based on diffuse prior					
Variable	Estimate	Std. Error	.05%	.95%		
Intercept	−2.81	1.60	−5.34	−.18		
ISS	.09	.03	.05	.13		
RTS	−.59	.17	−.86	−.32		
AGE	.06	.02	.03	.09		
TI	1.46	1.36	−.79	3.69		
AGE × TI	−.01	.03	−.07	.05		

obtained. The approximate posterior takes on the value $\{F^{-1}(\alpha) - \beta_0^r\}/\beta_1^r$ with probability $\tilde{q}_r$.

Table 13.4 presents the posterior median and central 90% intervals for LD_α using five values of α for the O-ring data. In particular, the Bayesian analysis gives 69.8 degrees as the posterior median temperature at which the chance of O-ring failure is .25. The tails of the LD_α's are very heavy due to a non-negligible probability of getting β_1 values near zero.

TABLE 13.4. Posterior Summaries for LD_α's

	Full Data Percentiles				Case 18 Deleted Percentiles		
α	5%	50%	95%	α	5%	50%	95%
.90	30.2	52.9	60.4	.90	39.8	55.1	61.2
.75	43.4	58.5	64.0	.75	48.9	59.4	64.0
.50	55.9	64.2	68.5	.50	57.5	63.8	67.5
.25	65.1	69.8	76.4	.25	64.1	68.1	73.0
.10	70.3	75.4	88.3	.10	68.3	72.4	80.9

13.3 Diagnostics

In this section, we examine a variety of influence diagnostics based on deleting cases. We also explore Box's (1980) method of model checking. Finally,

we consider the choice of an appropriate link function and an associated case deletion diagnostic.

13.3.1 Case Deletion Influence Measures

Case deletion diagnostics were pioneered by Cook (1977), Belsley, Kuh and Welsch (1980), and Pregibon (1981). Johnson and Geisser (1982, 1983, 1985) introduced Bayesian predictive and estimative case deletion diagnostics for the linear model and Johnson (1985) introduced diagnostics for the estimation of probabilities in logistic regression. Here we present the Johnson-Geisser influence measures for this nonlinear Bayesian setting. Our purpose is to detect those cases that, upon deletion from the data, noticeably affect inferences. For example, if the predictive probability of O-ring failure were to change radically upon deletion of a single case, it is incumbent upon us to report and quantify that fact. It may or may not be appropriate to delete such cases in a final analysis.

The effect of case deletion on the posterior of β is easily formulated. Recalling (13.1.1), the likelihood for β based on all the data except y_i is

$$L(\beta|Y_{(i)}) = \frac{L(\beta|Y)}{L(\beta|y_i)}$$

where $Y_{(i)}$ denotes the data Y with y_i deleted. It follows that

$$\pi(\beta|Y_{(i)}) = \frac{L(\beta|Y_{(i)})\pi(\beta)}{\int L(\beta|Y_{(i)})\pi(\beta)d\beta} = \frac{\pi(\beta|Y)/L(\beta|y_i)}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta}. \tag{1}$$

If we renormalize the probability weights in our discrete approximation,

$$\tilde{q}_{r(i)} = \frac{\tilde{q}_r/L(\beta^r|y_i)}{\sum_{k=1}^t \tilde{q}_k/L(\beta^k|y_i)},$$

then the distribution taking values β^r with probability $\tilde{q}_{r(i)}$ gives a discrete approximation to the posterior (1). Expectations with respect to $\pi(\beta|Y_{(i)})$ are evaluated using this approximate distribution.

Estimative Influence

Kullback-Leibler (KL) divergences can be used as in Johnson and Geisser (1985) and Pettit and Smith (1985) to measure the discrepancy between full and reduced data posteriors. The KL divergence with respect to the posterior density with the ith case deleted is defined as

$$D_{1i}^{\beta} \equiv \int \log\left[\frac{\pi(\beta|Y_{(i)})}{\pi(\beta|Y)}\right] \pi(\beta|Y_{(i)})d\beta \geq 0.$$

A large value of D_{1i}^{β} indicates that deletion of case i results in a different posterior for β than if it were retained, possibly resulting in different inferences for β.

We now present a computational formula for D_{1i}^{β}. The predictive probability that a future binomial observation y with covariate vector x_i equals the observed y_i value, given $Y_{(i)}$, can be expressed in two equivalent ways:

$$\Pr(y = y_i|Y_{(i)}, x_i) = \int L(\beta|y_i)\pi(\beta|Y_{(i)})d\beta = \frac{L(\beta|y_i)\pi(\beta|Y_{(i)})}{\pi(\beta|Y)}. \quad (2)$$

To see this, note that from (1),

$$\frac{L(\beta|y_i)\pi(\beta|Y_{(i)})}{\pi(\beta|Y)} = \frac{1}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta}.$$

Also, from (1),

$$\int L(\beta|y_i)\pi(\beta|Y_{(i)})d\beta = \frac{\int L(\beta|y_i)\pi(\beta|Y)/L(\beta|y_i)d\beta}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta} = \frac{1}{\int \pi(\beta|Y)/L(\beta|y_i)d\beta}.$$

Equation (2) gives

$$\begin{aligned} D_{1i}^{\beta} &= \int \log\left[\frac{\Pr(y = y_i|Y_{(i)}, x_i)}{L(\beta|y_i)}\right]\pi(\beta|Y_{(i)})d\beta \\ &= \log\Pr(y = y_i|Y_{(i)}, x_i) - \int \log L(\beta|y_i)\pi(\beta|Y_{(i)})d\beta \\ &\doteq \log\left\{\sum_{r=1}^{t} L(\beta^r|y_i)\tilde{q}_{r(i)}\right\} - \sum_{r=1}^{t} \log L(\beta^r|y_i)\tilde{q}_{r(i)}. \end{aligned}$$

The KL divergence with respect to the posterior based on all observations is defined as

$$D_{2i}^{\beta} \equiv \int \log\left[\frac{\pi(\beta|Y)}{\pi(\beta|Y_{(i)})}\right]\pi(\beta|Y)d\beta.$$

Using equation (2),

$$D_{2i}^{\beta} \doteq \sum_{r=1}^{t} \log L(\beta^r|y_i)\tilde{q}_r - \log\left\{\sum_{r=1}^{t} L(\beta^r|y_i)\tilde{q}_{r(i)}\right\}.$$

The symmetric divergence is defined to be the sum of the divergences for the deleted and full posteriors, $D_i^{\beta} \equiv D_{1i}^{\beta} + D_{2i}^{\beta}$.

Predictive Influence

The predictive distribution for a single trial is Bernoulli, i.e., takes on the values 0 and 1. The symmetric KL divergence is used to measure the discrepancy between full and reduced data predictive distributions. The symmetric KL divergence between two Bernoulli distributions with probabilities p and q reduces to

$$J(p,q) \equiv (p-q)\log\left(\frac{p(1-q)}{(1-p)q}\right).$$

As in Johnson (1985), we define a symmetric predictive divergence diagnostic for predicting new observations at the original data locations when case i is deleted:

$$D_i^p \equiv \sum_{j=1}^{n} J\big(\Pr(y=1|Y,x_j), \Pr(y=1|Y_{(i)},x_j)\big).$$

Here, $\Pr(y=1|Y,x)$ is the predictive probability of success from all the data as defined in (13.2.2), and $\Pr(y=1|Y_{(i)},x)$ is the predictive probability of success based on all the data except case i:

$$\Pr(y=1|Y_{(i)},x) = \int F(x'\beta)\pi(\beta|Y_{(i)})d\beta \doteq \sum_{r=1}^{t} F(x'\beta^r)\tilde{q}_{r(i)}.$$

The symmetric predictive divergence diagnostic for predicting observations at an arbitrary set of locations, say x_j^f, $j=1,\ldots,r$, is

$$D_i^f \equiv \sum_{j=1}^{r} J\Big(\Pr(y=1|Y,x_j^f), \Pr(y=1|Y_{(i)},x_j^f)\Big).$$

A large value of D_i^p or D_i^f indicates that deletion of case i results in different predictive probabilities than if it were retained, possibly resulting in different inferences or decisions.

EXAMPLE 13.3.1. *O-Ring Data.*
Figure 13.9 gives index plots of D_i^p and D_i^f for the O-ring data. The new locations used in defining D_i^f were $x_j^f = 31, 33, 35, \ldots, 51$. The plots for D_{1i}^β and D_{2i}^β were similar to the plot of D_i^p, so they are not included. Case 18, which corresponds to the flight where O-rings failed at the highest launch temperature, consistently stands out. Note that the values of D_i^f are larger for cases with low temperatures. This occurs because the predictions being made are also at low temperatures. The estimative measures and the predictive measure D_i^p are qualitatively similar for these data, although they need not be in general.

