36-720: Discrete Multivariate Analysis
HW04, Due Wednesday, October 17, 2007

Announcements:

- The last part of the course is on topics that are not really covered in Christensen's text: generalized linear mixed models, Rasch model, latent class models. Please see the lecture notes, for information and references.
- Please remember that, although you are free to talk with one another about HW's, the work you turn in should be your own.
- Best to use a word processor that can handle mathematics (like LaTeX) and can include graphics from R and other programs. Next best is neat handwriting with neatly cut-and-pasted tables and graphs.

Problems:

1. Consider the log-linear model for [12][13][23] for a three-way table,

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \ ,$$

under Poisson sampling, identified by setting the first level of every u-term equal to zero.

(a) Find the sufficient statistics and show that they are among the margins in the tables $\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$

(b) Show that all 12 entries in the three tables $\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$ are functions of the sufficient statistics you found.

2. Consider again the log-likelihood for a log-linear model $\mathcal{M}_1$ for the table $\{n_c : c = 1, 2, \ldots, C\}$ under Poisson sampling,

$$\ell(n|\mathcal{M}_1) = \sum_c (n_c \log m_c^{(1)} - m_c^{(1)}) - \sum_c \log n_c!$$

Assume $\mathcal{M}_1$ and all models in this problem contain an intercept, so that n_+ is part of the sufficient statistics. Also, $m_c^{(k)}$ are the (fitted) expected cell counts for model $\mathcal{M}_k$.

(a) Let $\mathcal{M}_\infty$ be the saturated model with (fitted) expected cell counts $m_c^{(\infty)} = n_c$. Show that the likelihood ratio test for testing $\mathcal{M}_1$ vs $\mathcal{M}_\infty$ is the deviance statistic

$$G^2(\mathcal{M}_1) = -2 \sum_c n_c \log n_c / m_c^1$$

(b) Suppose $\mathcal{M}_1$ is a submodel of $\mathcal{M}_2$. Show that the likelihood ratio test statistic for testing $\mathcal{M}_1$ vs $\mathcal{M}_2$ can be written as

$$G^2(\mathcal{M}_1) - G^2(\mathcal{M}_2)$$

and argue that this quantity *must* be non-negative.

(c) Use the previous result to show that model selection using G^2 statistics satisfies *coherence*.

(d) Give a numerical example, or make a theoretical argument, that model selection using AIC does not satisfy coherence, in general. How about BIC?

3. It has been well-known for years that high-school aged females tend to take fewer mathematics courses than males. The 1979 Women and Mathematics (WAM) Secondary Scholarship Program was designed both to encourage more interest in mathematics by females and to show positive role models by presenting lectures, all given by women in the mathematical sciences. 1190 students (males and females) at eight high schools (four urban and four suburban) responded to a questionnaire including the item in Table 1 on p. 4 below. These six variables were cross-classified in a single contingency table, shown in Table 2 on p. 4. (the raw table has been saved in `wam.txt` for you, but it will require some massaging to get in a useful form for R).

   (a) Find the best hierarchical log-linear model for this data.

   (b) Find the best decomposable log-linear model for this data.

   (c) Compare the two models in parts (a) and (b). Say which model is better (defend your choice with evidence and/or careful reasoning!), and interpret that model in terms that would be of interest to the people conducting the study.

4. Continuing with Exercise 3…

   (a) We can consider "Response" (Agree/Disagree) to be a binary outcome variable. Considered in this way, is this better thought of as a prospective or retrospective study?

   (b) Reanalyze the data with a logistic regression model. Select the best variables for the logistic regression model (provide evidence and/or careful reasoning).

   (c) Interpret your final model

   - As a log-linear model;
   - In terms that would be of interest to the people conducting the study.

