Lecture 9: Loglinear models

Claudia Czado

TU München

Overview

- Introduction to loglinear models
- Two dimensional loglinear models
- Three dimensional loglinear models

Introduction

Reference: Agresti (1990), chapters 5, 6

Loglinear models are used to determine relationships like

- independance
- conditional independance
- dependance

between variables. Particularly, no variable is characterized as response.

Example: Classification of medical test results

Contingency table, source: Agresti (1990), p.158

Age	Smoking status	Test result	
		normal	not normal
< 40	non-smoker	577	34
	smoker	682	57
$40 - 59$	non-smoker	164	4
	smoker	245	74

Variables (here: Age, Smoking status, Test result) are called classification variables.
Each of them has a certain number of categories.

Relationship between variables?

Model frame

d = # of classification variables
L_j = # of categories of classification variable j $(j = 1, \ldots, d)$
i = $(i_1, \ldots, i_d)$, with $1 \leq i_j \leq L_j$ multiple index for cell identification
I = $\{i = (i_1, \ldots, i_d), 1 \leq i_j \leq L_j, j = 1, \ldots d\}$

In our example:

$$d = 3, \quad \underbrace{L_1}_{\text{Age}} = \underbrace{L_2}_{\text{Smoking status}} = \underbrace{L_3}_{\text{Test result}} = 2$$

$\Rightarrow i = (1, 2, 1)$ means

$$\text{Age} < 40, \text{ Smoker, Normal test result}$$

Distribution of observations

Y_i $= \#$ of observations in cell i

$\{Y_i, i \in I\}$ $\sim multinomial\ \mathcal{M}_t(m, \boldsymbol{p})$

$t := \prod_{j=1}^{d} L_j = \#$ of cells

$m := \sum_{i \in I} Y_i =$ total $\#$ of observations

$p_i := P(\text{observation falls in cell } i),\ i \in I,\ \boldsymbol{p} := (p_i, i \in I)$

$$\Rightarrow \quad P(Y_i = y_i, i \in I) = \frac{m!}{\prod_{i \in I} y_i!} \prod_{i \in I} p_i^{y_i},$$

$$E(Y_i) = mp_i,\ Var(Y_i) = mp_i(1 - p_i),\ Cov(Y_i, Y_j) = -mp_ip_j$$

General loglinear models

$$\{Y_i, i \in I\} \sim \mathcal{M}_t(m, \boldsymbol{p})$$

$$l_i := log(E(Y_i)) = log(mp_i) = l_{(i_1,\ldots,i_d)}$$

$$= \lambda + \sum_{j=1}^{d} \lambda_{i_j}^{j} + \sum_{1\le j<k\le d}\sum \lambda_{i_j i_k}^{jk} + \sum_{1\le j<k<l\le d}\sum\sum \lambda_{i_j i_k i_l}^{jkl} + \ldots + \lambda_{i_1\cdots i_d}^{12\cdots d}$$

ANOVA-like decomposition of expected value

Since $\sum_{i\in I} p_i = 1 \quad \Rightarrow \sum_{i\in I} mp_i = m$

$$\Rightarrow l_i = log(mp_i) \Rightarrow \prod_{i\in I} l_i = \log\left(\sum_{i\in I} mp_i\right) = \log(m)$$

Two dimensional contingency tables

$$d = 2, \quad L_1 = R, \quad L_2 = C, \quad R \times C \text{ table}$$

	1	2	$\cdots$	C
1	p_{11}	p_{12}	$\cdots$	p_{1C}
2	p_{21}	p_{22}	$\cdots$	p_{2C}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
R	p_{R1}	p_{R2}	$\cdots$	p_{RC}

$\boldsymbol{p}_{RC}$

$$\Rightarrow l_{ij} = \log(mp_{ij}) = \lambda + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12}$$

Identifiability conditions

$l_{ij} = \log(mp_{ij}) = \lambda + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12}$

$\lambda, \lambda_i^1, \lambda_j^2, \lambda_{ij}^{12}, \quad (i = 1, \ldots, R, \quad j = 1, \ldots, C)$ have to be estimated

$\Rightarrow 1 + R + C + RC$ parameters, but only RC different l_{ij}'s

Need identifiability conditions, e. g.

$$\sum_{i=1}^{R} \lambda_i^1 = 0, \quad \sum_{j=1}^{C} \lambda_j^2 = 0, \quad \sum_{i=1}^{R} \lambda_{ij}^{12} = 0 \quad , (i = 1, \ldots C)$$

$$\sum_{j=1}^{L} \lambda_{ij}^{12} = 0 \quad , (j = 1, \ldots L)$$

$R + C + 2$ identifiability conditions, RC different l_{ij}'s with $\prod_{i=1}^{R} \prod_{j=1}^{R} l_{ij} = \log(m)$

$\Rightarrow$ parameters can be identified

Parameter specification

The previous identifiability conditions are satisfied for

$$\lambda := l_{\bullet\bullet} := \frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C} l_{ij}$$

$$\lambda_i^1 := l_{i\bullet} - \lambda \qquad l_{i\bullet} := \frac{1}{C}\sum_{j=1}^{C} l_{ij} \quad (i = 1, \ldots R)$$

$$\lambda_j^2 := l_{\bullet j} - \lambda \qquad l_{\bullet j} := \frac{1}{R}\sum_{i=1}^{R} l_{ij} \quad (j = 1, \ldots C)$$

$$\lambda_{ij}^{12} := l_{ij} - l_{i\bullet} - l_{\bullet j} + l_{\bullet\bullet}$$

In the following, all possible model specifications will be interpreted

Model 1: $l_{ij} = \lambda$

$l_{ij} = \lambda$

$\Rightarrow \quad l_{ij} = \log(mp_{ij}) = \lambda$

$\Rightarrow \quad p_{ij} = p$ and $1 = \sum_{i=1}^{R} \sum_{j=1}^{C} p_{ij} = RCp$

W^1 := rv of row classification (i. e. W^1 takes values in $1, \ldots, R$)
W^2 := rv of column classification (i. e. W^2 takes values in $1, \ldots, C$)

$$\Rightarrow p_{ij} = P(W^1 = i, W^2 = j)$$

- $p_{ij} = P(W^1 = i, W^2 = j)$

$$\Rightarrow P(W^1 = i) = \sum_{j=1}^{C} p_{ij} = Cp = \frac{1}{R} \quad \text{uniform distribution}$$

$$\Rightarrow P(W^2 = j) = \sum_{i=1}^{R} p_{ij} = Rp = \frac{1}{C} \quad \text{uniform distribution}$$

- W^1 and W^2 are independent:

$$P(W^1 = i, W^2 = j) = p_{ij} = p = \underbrace{(RCp)}_{=1} p = (Cp)(Rp) = P(W^1 = i)P(W^2 = j)$$

$\Rightarrow (W^1, W^2)$ are independent and uniform

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley & Sons.