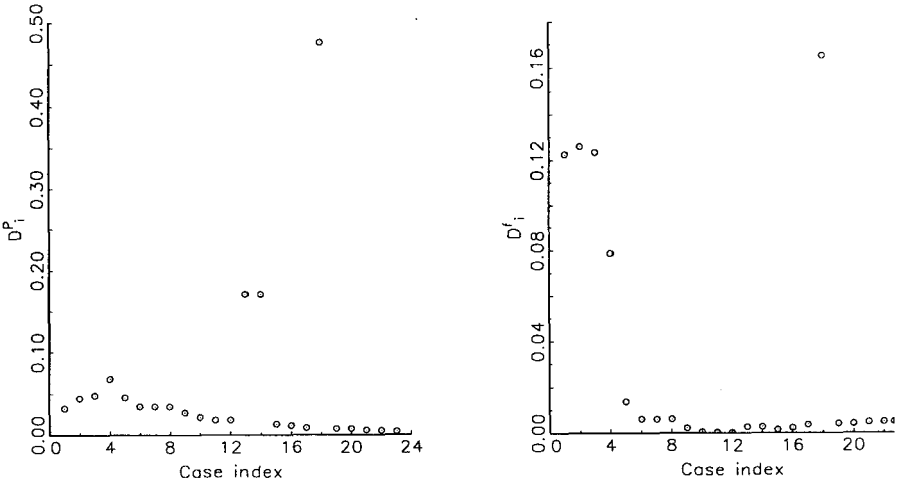


FIGURE 13.9. O-Ring Data: Index Plots of D_i^p and D_i^f

The influence of case 18 was evaluated by repeating the analysis with case 18 deleted. Summary statistics for the Bayesian analysis are provided in Tables 13.2 and 13.4. Omitting case 18, the probability of O-ring failure increased at low temperatures and decreased at high temperatures. The difference in the predictive probabilities for the analyses with and without case 18 are not dramatic. The actual influence of case 18 on the posterior summaries is minor.

EXAMPLE 13.3.2. *Trauma Data.*
Computing the D_i^p's we found cases 52 and 232 to be most influential. Case 52 is a 66-year-old person who had very little wrong with him ($ISS = 9$, $RTS = 7.84$) with a penetration injury who died. Case 232 is a very sick and damaged ($RTS = 2.19$, $ISS = 50$) 50-year-old person with a blunt injury who managed to survive. An ISS score of 50 is characteristic of a person who has very severe injuries to two different parts of the body. The actual statistics are $D_{52}^p = .46$ and $D_{232}^p = .41$, with the next highest value being $D_{173}^p = .25$. Figure 13.10 contains an index plot of the difference in the predictive probabilities of death, $p(y = 1|Y, x_j) - p(y = 1|Y_{(52)}, x_j)$. These probabilities depend on the specified prior. Note that having deleted a case in which a relatively healthy person died, most of the probability differences are very near 0 but slightly positive; e.g., most peoples probabilities of death have decreased. Moreover, all of the changes in probabilities are relatively small. Deletion of case 232 seems to change the regression coefficients even less and would seem to have even less effect on the fitted probabilities.

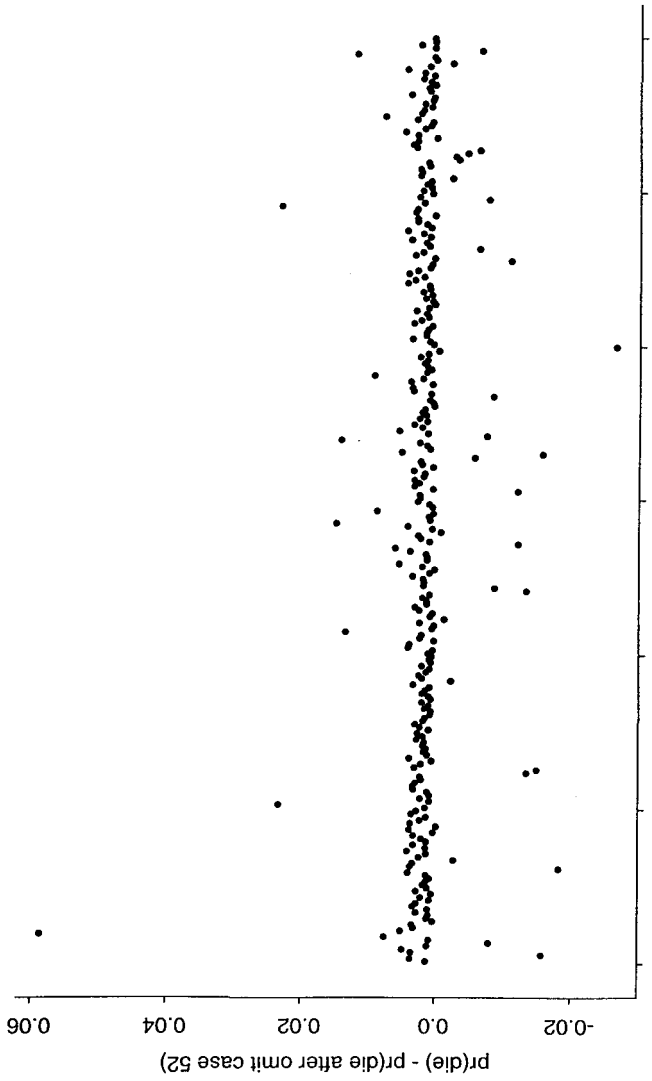


FIGURE 13.10. Trauma Data: Index Plot of $p(y = 1|Y, x_j) - p(y = 1|Y_{(52)}, x_j)$

13.3.2 Model Checking

We consider two methods for model checking. The first is a global model check due to Box (1980). This involves finding the probability that a new vector Y_* has a marginal probability smaller than that of the vector Y that we actually observed, i.e.,

$$\Pr\left[p(Y_*) \leq p(Y)\right],$$

where

$$p(Y) = \int L(\beta|Y)\pi(\beta)d\beta.$$

This is essentially a *P value*, so small values are of significance. For the O-ring data, this value is approximated as .58. The probability is large, so there is no indication of a substantial problem with the model. If the improper diffuse prior $\pi(\beta) = 1$ is used, the required marginal distribution of the data may not exist.

Another model check considers the criterion for one element of the Y vector at a time, i.e.,

$$\Pr\left[p(y_{i*}) \leq p(y_i)\right].$$

This can be viewed as a Bayesian outlier check because we are assessing whether each observation is unusual relative to the model. For the O-ring data, all of these values are 1 except the two identical cases 13 and 14 that give .43 and case 18 that gives .37. This diagnostic gives no indication of substantial problems with the model.

The model checking computations were performed by sampling from the prior distribution. We sample pairs $(\tilde{p}_1, \tilde{p}_2)$ and solve the equations $F^{-1}(\tilde{p}_i) = \beta_0 + \beta_1\tilde{x}_i$, $i = 1, 2$, to obtain samples of β_0 and β_1. Sampling the pairs $(\tilde{p}_1, \tilde{p}_2)$ is easy with our prior because the $\tilde{p}_i$'s have independent beta distributions. Given a sample $\beta^r_\#$, $r = 1, \ldots, v$, from the prior,

$$p(Y) \doteq \frac{1}{v}\sum_{r=1}^{v} L(\beta^r_\#|Y).$$

For an individual component,

$$p(y_i) \doteq \frac{1}{v}\sum_{r=1}^{v} L(\beta^r_\#|y_i).$$

Computing $\Pr[p(Y_*) \leq p(Y)]$ for a new vector Y_* requires an additional round of sampling. For each $\beta^r_\#$, $r = 1, \ldots, v$, generate new independent random variables y_{ir*}, $i = 1, \ldots, n$, that are $\text{Bin}(N_i, F(x_i'\beta^r_\#))$, respectively. The y_{ir*}'s form vectors Y_{r*} for which we can compute $p(Y_{r*})$ as above. $\Pr[p(Y_*) \leq p(Y)]$ is approximated by the proportion of $p(Y_{r*})$'s that are no greater than $p(Y)$. Computation of $\Pr[p(y_{i*}) \leq p(y_i)]$ is similar.

Rubin (1988) advocated Bayesian model checks using predictive rather than marginal distributions. On the O-ring data, Rubin's analogues of the global and local model checks lead to identical conclusions. Chaloner and Brant (1988) check for outliers using the posterior of β. Similar methods also apply to the trauma data.