5. In 1999 a small study was conducted of middle-school students' ability to solve "proportional reasoning" problems. For a part of this study, $N = 67$ students were asked to solve $J = 7$ problems in proportional reasoning. The problems were of three types:

Missing Value Problems: The student is given a "story problem" whose solution is tantamount to solving an equation like

$$\frac{a}{b} = \frac{c}{d} \tag{1}$$

in which one of *a*, *b*, *c* or *d* is missing. The student has to solve for the missing value, given the other three. *Problems 1, 2, 3, and 6 were of this type.*

Relative/Absolute Problems: The student is given a "story problem" and the student has to decide whether the solution involves a relation among ratios as in (1) above (called a "relative" missing value problem) or a relation among differences like (2) below (called an “absolute” missing value problem)

$$a - b = c - d \tag{2}$$

Problems 4 and 5 were of this type.

Similarity Problems: These are like Missing-Value problems, but the problem is given as a figure in which two geometric objects (typically triangles or rectangles) are asserted to be similar (same shape and relative proportions). *Problem 7 is of this type.*

The data set for this problem consists of a matrix `y[i,j]` in the file `pr-responses.txt`, for which

$$\texttt{y[i,j]} = \begin{cases} 1, & \text{if student } i \text{ answered question } j \text{ correctly;} \\ 0, & \text{otherwise} \end{cases}$$

Letting $p_{ij} = P[y_{ij} = 1]$, two possible models to consider for this data are

- The *Rasch* model,

$$\log p_{ij}/(1 - p_{ij}) = u_i - \beta_j$$

- The *linear logistic test model (LLTM)*, which simply places linear regression structure on the β_j's corresponding to the three problem types above:

$$\log p_{ij}/(1 - p_{ij}) = u_i - X_j \alpha$$

where

$$X_j = \begin{cases} (1,0,0) & \text{if } j \text{ is a missing value problem} \\ (0,1,0) & \text{if } j \text{ is a relative/absolute problem} \\ (0,0,1) & \text{if } j \text{ is a similarity problem} \end{cases}$$

and $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T$ are the appropriate fixed-effects regression coefficients.

(a) Using `lmer()` and any other appropriate tools at your disposal, fit the Rasch and LLTM models as GLMM's, explore and compare the fits of each, and decide which (if either) model provides adequate fit to the data. Provide evidence to defend your choices.

(b) Repeat part (a), but now treating the marginal versions of the models as log-linear models for the (rather sparse!) 2^7 table cross-classifying responses to the seven proportional reasoning questions. Explore and compare the fits of the Rasch and LLTM models as log-linear models and decide which (if either) provides adequate fit to the data; provide evidence to defend your choices.

In either (a) or (b) you may (or may not) get stuck and be unable to continue. If you do get stuck, explain *clearly* where and why you are stuck, and provide evidence to defend the assertion that the problem cannot be solved as stated.

Name	Meaning	Levels
Need	“I’ll need mathematics in my future work”	Agree Disagree
WAM	Attended WAM lectures?	Yes No
Sex	–	Female Male
School	Location of High School	Suburban Urban
Prefs	Course preferenes	Math/Sci Lib. Arts
Plans	Future plans	College Job

Table 1: The variables in the WAM survey [questions 3 and 4].

	Suburban School				Urban School			
	Female		Male		Female		Male	
	WAM	No WAM	WAM	No WAM	WAM	No WAM	WAM	No WAM
Plans: College; Prefs: Math/Sci								
Agree	37	27	51	48	51	55	109	86
Disagree	16	11	10	19	24	28	21	25
Plans: College; Prefs: Lib. Arts								
Agree	16	15	7	6	32	34	30	31
Disagree	12	24	13	7	55	39	26	19
Plans: Job; Prefs: Math/Sci								
Agree	10	8	12	15	2	1	9	5
Disagree	9	4	8	9	8	9	4	5
Plans: Job; Prefs: Lib. Arts								
Agree	7	10	7	3	5	2	1	3
Disagree	8	4	6	4	10	9	3	6

Table 2: The data from the WAM survey [questions 3 and 4].