13.3.3 Link Selection

We now allow the Bayesian paradigm to indicate which of the three link function models is most appropriate for the data: logistic (M_1), probit (M_2), or complementary log-log (M_3). Bayes factors for comparing models M_j and M_k are numbers BF_{jk} such that

$$\frac{P(M_j|Y)}{P(M_k|Y)} = \Big[BF_{jk}\Big]\frac{P(M_j)}{P(M_k)}.$$

The Bayes factor is the multiplier that changes the prior odds for the models into the posterior odds. It is a simple application of Bayes's theorem to show that

$$BF_{jk} = \frac{p(Y|M_j)}{p(Y|M_k)}.$$

$p(Y|M_1)$ was computed previously as $p(Y)$; it is the marginal probability of obtaining Y from the logistic model. Computing $p(Y|M)$ for an alternative model M involves integrating the corresponding likelihood function with respect to the induced prior on β for that model. As in the logistic case, $p(Y|M)$ is estimated using samples generated from the prior on the $\tilde{p}_j$'s.

EXAMPLE 13.3.1 CONTINUED. *O-Ring Data.*
For the O-ring data, the Bayes factors under our prior are $BF_{21} = 1.086$, $BF_{31} = 1.403$, and, thus, $BF_{32} = BF_{31}/BF_{21} = 1.403/1.086 = 1.292$. None of these values is large enough to suggest a serious preference for one of the three models. In particular, if the prior odds for the probit versus logit models are 1, the posterior odds are merely 1.086.

EXAMPLE 13.3.2 CONTINUED. *Trauma Data.*
For the trauma data, the Bayes factors under our prior are $BF_{21} = 1.05$, $BF_{13} = 20.72$, and, thus,

$$BF_{23} = BF_{21}/BF_{31} = BF_{21}BF_{13} = 1.05(20.72) = 21.83\,.$$

There is a suggestion against the complementary log-log model, but there is little to choose from between the logistic and probit models. If the prior odds for the probit versus logit models are 1, the posterior odds are merely 1.05. (These numbers were based on an importance sample of 10,000 observations. Based on only 2000 observations, we got $BF_{21} = 1.97$, BF_{13}

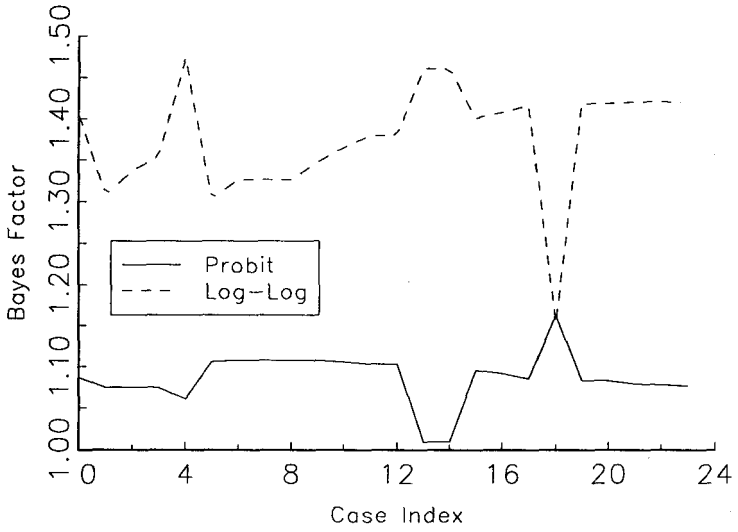


FIGURE 13.11. O-Ring Data: Bayes Factors with Case Deletion

was about the same. Similarly, the values D^p_{52}, D^p_{232}, and D^p_{173} were .525, .519, and .234 based on 2000 samples rather than those reported earlier. The numbers changed, but the message did not!)

We also formed a link selection case deletion diagnostic by computing

$$BF_{jk(i)} = \frac{p(Y_{(i)}|M_j)}{p(Y_{(i)}|M_k)},$$

where

$$p(Y_{(i)}|M) = \int L(\beta|Y_{(i)}, M)\pi(\beta|M)\, d\beta.$$

Figure 13.11 contains a simultaneous plot of $BF_{21(i)}$ versus i and $BF_{31(i)}$ versus i with $i = 1, \ldots, 23$ for the O-ring data. The full data Bayes factors are given by the intercept at "case index 0." Case 18 has the largest effect on the Bayes factors. The deletion of case 18 decreases the posterior odds for the complementary log-log relative to the logistic link, and increases the posterior odds for the probit over the logistic. The actual effect of this case on the Bayes factors is small, and so our decision to use the logistic link is not altered by case deletion.

13.3.4 Sensitivity Analysis

The sensitivity of posterior inferences to the choice of the prior can be evaluated by recalculating posterior summaries based on alternative priors. In situations where the prior changes radically, Monte Carlo samples

from the new posteriors might be needed. When changes in the prior are not dramatic, renormalization of the original Monte Carlo weights might be sufficient. For example, the posterior based on a prior $\pi^*(\beta)$ is approximated by the discrete distribution taking values β^r with probabilities $\tilde{q}_r^*$, where

$$\tilde{q}_r^* = \frac{\pi^*(\beta^r)\tilde{q}_r/\pi(\beta^r)}{\sum_{k=1}^t \pi^*(\beta^k)\tilde{q}_k/\pi(\beta^k)}.$$

EXAMPLE 13.3.1 CONTINUED. *O-Ring Data.*
We used two additional priors to evaluate the sensitivity of our analysis. Each of the priors is a product of independent beta distributions placed at $\tilde{\tau}_1 = 55$ and $\tilde{\tau}_2 = 75$ degrees. Prior II [$\tilde{p}_1 \sim$ Beta(.9,.1) and $\tilde{p}_2 \sim$ Beta(.1,.9)] places a prior (mean) probability of .9 for O-ring failure at 55 degrees and prior probability of .1 for O-ring failure at 75 degrees, while making the beliefs equivalent to one prior observation. Prior III placed Jeffrey's "noninformative" Beta(.5,.5) priors on the $\tilde{p}$'s. The posteriors using the original prior and Prior III were similar, whereas the posterior using Prior II was similar to the posterior obtained from the original prior after omitting case 18. Given the small effect of case 18 on our original analysis, we felt that our posterior analysis was not overly sensitive to these changes in the prior.

EXAMPLE 13.3.2 CONTINUED. *Trauma Data.*
To examine sensitivity to the prior specifications, we considered case deletions of the "prior observations." In Figure 13.12 we present plots of $p(y = 1|Y, x_j, \tilde{Y}) - p(y = 1|Y, x_j, \tilde{Y}_{(i)})$, where each is a predictive probability of success but based on different prior information. Here, the data are the same and the priors involve case deletion. In Figure 13.10, the data involve case deletion but the priors are the same. Note that $\tilde{Y}_{(i)}$ represents partial prior information in the sense of BCJ.

13.4 Posterior Computations and Sample Size Calculation

In recent years, Bayesian analysis has been performed by using numerical integrations (Naylor and Smith, 1982; Smith et al., 1985), by using the analytic Laplace approximation (Leonard,1982; Tierney and Kadane, 1986; Kass et al., 1988), and by using Monte Carlo methods (Zellner and Rossi, 1984; Gelfand and Smith, 1990; Dellaportas and Smith, 1993). See Gelman et al. (1995, Chaps. 9-11) for a nice summary of these methods. We prefer Monte Carlo methods to Laplace approximations in regression problems because when performing many predictions, only a single Monte Carlo sample is necessary to perform all predictions, while the Laplace method

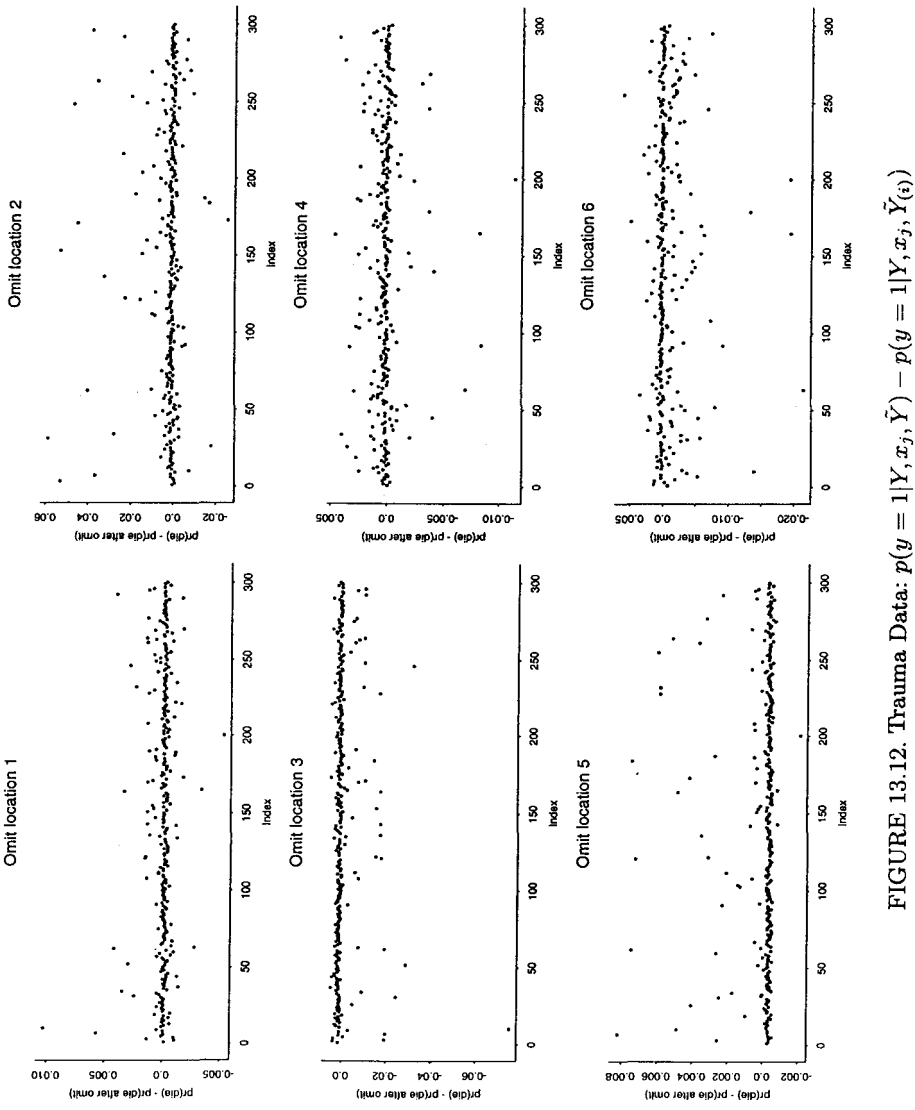


FIGURE 13.12. Trauma Data: $p(y = 1|Y, x_j, \tilde{Y}) - p(y = 1|Y, x_j, \tilde{Y}_{(i)})$

requires a separate analytic approximation for each prediction. We prefer Monte Carlo methods to numerical integration because of their potential to deal with high-dimensional problems. Monte Carlo methods provide a discrete approximation to the posterior distribution. We discuss a variant of importance sampling that is especially simple when used with a DAP.

In importance sampling, one chooses a density function $g(\beta)$ that is similar in shape to the known kernel of the posterior $L(\beta|Y)\pi(\beta)$ with tails that do not decay more rapidly than the tails of the posterior. Then sample $\beta^1, \ldots, \beta^t$ from the distribution with density $g(\beta)$. For $r = 1, \ldots, t$, compute the weights

$$q_r = q(\beta^r) = \frac{L(\beta^r|Y)\pi(\beta^r)}{g(\beta^r)} \tag{1}$$

and

$$\tilde{q}_r = q_r \Big/ \sum_{k=1}^{t} q_k.$$

The discrete approximation to the posterior distribution takes values β^r with probability $\tilde{q}_r$.

Under fairly weak assumptions (cf. Geweke, 1989), the approximation in (13.1.2) has a large sample normal distribution with estimated variance

$$\hat{\sigma}_h^2 = \sum_{r=1}^{t} \{h(\beta^r) - \bar{\theta}_h\}^2 \tilde{q}_r \tag{2}$$

where $\bar{\theta}_h$ is the approximation from (13.1.2). The variance of $\bar{\theta}_h$ depends critically on the tails of $g(\beta)$ through the weight function $q(\beta)$ of (1). Geweke (1989) concluded that to attain high efficiency across a variety of functions, $q(\beta)$ should be reasonably constant with small tails. If the tails of the importance function were allowed to decrease much more rapidly than the tails of the posterior density, the normalized weights $\tilde{q}_r$ could be dominated by individual importance samples in the tail of the approximate posterior. This needlessly inflates the variance of $\bar{\theta}_h$. Similar difficulties can arise with any renormalization of the weights for dealing with case deletions or different priors.

A natural choice for the importance density $g(\beta)$ is a multivariate Student's t density with v degrees of freedom, with location equal to the posterior mode β_M, and dispersion proportional to $\Sigma(\beta_M)$, the asymptotic posterior covariance matrix evaluated at the mode. The approximate $N(\beta_M, \Sigma(\beta_M))$ posterior density is an alternative possibility, but the thin tails of the normal often cause problems; see Zellner and Rossi (1984). Johnson (1987) gives simple algorithms for generating the multivariate normal and $t(v)$ distributions.

Prior to selecting the importance sampling density $g(\beta)$, plot the kernel of $\pi(\beta|Y)$ along the asymptotic principal component directions and choose

the degrees of freedom v so that the tails of $g(\beta)$ are at least as heavy as those of $\pi(\beta|Y)$. Specifically, with $\Sigma(\beta_M) = TT'$, where the columns of T are orthogonal, plot $g(\beta_M + \delta T e_i)$ as a function of δ in each of the unit directions e_i, $i = 1, ..., k$, and similarly for the kernel of the posterior. (e_i is a vector of 0s except for a 1 in the ith place.) In cases of extreme asymmetry along these directions, we recommend sampling from split-t distributions. The split-t distributions allow for different tail heights in each direction, in addition to asymmetry about the mode; see Geweke (1989) for details.

Figure 13.13 gives a plot of the posterior kernel and the normal, $t(6)$, and split-$t(6)$ densities in the direction of the first principal component for the O-ring data. Each function was normalized to have a maximum value of 1. The plot of the posterior reflects the skewness seen in Figure 13.2. The normal density is inferior to the $t(6)$ density as an importance function because the normal underestimates the posterior upper tail in this direction. The weights $q(\beta)$ in this direction at 3, 4, and 4.5 standard deviations above zero are 5, 40, and 150 times greater than the weight at zero for the normal density. For the $t(6)$ density, this ratio is below 3. The corresponding plot along the second principal component was similar, with the exception that the posterior is skewed to the left. The split-$t(6)$ density has heavier tails than the posterior in each direction and reproduces the shape in the center of the posterior. We concluded that the split-$t(6)$ is best among the three importance functions, with the $t(6)$ a close second. Heavier tails on the t distribution could have been obtained by reducing the degrees of freedom, but this was unnecessary. The posterior summaries based on both $t(6)$ and split-$t(6)$ sampling were obtained; they were similar.

For the O-ring data, we decided on the importance sample size by first generating a pilot study of 500 samples. Prediction was a primary goal. We decided that the estimates for the probability of O-ring failure $F(x'\beta)$ and success $1 - F(x'\beta)$ at the 23 observed lift-off temperatures must be accurate. The maximum coefficient of variation across estimates under our prior was 4.4%. To reduce this to a target value 2%, the sample size needed to be increased by a factor of $(2.2)^2 = 4.84$, to approximately 2500. We decided to sample 5000 observations. The estimated maximum coefficient of variation for the parameters of interest based on 5000 samples was 1.4%. Similar methods were used for the trauma data, with a pilot study of 2500 samples and a total sample of 10,000 from a split-$t(6)$.

We noted earlier that β_M and $\Sigma(\beta_M)$ are easily computed using standard software when the prior is a DAP. An interesting special case is the improper prior $\pi(\beta) = 1$, where β_M is the MLE $\hat{\beta}_{ml}$ based on the original data and $\Sigma(\beta_M)^{-1}$ is the observed Fisher information evaluated at $\hat{\beta}_{ml}$. For non-DAP priors, the posterior mode β_M must be computed using specialized software for numerical maximization. Typically, β_M is the solution to $S(\beta) = 0$, where $S(\beta)$ is the vector of partial derivatives of the log of the posterior kernel, i.e., $\log\{L(\beta|Y)\pi(\beta)\}$. The inverse of minus one times

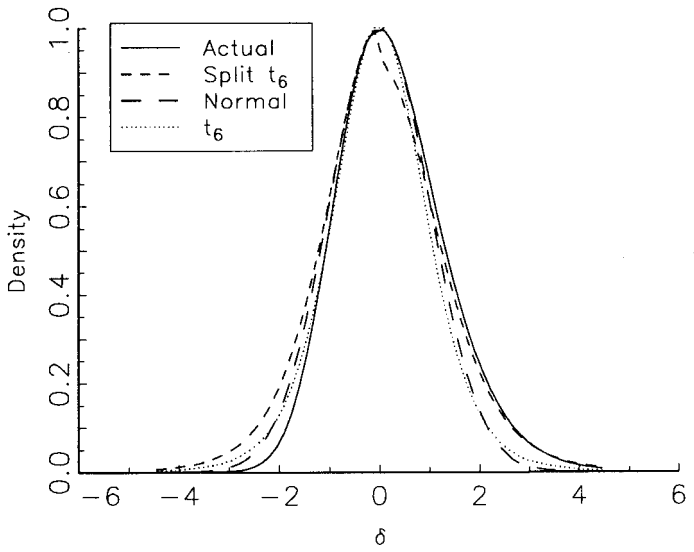


FIGURE 13.13. O-Ring Data: Importance Function Diagnostic Plots

the matrix of second partial derivatives of the log kernel evaluated at β_M serves as $\Sigma(\beta_M)$. Alternatively, the maximum likelihood estimate and the inverse of either the observed or expected Fisher information can be used in place of β_M and $\Sigma(\beta_M)$, cf. Berger (1985, p. 224).

Except for this added computational component, there is no intrinsic problem with using importance sampling with arbitrary priors. Importance sampling can be inefficient when the shape of the posterior density is hard to match. This difficulty might arise when the posterior is highly non-concave or restricted to a subset of the natural parameter space. However, this problem usually does not occur when normal, diffuse, and DAP priors are used with logistic or probit models because the posterior densities are concave. Dellaportas and Smith's (1993) rejection method is also attractive in these cases because the posterior mode is not needed to sample the posterior. Unfortunately, no single method works well, regardless of the prior and link function.

In related work, Smith and Gelfand (1992) presented an introduction to the use of importance sampling and the rejection method. They use the prior distribution as an importance function, while we use the posterior mode β_M and asymptotic posterior dispersion matrix $\Sigma(\beta_M)$ to determine an importance function. Casella and George (1992) explained the Gibbs sampler. When applicable, this provides a random sample of size t from the posterior, so $\tilde{q}_r = 1/t$ for all r. Dellaportas and Smith (1993) combine the Gibbs sampler with the rejection method to obtain samples from the posterior in generalized linear models.

The recent books by Carlin and Louis (1996), Gelman et al. (1995), and Gilks, Richardson and Spiegelhalter (1996) examine a variety of complex modeling problems that can be handled easily using Bayesian methods. Standard references for Bayesian prediction are Aitchison and Dunsmore (1975) and Geisser (1993).

Appendix: Tables

A.1 The Greek Alphabet

TABLE .1. The Greek alphabet

Capital	Small	Name	Capital	Small	Name
A	α	alpha	N	ν	nu
B	β	beta	Ξ	ξ	xi
Γ	γ	gamma	O	o	omicron
Δ	δ, ∂	delta	Π	π	pi
E	ϵ, ε	epsilon	P	ρ	rho
Z	ζ	zeta	Σ	σ	sigma
H	η	eta	T	τ	tau
Θ	θ	theta	Υ	υ	upsilon
I	ι	iota	Φ	ϕ	phi
K	κ	kappa	X	χ	chi
Λ	λ	lambda	Ψ	ψ	psi
M	μ	mu	Ω	ω	omega

A.2 Tables of the χ^2 Distribution

TABLE .2. Percentiles of the χ^2 Distribution

	Percentiles							
df	0.80	0.90	0.95	0.975	0.98	0.99	0.995	0.999
1	1.64	2.71	3.84	5.02	5.41	6.63	7.88	10.83
2	3.22	4.61	5.99	7.38	7.82	9.21	10.60	13.82
3	4.64	6.25	7.81	9.35	9.84	11.35	12.84	16.27
4	5.99	7.78	9.49	11.14	11.67	13.28	14.86	18.47
5	7.29	9.24	11.07	12.83	13.39	15.09	16.75	20.51
6	8.56	10.65	12.59	14.45	15.03	16.81	18.55	22.46
7	9.80	12.02	14.07	16.01	16.62	18.47	20.28	24.32
8	11.03	13.36	15.51	17.53	18.17	20.09	21.95	26.13
9	12.24	14.68	16.92	19.02	19.68	21.67	23.59	27.88
10	13.44	15.99	18.31	20.48	21.16	23.21	25.19	29.59
11	14.63	17.27	19.67	21.92	22.62	24.73	26.76	31.26
12	15.81	18.55	21.03	23.34	24.05	26.22	28.30	32.91
13	16.99	19.81	22.36	24.74	25.47	27.69	29.82	34.53
14	18.15	21.06	23.69	26.12	26.87	29.14	31.32	36.12
15	19.31	22.31	25.00	27.49	28.26	30.58	32.80	37.70
16	20.47	23.54	26.30	28.85	29.63	32.00	34.27	39.25
17	21.61	24.77	27.59	30.19	30.99	33.41	35.72	40.79
18	22.76	25.99	28.87	31.53	32.35	34.81	37.16	42.31
19	23.90	27.20	30.14	32.85	33.69	36.19	38.58	43.82
20	25.04	28.41	31.41	34.17	35.02	37.57	34.00	45.31
21	26.17	29.61	32.67	35.48	36.34	38.93	41.40	46.80
22	27.30	30.81	33.92	36.78	37.66	40.29	42.80	48.27
23	28.43	32.01	35.17	38.08	38.97	41.64	44.18	49.73
24	29.55	33.20	36.41	39.36	40.27	42.98	45.56	51.18
25	30.67	34.38	37.65	40.65	41.57	44.31	46.93	52.62
26	31.79	35.56	38.89	41.92	42.86	45.64	48.29	54.05
27	32.91	36.74	40.11	43.19	44.14	46.96	49.65	55.48
28	34.03	37.92	41.34	44.46	45.42	48.28	50.99	56.89
29	35.14	39.09	42.56	45.72	46.69	49.59	52.34	58.30
30	36.25	40.26	43.77	46.98	47.96	50.89	53.67	59.70
31	37.36	41.42	44.99	48.23	49.23	52.19	55.00	61.10
32	38.47	42.59	46.19	49.48	50.49	53.49	56.33	62.49
33	39.57	43.75	47.40	50.73	51.74	54.77	57.65	63.87
34	40.68	44.90	48.60	51.97	52.99	56.06	58.96	65.25
35	41.78	46.06	49.80	53.20	54.24	57.34	60.27	66.62

TABLE .3. Percentiles of the χ^2 Distribution

	Percentiles							
df	0.80	0.90	0.95	0.975	0.98	0.99	0.995	0.999
36	42.88	47.21	51.00	54.44	55.49	58.62	61.58	67.99
37	43.98	48.36	52.19	55.67	56.73	59.89	62.89	69.35
38	45.08	49.51	53.38	56.90	57.97	61.16	64.18	70.71
39	46.17	50.66	54.57	58.12	59.20	62.43	65.48	72.06
40	47.27	51.81	55.76	59.34	60.44	63.69	66.77	73.41
41	48.36	52.95	56.94	60.56	61.67	64.95	68.05	74.75
42	49.46	54.09	58.12	61.78	62.89	66.21	69.34	76.09
43	50.55	55.23	59.30	62.99	64.11	67.46	70.62	77.42
44	51.64	56.37	60.48	64.20	65.34	68.71	71.89	78.75
45	52.73	57.51	61.66	65.41	66.55	69.96	73.17	80.08
46	53.82	58.64	62.83	66.62	67.77	71.20	74.44	81.40
47	54.91	59.77	64.00	67.82	68.99	72.44	75.70	82.72
48	55.99	60.91	65.17	69.02	70.20	73.68	76.97	84.03
49	57.08	62.04	66.34	70.22	71.41	74.92	78.23	85.35
50	58.16	63.17	67.51	71.42	72.61	76.15	79.49	86.66
51	59.25	64.29	68.67	72.62	73.82	77.39	80.75	87.97
52	60.33	65.42	69.83	73.81	75.02	78.62	82.00	89.27
53	61.41	66.55	70.99	75.00	76.22	79.84	83.25	90.57
54	62.50	67.67	72.15	76.19	77.42	81.07	84.50	91.88
55	63.58	68.80	73.31	77.38	78.62	82.29	85.75	93.17
56	64.66	69.92	74.47	78.57	79.81	83.51	87.00	94.47
57	65.74	71.04	75.62	79.75	81.01	84.73	88.24	95.75
58	66.82	72.16	76.78	80.93	82.200	85.95	89.47	97.03
59	67.90	73.28	77.93	82.12	83.39	87.17	90.72	98.34
60	68.97	74.40	79.08	83.30	84.58	88.38	91.96	99.62
70	79.72	85.53	90.53	95.02	96.39	100.42	104.21	112.31
80	90.41	96.58	101.88	106.63	108.07	112.33	116.32	124.84
90	101.05	107.57	113.15	118.13	119.65	124.11	128.30	137.19
100	111.67	118.50	124.34	129.56	131.14	135.81	140.18	149.48
110	122.25	129.39	135.48	140.92	142.56	147.42	151.95	161.59
120	132.81	140.23	146.57	152.21	153.92	158.95	163.65	173.62
150	164.35	172.58	179.58	185.80	187.67	193.20	198.35	209.22
200	216.61	226.02	234.00	241.06	243.19	249.45	255.28	267.62
250	268.60	279.05	287.88	295.69	298.05	304.95	311.37	324.93
300	320.40	331.79	341.39	349.87	352.42	359.90	366.83	381.34
350	372.05	384.31	394.62	403.72	406.45	414.47	421.89	437.43
400	423.59	436.65	447.63	457.308	460.20	468.71	476.57	492.99

References

Agresti, Alan (1984). *Analysis of Ordinal Categorical Data*. New York: John Wiley and Sons.

Agresti, Alan (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.

Agresti, Alan (1992). A survey of exact inference for contingency tables, with discussion. *Statistical Science*, **7**, 131-153.

Agresti, Alan, Wackerly, Dennis, and Boyett, J.P. (1979). Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika*, **44**, 75-88.

Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge, MA: Cambridge University Press.

Aitkin, Murray (1978). The analysis of unbalanced cross-classifications. *Journal of the Royal Statistical Society, Series B*, **141**, 195-223.

Aitkin, Murray (1979). A simultaneous test procedure for contingency tables. *Applied Statistics*, **28**, 233-242.

Akaike, Hirotugu (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information*, edited by B.N. Petrov and F. Czaki. Budapest: Akademiai Kiado.

Anderson, E.B. (1980). *Discrete Statistical Models with Social Science Application*. Amsterdam: North-Holland.

Anderson, E.B. (1991). *The Statistical Analysis of Categorical Data*, Second Edition. Berlin: Springer-Verlag

Anderson, Erling B. (1992). Diagnostics in categorical data analysis. *Journal of the Royal Statistical Society, Series B*, **54**, 781-791.

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**, 19-35.

Anderson, J.A. and Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, **69**, 123-136.

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition. New York: John Wiley and Sons.

Arnold, Steven F. (1981). *The Theory of Linear Models and Multivariate Analysis*. New York: John Wiley and Sons.

Asmussen, S. and Edwards, D.G. (1983). Collapsibility and response variables in contingency tables. *Biometrika*, **70**, 567-578.

Balmer, D.W. (1988). Recursive enumeration of $r \times c$ tables for exact likelihood evaluation. *Applied Statistics*, **37**, 290-301.

Bartle, Robert G. (1964). *The Elements of Real Analysis*. New York: John Wiley and Sons.

Bartlett, M.S. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Supplement*, **2**, 248-252.

Bedrick, Edward J. (1983). Adjusted chi–squared tests for cross-classified tables of survey data. *Biometrika*, **70**, 591-595.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, **91**, 1450-1460.

Bedrick, E. J., Christensen, R., and Johnson, W. (1997). Bayesian binomial regression. *The American Statistician*, in press.

Bedrick, Edward J. and Hill, Joe R. (1990). Outlier tests for logistic regression: a conditional approach. *Biometrika*, **77**, 815-827.

Belsley, D.A. (1991). *Collinearity Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley and Sons.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics*. New York: John Wiley and Sons.

Benedetti, Jacqueline K. and Brown, Morton B. (1978). Strategies for the selection of log-linear models. *Biometrics*, **34**, 680-686.

Berge, C. (1973). *Graphs and Hypergraphs*. Amsterdam: North-Holland.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition. New York: Springer-Verlag

Berger, James O. and Wolpert, Robert (1984). *The Likelihood Principle*. Hayward, CA: Institute of Mathematical Statistics Monograph Series.

Berkson, Joseph (1978). In dispraise of the exact test. *Journal of Statistical Planning and Inference*, **2**, 27-42.

Bickel, P.J., Hammel, E.A., and O'Conner, J.W. (1975). Sex bias in graduate admissions: data from Berkeley. *Science*, **187**, 398-404.

Binder, D.A., Gratton, M., Hidiroglou, M.A., Kumar, S., and Rao, J.N.K. (1984). Analysis of categorical data from surveys with complex designs: Some Canadian experiences. *Survey Methodology*, **10**, 141-156.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B*, **25**, 220-233.

Bishop, Yvonne M.M., Fienberg, Steven E., and Holland, Paul W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Bortkiewicz, L. von (1898). *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**, 383-404.

Boyett, James M. (1979). Random $R \times C$ tables with given row and column totals. *Applied Statistics*, **29**, 329-332.

Bradley, R.A. and Terry, M.E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrics*, **39**, 324-345.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research*. Lyon: International Agency for Research on Cancer.

Brier, Stephen S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika*, **67**, 591-596.

Brown, Byron Wm., Jr. (1980). Prediction analyses for binary data. In *Biostatistics Casebook*, edited by R.G. Miller, Jr. et al. New York: John Wiley and Sons.

Brown, Morton B. (1976). Screening effects in multidimensional contingency tables. *Applied Statistics*, **25**, 37-46.

Brunswick, A.F. (1971). Adolescent health, sex, and fertility. *American Journal of Public Health*, **61**, 711-720.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, New York: Chapman and Hall.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651-660.

Cheng, Bing and Titterington, D.M. (1994). Neural networks: A review from a statistical perspective, with discussion. *Statistical Science*, **9**, 2-30.

Christensen, Ronald (1987). *Plane Answers to Complex Questions: The Theory of Linear Models*, First Edition. New York: Springer-Verlag.

Christensen, Ronald (1990). *Linear Models for Multivariate, Time Series and Spatial Data*. New York: Springer-Verlag.

Christensen, Ronald (1996a). *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. London: Chapman and Hall.

Christensen, Ronald (1996b). *Plane Answers to Complex Questions: The Theory of Linear Models*, Second Edition. New York: Springer-Verlag.

Christensen, Ronald and Utts, Jessica (1992). Testing for nonadditivity in log-linear and logit models. *Journal of Statistical Planning and Inference*, **33**, 333-343.

Chuang, Christy (1983). Multiplicative-interaction logit models for $I \times J \times 2$ three-way tables. *Communications in Statistics, Theory and Methods*, **12**, 2871-2885.

Clayton, Murray K., Geisser, Seymour, and Jennings, Dennis E. (1986). A comparison of several model selection procedures. In *Bayesian Inference and Decision Techniques*, edited by P. Goel and A. Zellner. Amsterdam: North-Holland

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19** 15-18.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D.R. and Snell, E.J.(1989). *Analysis of Binary Data*, Second Edition. London: Chapman and Hall.

Cramér, Harald (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440-464.

Dalal, S.R., Fowlkes, E.B., and Hoadley, B. (1989) Risk analysis of the space shuttle: Pre-*Challenger* prediction of failure. *Journal of the American Statistical Association*, **84**, 945-957.

Darroch, J.N., Lauritzen, S.L., and Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, **8**, 522-539.

David, Herbert A. (1988). *The Method of Paired Comparisons*, Second Edition. New York: Oxford University Press.

Davison, A.C. (1988). Approximate conditional inference in generalized linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 445-461.

del Pino, Guido (1989). The unifying role of iterative generalized least squares in statistical algorithms. *Statistical Science*, **4**, 394-403.

Dellaportas, P. and Smith, A.F.M. (1993). Bayesian inference for generalized linear models and proportional hazards via Gibbs sampling. *Applied Statistics*, **42**, 443-459.

Deming, W. Edwards and Stephan, Frederick F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, 427-444.

Dixon, Wilfrid J. and Massey, Frank J., Jr. (1983). *Introduction to Statistical Analysis*. New York: McGraw-Hill.

Draper, Norman and Smith, Harry (1981). *Applied Regression Analysis*, Second Edition. New York: John Wiley and Sons.

Duncan, Otis Dudley (1975). Partitioning polytomous variables in multiway contingency analysis. *Social Science Research*, **4**, 167-182.

Duncan, O.D. and McRae, J.A., Jr. (1979). Multiway contingency analysis with a scaled response or factor. In *Sociological Methodology*. San Francisco: Jossey-Bass.

Duncan, O.D., Schuman, H., and Duncan, B. (1973). *Social Change in a Metropolitan Community*. New York: Russell Sage Foundation.

Edwards, David (1995). *Introduction to Graphical Modeling*. Berlin: Springer-Verlag.

Edwards, David and Havranek, Tomas (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339-351.

Edwards, David and Kreiner, Sven (1983). The analysis of contingency tables by graphical models. *Biometrika*, **70**, 553-565.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association,* **70**, 892-898.

Everitt, B.S. (1977). *The Analysis of Contingency Tables*. London: Chapman and Hall.

Farewell, V.T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**, 27-32.

Fay, Robert E. (1985). A jackknifed chi–squared test for complex samples. *Journal of the American Statistical Association*, **80**, 148-157.

Feigl, P. and Zelen, M. (1965). Estimation of exponential probabilities with concomitant information. *Biometrics*, **21**, 826-838.

Fienberg, S.E. (1968). *The Estimation of Cell Probabilities in Two-Way Contingency Tables*. Ph. D. Thesis, Department of Statistics, Harvard University.

Fienberg, Stephen E. (1979). The use of chi–squared statistics for categorical data problems. *Journal of the Royal Statistical Society, Series B*, **41**, 54-64.

Fienberg, Stephen E. (1980). *The Analysis of Cross-Classified Categorical Data*, Second Edition. Cambridge, MA: MIT Press.

Fienberg, Stephen E. and Gong, Gail D. (1984). Comment on a paper by Landwehr et al. *Journal of the American Statistical Association*, **79**, 72-77.

Fienberg, Stephen E. and Larntz, K. (1976). Loglinear representation for paired and multiple comparisons models. *Biometrika*, **63**, 245-254.

Finney, D.J. (1941). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, **34**, 320-334.

Finney, D.J. (1971). *Probit Analysis*, Third Edition. Cambridge: Cambridge University Press.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*. New York: Hafner Press.

Fisher, Ronald A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics*. New York: W. W. Norton.

Freeman, M.F. and Tukey, J.W. (1950). Transformations related to the angular and the square root. *Annals of Statistics*, **21**, 607-611.

Geisser, Seymour (1977). The inferential use of predictive distributions. In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott. Toronto: Holt, Rinehart and Winston.

Geisser, S. (1982). Aspects of the predictive and estimative approaches to the determination of probabilities. *Biometrics Supplement: Current Topics in Biostatistics and Epidemiology*, **38**, 75-93.

Geisser, S. (1993). *Predictive Inference: An Introduction*. New York: Chapman and Hall.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*, New York: Chapman and Hall.

Geweke, J. (1989). Bayesian Inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317-1339.

Gilby, Walter H. (1911). On the significance of the teacher's appreciation of general intelligence. *Biometrika*, **VII**, 79-93.

Gilks, W. R., Richardson, S., and Spiegelhalter, D., eds. (1996). *Practical Markov Chain Monte Carlo,* New York: Chapman and Hall.

Glymour, Clark, Scheines, Richard, Spirtes, Peter, and Kelly, Kevin (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. New York: Academic.

Gong, Gail (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, **81**, 108-113.

Good, I.J. (1975). The number of hypotheses of independence for a random vector or a multidimensional contingency table, and the Bell numbers. *Iranian Journal of Science and Technology*, **4**, 77-83.

Goodman, L.A. (1970). The multivariate analysis of qualitative data: Interaction among multiple classifications. *Journal of the American Statistical Association*, **65**, 226-256.

Goodman, L.A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional tables. *Journal of the American Statistical Association*, **66**, 339-344.

Goodman, L.A. (1973). The analysis of contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, **60**, 179-192.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537-552.

Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, **76**, 320-334.

Grieve, A.P. (1988). A Bayesian approach to the analysis of $LD50$ experiments. In *Bayesian Statistics 3*, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith. Oxford University Press, Oxford, 617-630.

Grizzle, James E., Starmer, C. Frank, and Koch, Gary G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Grizzle, James E. and Williams, O. Dale (1972). Log linear models and tests of independence for contingency tables. *Biometrics*, **28**, 137-156.

Gross, A.J. (1984). A note on "chi–squared tests with survey data." *Journal of the Royal Statistical Society, Series B*, **46**, 270-272.

Haberman, Shelby J. (1974a). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Haberman, Shelby J. (1974b). Loglinear models for frequency tables with ordered classifications. *Biometrics*, **30**, 589-600.

Haberman, Shelby J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, **5**, 1148-1169.

Haberman, Shelby J. (1978). *The Analysis of Qualitative Data*, Volume 1. New York: Academic Press.

Haberman, Shelby J. (1979). *The Analysis of Qualitative Data*, Volume 2. New York: Academic Press.

Hand, D.J. (1981). *Discrimination and Classification*. New York: John Wiley and Sons.

Havranek, Tomas (1984). A procedure for model search in multidimensional contingency tables. *Biometrics*, **40**, 95-100.

Hirji, K.F., Mehta, C.R., and Patel, N.R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, **82**, 1110-1117.

Holland, Paul W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945-960

Holt, D., Scott, A.J., and Ewings, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Series A*, **143**, 302-320.

Hosmer, David W. and Lemeshow, Stanley (1989). *Applied Logistic Regression*. New York: John Wiley and Sons.

Imrey, P.B., Johnson, W.D., and Koch, G.G. (1976). An incomplete contingency table approach to paired comparison experiments. *Journal of the American Statistical Association*, **71**, 614-623.

Irwin, J.O. (1949). A note on the subdivision of χ^2 into components. *Biometrika*, **36**, 130-134.

Jennings, Dennis E. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association*, **81**, 987-990.

Johnson, D.E. and Graybill, F.A. (1972). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association*, **67**, 862-868.

Johnson, M.E. (1987). *Multivariate Statistical Simulation*. New York: John Wiley and Sons.

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika*, **72**, 59-65.

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. In *Statistics and Probability: Essays in Honor of C.R. Rao*, edited by G. Kallianpur, P.B. Krishnaiah, and J.K. Ghosh. Amsterdam: North-Holland.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression. *Journal of the American Statistical Association*, **78**, 137-144.

Johnson, W. and Geisser, S. (1985). Estimative influence measures for the multivariate general linear model. *Journal of the Statistical Planning and Inference*, **11**, 33-56

Kass, R., Tierney, L., and Kadane, J. (1988). Asymptotics in Bayesian computation. In *Bayesian Statistics 3*, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A.F.M. Smith. Oxford: Oxford University Press.

Kempthorne, Oscar (1979). On dispraise of the exact test: reactions. *Journal of Statistical Planning and Inference*, **3**, 199-213.

Kihlberg, J.K., Narragon, E.A., and Campbell, B.J. (1964). Automobile crash injury in relation to car size. Cornell Aero. Lab. Report No. VJ-1823-Rll.

Kiiveri, H. and Speed, T.P. (1982). Structural analysis of multivariate data: A review. In *Sociological Methodology*, edited by S. Leinhart. San Francisco: Jossey–Bass.

Kiiveri, H., Speed, T.P., and Carlin, J.B. (1984). Recursive causal models. *Journal of the Australian Mathematics Society*, **36**, 30-52.

Kish (1965). *Survey Sampling*. New York: John Wiley and Sons.

Koch, G.G., Freeman, D.H., and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, **43**, 59-78.

Koch, Gary G., Imrey, Peter B., Freeman, Daniel H., Jr., and Tolley, H. Dennis (1976). The asymptotic covariance structure of estimated parameters from contingency table log-linear models. *Proceedings of the 9th International Biometric Conference*, vol. 1, 317-336.

Koehler, Kenneth J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, **81**, 483-493.

Koehler, Kenneth J. and Larntz, Kinley (1980). An assessment of several asymptotic approximations for sparse multinomials. *Journal of the American Statistical Association*, **75**, 336-344.

Kreiner, Sven (1987). Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scandinavian Journal of Statistics*, **14**, 97-112.

Lachenbruch, P.A. (1975). *Discriminate Analysis*. New York: Hafner Press.

Lachenbruch, P.A., Sneeringeer, C., and Revo, L.T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics*, **1**, 39-57.

Lancaster, H.O. (1949). The derivation and partition of χ^2 in certain discrete distributions. *Biometrika*, **36**, 117-129.

Landwehr, James M., Pregibon, Daryl, and Shoemaker, Anne C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, **79**, 61-71.

Larntz, Kinley (1978). Small sample comparisons of exact levels for chi-square goodness-of-fit statistics. *Journal of the American Statistical Association*, **73**, 253-263.

Lauritzen, Steffen L. (1996). *Graphical Models*. Oxford: Oxford University Press.

Lavine, M. (1991). Problems in extrapolation illustrated with pace shuttle o-ring data (with discussion). *Journal of the American Statistical Association*, **86**, 919-921.

Lazerwitz, Bernard (1961). A comparison of major United States religious groups. *Journal of the American Statistical Association*, **56**, 568-579.

Lee, Elisa T. (1980). *Statistical Methods for Survival Data Analysis*. Belmont, CA: Lifetime Learning Publications.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, Second Edition. New York: John Wiley and Sons.

Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika*, **59**, 581-589.

Leonard, T. (1982). Comment on a paper by Lejeune and Faulkenberry. *Journal of the American Statistical Association*, **77**, 657-658.

Lindgren, B.W. (1993). *Statistical Theory*, Fourth Edition. New York: Macmillan.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association*, **81**, 104-107.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Second Edition. London: Chapman and Hall.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley and Sons.

McNemar, Q. (1947). Note on the sampling error of differences between correlated proportions or percentages. *Psychometrika*, **12**, 153-157.

Mandel, J. (1961). Nonadditivity in two-way analysis of variance. *Journal of the American Statistical Association*, **56**, 878-888.

Mandel, J. (1971). A new analysis of variance model for nonadditive data. *Technometrics*, **13**, 1-18.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Martz, H. F. and Zimmer, W. J. (1992). The risk of catastrophic failure of the solid rocket boosters on the space shuttle. *The American Statistician*, **46**, 42-47.

Mehta, C.R. (1994). The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research*, **3**, 135-156.

Mehta, C.R. and Patel, N.R. (1980). A network algorithm for the exact treatment of the $2 \times k$ contingency table. *Communications in Statistics–Part B*, **9**, 649-664.

Mehta, C.R. and Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, **78**, 427-434.

Mehta, C.R., Patel, N.R., and Gray, R. (1983). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *Journal of the American Statistical Association*, **80**, 969-973.

Mehta, C.R., Patel, N.R., and Tsiatis, A.A. (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics*, **40**, 819-825.

Meyer, Michael M. (1982). Transforming contingency tables. *Annals of Statistics*, **10**, 1172-1181.

Milliken, G.A. and Graybill, F.A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, **65**, 797-807.

Mosteller, Frederick, and Tukey, John W. (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214-225.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

Nordberg, Lennart (1981). Stepwise selection of explanatory variables in the binary logit model. *Scandinavian Journal of Statistics*, **8**, 17-26.

Nordberg, Lennart (1982). On variable selection in generalized linear and related regression models. *Communications in Statistics, A*, **11**, 2427-2449.

O'Neill, Terence J. (1994). The bias of estimating equations with application to the error rate in logistic discrimination. *Journal of the American Statistical Association*, **89**, 1492-1498.

Pagano, M. and Taylor-Halvorson, K. (1981). An algorithm for finding the exact significance levels of $r \times c$ contingency tables. *Journal of the American Statistical Association*, **76**, 931-934.

Patefield, W.M. (1981). An efficient method of generating random $R \times C$ tables with given row and column totals. *Applied Statistics*, **30**, 91-97.

Pettit, L.I. and Smith, A.F.M. (1985). Outliers and influential observations in linear models. *Bayesian Statistics*, **2**, Edited by J.M. Bernardo. Amsterdam: North-Holland, 473-494.

Plackett, R.L. (1981). *The Analysis of Categorical Data*, Second Edition. London: Griffin.

Pregibon, Daryl (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705-724.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-411.

Press, S. James (1984). *Applied Multivariate Analysis*, Second Edition. New York: Holt, Rinehart and Winston.

Press, S. James and Wilson, Sandra (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, **73**, 699-705.

Quine, S. (1975). *Achievement Orientation of Aboriginal and White Australian Adolescents*. Ph. D. Thesis, Australian National University.

Radelet, M. (1981). Racial characteristics and the imposition of the death penalty. *American Sociological Review*, **46**, 918-927.

Rao, C. Radhakrishna (1965). *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons.

Rao, C. Radhakrishna (1973). *Linear Statistical Inference and Its Applications*, Second Edition. New York: John Wiley and Sons.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex survey samples: Chi-squared tests for goodness-of-fit and independence in two-way tables. *Journal of the American Statistical Association*, **76**, 221-230.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, **12**, 46-60.

Rao, J.N.K. and Scott, A.J. (1987). On simple adjustments to chi-squared tests with sample survey data. *Annals of Statistics*, **15**, 385-397.

Read, Timothy R.C. and Cressie, Noel A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Reiss, I.L., Banwart, A., and Foreman, H. (1975). Premarital contraceptive usage: A study and some theoretical explorations. *Journal of Marriage and the Family*, **37**, 619-630.

Rosenberg, M. (1962). Test factor standardization as method interpretation. *Social Forces*, **41**(October), 53-61.

Rubin, D.B. (1987). The SIR algorithm – A discussion of Tanner and Wong's: The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 543-546.

Rubin, D.B. (1988). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151-1172.

Santner, Thomas J. and Duffy, Diane E. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag.

Schwarz, Gideon (1978). Estimating the dimension of a model. *Annals of Statistics*, **16**, 461-464.

Seber, G.A.F. (1984). *Multivariate Observations*. New York: John Wiley and Sons.

Simonoff, Jeffrey S. (1983). A penalty function approach to smoothing large sparse multinomials contingency tables. *Annals of Statistics*, **11**, 208-218.

Simonoff, Jeffrey S. (1985). An improved goodness-of-fit statistic for sparse multinomials. *Journal of the American Statistical Association*, **80**, 671-677.

Simonoff, Jeffrey S. (1986). Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *Journal of the American Statistical Association*, **81**, 1005-1012.

Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.

Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, **46**, 84-88.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics – Theory and Methods*, **14**, 1079-1102.

Thomas, D. Roland and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, **82**, 630-636.

Thomas, William and Cook, R. Dennis (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, **76**, 741-749.

Thomas, William and Cook, R. Dennis (1990). Assessing influence on predictions for generalized linear models. *Technometrics*, **32**, 59-65.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82-86.

Tsiatis, Anasatasios A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, **67**, 250-251.

Tsutakawa, R.K. (1975). Bayesian inference for bioassay. Technical Report No. 52. Department of Statistics, Univ. of Missouri – Columbia.

Tsutakawa, R.K. and Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, **51**, 251-267.

Tukey, J.W. (1949). One degree of freedom for nonadditivity. *Biometrics*, **5**, 232-242.

Waite, H. (1911). The teacher's estimation of the general intelligence of school children. *Biometrika*, **VII**, 79-93.

Weisberg, Sanford (1975). Inference for tables of counts: Contingency tables. Unpublished Stat 5021 Handout. School of Statistics, University of Minnesota, Minneapolis.

Weisberg, Sanford (1985). *Applied Linear Regression*, Second Edition. New York: John Wiley and Sons.

Wermuth, Nanny (1976). Model search among multiplicative models. *Biometrics*, **32**, 253-263.

Wermuth, Nanny and Lauritzen, Steffen L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537-552.

Whittaker, Joe (1990). *Graphical Models in Applied Multivariate Statistics*. New York: John Wiley and Sons.

Wing, J.K. (1962). Institutionalism in mental hospitals. *British Journal of Social and Clinical Psychology*, **1**, 38-51.

Woodward, G. Lange, S.W., Nelson, K.W., and Calvert, H.O. (1941). The acute oral toxicity of acetic, chloracetic, dichloracetic and trichloracetic acids. *Journal of Industrial Hygiene and Toxicology*, **23**, 78-81.

Yule, G.U. (1900). On the association of attributes in statistics: With illustration from the material of the childhood society, etc. *Philosophical Transactions of the Royal Society, Series A*, **194**, 257-319.

Zellner, A. and Rossi, P.E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, **25**, 365-393.

Zelterman, Daniel (1987). Goodness-of-fit statistics for large sparse multinomial distributions. *Journal of the American Statistical Association*, **82**, 624-629.

Author Index

Subject Index

Springer Texts in Statistics *(continued from page ii)*

Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II: Statistical Inference
Noether: Introduction to Statistics: The Nonparametric Way
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: A Decision-Theoretic Motivation
Robert and Casella: Monte Carlo Statistical Methods
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics
Shumway and Stoffer: Time Series Analysis and its Applications.
Terrell: Mathematical Statistics: A Unified Introduction
Whittle: Probability via Expectation, Third Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods