6 Analysis of Contingency Tables

Masahiro KURODA

6.1 Introduction

This chapter presents analysis of log-linear models for contingency tables. In analysis of contingency tables, we are interested in associations and interactions among the categorical variables or interpretations of the parameters describing the model structure. Then our goal is to find a best model such that the model structure is simple and the model has few parameters.

The log-linear models which is used in analysis of contingency tables are a generalized linear model for counted data and can easily describe the variety of associations and interactions among the variables. To search a best model, we assess the effects on interaction terms in log-linear models by goodness of fit tests. The methodology for analysis of contingency tables is described in many books, for example, Bishop et al. (1975), Everitt (1977) and Agresti (2002).

This chapter is organized as follows: Section 6.2 introduces log-linear models and generalized linear models. Moreover we provide the procedures to find the best model by using goodness of fit tests for independence and comparing two log-linear models. Section 6.3 presents the XploRe functions to make inferences for log-linear models. Contingency table analysis using XploRe are illustrated in Section 6.4.

6.2 Log-linear models

Let $Y = (Y_1, Y_2, \ldots, Y_D)$ be categorical variables. Then a rectangular $(N \times D)$ data matrix consisting of N observations on Y can be rearranged as a D-way contingency table with cells defined by joint levels of the variables. Let $n_{ij\ldots t}$

denote the frequency for a cell $Y = (i, j, \ldots, t)$ and $n = \{n_{ij\ldots t}\}$. Suppose that Y has a multinomial distribution with an unknown parameter $\theta = \{\theta_{ij\ldots t}\}$, where $\theta_{ij\ldots t} \geq 0$ and $\sum \theta_{ij\ldots t} = 1$. The log-linear model is expressed in the form

$$\log \theta = X\lambda, \tag{6.1}$$

where X is a $D \times r$ design matrix and λ is an $r \times 1$ parameter vector. When Y has a Poisson distribution, the log-linear model is re-written by

$$\log m = X\lambda, \tag{6.2}$$

where $m = \{m_{ij\ldots t} = N\theta_{ij\ldots t}\}$ is the vector of expected frequencies.

6.2.1 Log-linear models for two-way contingency tables

Consider an $I \times J$ contingency table. The log-linear model is represented by

$$\log \theta_{ij} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12}, \tag{6.3}$$

for all i and j, under the constraints of the λ terms to sum to zero over any subscript such as

$$\sum_{i=1}^{I} \lambda_i^1 = 0, \qquad \sum_{j=1}^{J} \lambda_j^2 = 0, \qquad \sum_{i=1}^{I} \lambda_{ij}^{12} = \sum_{j=1}^{J} \lambda_{ij}^{12} = 0. \tag{6.4}$$

The log-linear model given by (6.3) is called the *saturated model* or the *full model* that there is statistically dependence between Y_1 and Y_2.

By analogy with analysis of variance models, we define the overall mean by

$$\lambda_0 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log \theta_{ij},$$

the main effects of Y_1 and Y_2 by

$$\lambda_i^1 = \frac{1}{J} \sum_{j=1}^{J} \log \theta_{ij} - \lambda_0,$$

$$\lambda_j^2 = \frac{1}{I} \sum_{i=1}^{I} \log \theta_{ij} - \lambda_0,$$

and the two-factor effect between Y_1 and Y_2 by

$$\lambda_{ij}^{12} = \log\theta_{ij} - (\lambda_i^1 + \lambda_j^2) - \lambda_0.$$

Then the main and two-factor effects are determined by the odds and odds ratios, and can be written by

$$\lambda_i^1 = \frac{1}{IJ}\sum_{i'=1}^{I}\sum_{j=1}^{J}\log\frac{\theta_{ij}}{\theta_{i'j}},$$

$$\lambda_j^2 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j'=1}^{J}\log\frac{\theta_{ij}}{\theta_{ij'}}$$

and

$$\lambda_{ij}^{12} = \frac{1}{IJ}\sum_{i'=1}^{I}\sum_{j'=1}^{J}\log\frac{\theta_{ij}\theta_{i'j'}}{\theta_{i'j}\theta_{ij'}}.$$

For the *independence model* that Y_1 is statistically independent of Y_2, the cell probability θ_{ij} can be factorized into the product of marginal probabilities θ_{i+} and θ_{+j}, that is,

$$\theta_{ij} = \theta_{i+}\theta_{+j},$$

where $\theta_{i+} = \sum_{j=1}^{J}\theta_{ij}$ and $\theta_{+j} = \sum_{i=1}^{I}\theta_{ij}$. Then the two-factor effect is

$$\lambda_{ij}^{12} = \frac{1}{IJ}\sum_{i'=1}^{I}\sum_{j'=1}^{J}\log\frac{\theta_{i+}\theta_{+j}\theta_{i'+}\theta_{+j'}}{\theta_{i'+}\theta_{+j}\theta_{i+}\theta_{+j'}} = 0,$$

so that the log-linear model for the independence model is expressed by

$$\log\theta_{ij} = \lambda_0 + \lambda_i^1 + \lambda_j^2,$$

for all i and j.

6.2.2 Log-linear models for three-way contingency tables

For an $I \times J \times K$ contingency table, the saturated log-linear model for the contingency table is

$$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{ik}^{13} + \lambda_{jk}^{23} + \lambda_{ijk}^{123},$$

for all i, j and k. The λ terms are also satisfied the constraints that

$$\sum_{i=1}^{I} \lambda_i^1 = \sum_{j=1}^{J} \lambda_j^2 = \sum_{k=1}^{K} \lambda_k^3 = 0,$$

$$\sum_{i=1}^{I} \lambda_{ij}^{12} = \sum_{j=1}^{J} \lambda_{ij}^{12} = \cdots = \sum_{k=1}^{K} \lambda_{jk}^{23} = 0,$$

$$\sum_{i=1}^{I} \lambda_{ijk}^{123} = \sum_{j=1}^{J} \lambda_{ijk}^{123} = \sum_{k=1}^{K} \lambda_{ijk}^{123} = 0.$$

We define the λ terms as follows: The overall mean is given by

$$\lambda_0 = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log \theta_{ijk}.$$

The main effects of Y_1, Y_2 and Y_3 are

$$\lambda_i^1 = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \log \theta_{ijk} - \lambda_0,$$

$$\lambda_j^2 = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \log \theta_{ijk} - \lambda_0,$$

$$\lambda_k^3 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log \theta_{ijk} - \lambda_0.$$

Each interaction effect is given by

$$\lambda_{ij}^{12} = \frac{1}{K} \sum_{k=1}^{K} \log \theta_{ijk} - (\lambda_i^1 + \lambda_j^2) - \lambda_0,$$

$$\lambda_{ik}^{13} = \frac{1}{J} \sum_{j=1}^{J} \log \theta_{ijk} - (\lambda_i^1 + \lambda_k^3) - \lambda_0,$$

$$\lambda_{jk}^{23} = \frac{1}{I} \sum_{i=1}^{I} \log \theta_{ijk} - (\lambda_j^2 + \lambda_k^3) - \lambda_0$$

and

$$\lambda_{ijk}^{123} = \log \theta_{ijk} - (\lambda_{ij}^{12} + \lambda_{ik}^{13} + \lambda_{jk}^{23}) - (\lambda_i^1 + \lambda_j^2 + \lambda_k^3) - \lambda_0.$$

Like log-linear models for two-ways contingency tables, the λ terms in the log-linear models for three-way contingency tables directly relate to the odds and odds ratios.

Here we introduce an important class of independence models that are called *hierarchical log-linear models*. In the hierarchical models, if high-order λ terms with certain variables are contained in the model, all lower-order λ terms with these same variables are included. For instance, when a log-linear model contains $\{\lambda_{ij}^{12}\}$, the log-linear model also contains $\{\lambda_i^1\}$ and $\{\lambda_j^2\}$. We assume that all log-linear models are the hierarchical model. Table 6.1 is the list of the hierarchical log-linear models for three-way contingency tables. Then interpretations of parameters in the log-linear models refer to the highest-order terms.

In log-linear models for conditional independence models, the two-factor effects correspond to *partial associations*. For instance, the log-linear model $[Y_1Y_2][Y_2Y_3]$ permits two-factor terms for associations between Y_1 and Y_2, and Y_2 and Y_3, but does not contain a two-factor term for an association between Y_1 and Y_3. Then the log-linear model $[Y_1Y_2][Y_2Y_3]$ specifies conditional independence between Y_1 and Y_3 given Y_2. In the log-linear model $[Y_1Y_2][Y_1Y_3][Y_2Y_3]$ called the *no three-factor interaction model*, there exists conditional dependence for all three pairs. Then the no three-factor interaction model has equal conditional odds ratios between any two variables at each level of the third variable. For example, the conditional odds ratio of Y_1 to Y_2 in the k-th level of Y_3 does not depend on k, and is given by

$$\log \frac{m_{ijk}m_{IJk}}{m_{iJk}m_{Ijk}} = \lambda_{ij}^{12} + \lambda_{IJ}^{12} - \lambda_{iJ}^{12} - \lambda_{Ij}^{12},$$

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

With multi-way contingency tables, the independence models are more complicated than the models for three-way contingency tables. The log-linear models can also describe easily several models for multi-way contingency tables. The basic principles of log-linear models for three-way contingency tables can be extended readily to multi-way contingency tables.

6.2.3 Generalized linear models

The log-linear model is a special case of generalized linear models (McCullagh and Nelder, 1989). For cell frequencies in contingency tables, the generalized

Table 6.1: Independence models for three-way contingency tables

Symbol	Log-linear model
Mutual independence	
$[Y_1]\,[Y_2]\,[Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3$
Joint independence from two-factors	
$[Y_1]\,[Y_2Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{jk}^{23}$
$[Y_1Y_2]\,[Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12}$
$[Y_1Y_3]\,[Y_2]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ik}^{13}$
Conditional independence	
$[Y_1Y_2]\,[Y_1Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{ik}^{13}$
$[Y_1Y_3]\,[Y_2Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ik}^{13} + \lambda_{jk}^{23}$
$[Y_1Y_2]\,[Y_2Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{jk}^{23}$
No three-factor interaction	
$[Y_1Y_2]\,[Y_1Y_3]\,[Y_2Y_3]$	$\log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{ik}^{13} + \lambda_{jk}^{23}$

linear models assume a Poisson distribution as the link function. Thus the log-linear models are given by equation (6.2).

Consider a 2×3 contingency table. From the constraints

$$\sum_{i=1}^{2}\lambda_i^1 = 0, \qquad \sum_{j=1}^{3}\lambda_j^2 = 0, \qquad \sum_{i=1}^{2}\lambda_{ij}^{12} = \sum_{j=1}^{3}\lambda_{ij}^{12} = 0, \tag{6.5}$$

the parameter vector is identified by

$$\lambda = \left[\lambda_0, \lambda_1^1, \lambda_1^2, \lambda_2^2, \lambda_{11}^{12}, \lambda_{12}^{12}\right]^T .$$

Thus the log-linear (6.2) can be written as

$$\begin{bmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{13} \\ \log m_{21} \\ \log m_{22} \\ \log m_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1^1 \\ \lambda_1^2 \\ \lambda_2^2 \\ \lambda_{11}^{12} \\ \lambda_{12}^{12} \end{bmatrix} .$$

When the maximum likelihood estimates (MLEs) of λ can not be found directly, iterative algorithms such as the Newton-Raphson and Fisher-scoring algorithms or the iterative proportional fitting procedure are applied. To compute the MLEs $\hat{\lambda}$ in log-linear models, XploRe uses the Newton-Raphson and Fisher-scoring algorithms.

6.2.4 Fitting to log-linear models

Chi-squared goodness of fit tests

To test a log-linear model against the saturated model, we estimate the expected frequencies of the log-linear model and evaluate the adequacy by the Pearson chi-squared statistic. When the MLEs $\hat{\lambda}$ in a log-linear model are obtained, the expected frequencies are estimated from

$$\hat{m} = \exp(X\hat{\lambda}).$$

To assess a log-linear model fitting to the data by comparing n to $\hat{m}$, the Pearson chi-squared statistic

$$\chi^2 = \sum_{i,j,\ldots,t} \frac{(n_{ij\ldots t} - \hat{m}_{ij\ldots t})^2}{\hat{m}_{ij\ldots t}}$$

is computed. As another measure of goodness of fit, the likelihood ratio test statistic is used. This test statistic is computed from

$$G^2 = 2 \sum_{i,j,\ldots,t} n_{ij\ldots t} \log \frac{n_{ij\ldots t}}{\hat{m}_{ij\ldots t}}.$$

If the sample size is sufficiently large, χ^2 and G^2 have an asymptotic chi-squared distribution with degrees of freedom (*df*) equal to the difference of the number of free parameters in the saturated model and a log-linear model. Then the chi-squared goodness of fit test can be conducted by the value of χ^2 or G^2.

Moreover the likelihood ratio test statistic G^2 can be used to compare two log-linear models M_1 and M_2. Then M_2 is nested in M_1, such that every nonzero λ terms in M_2 is contained in M_1. For example, the log-linear model $[Y_1Y_2][Y_3]$ is the nested model in the log-linear model $[Y_1Y_2][Y_2Y_3]$ and these models are expressed by

$$\begin{array}{ll} M_1: & \log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{jk}^{23}, \\ M_2: & \log\theta_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12}. \end{array}$$

Thus M_2 is simpler than M_1 and M_1 must hold when M_2 holds. Assuming that M_1 holds, we test whether M_2 fits the data as well as M_1. To compare two models, the following test statistic is used:

$$\triangle G^2 = G_2^2 - G_1^2, \tag{6.6}$$

where G_1^2 and G_2^2 are the likelihood ratio test statistics for M_1 and M_2. Then $\triangle G^2$ also has an asymptotic chi-squared distribution with *df* equal to *df* for M_1- *df* for M_2.

When the value of $\triangle G^2$ is in a critical region, we conclude that M_2 provides a better description of the data. Furthermore $\triangle G^2$ is computed to compare a nested model in M_2 with M_2. If the value of $\triangle G^2$ is out a critical region, we re-compare another nested model in M_1 with M_1. Repeating goodness of fit tests to compare nested models, we find a best model.

As another criteria to compare nested models, the Akaike information criteria (AIC) and Bayesian information criteria (BIC) are well known.

Model residuals

The goodness of fit statistic gives the summary of how a log-linear model fits to the data. We examine lack of fit by comparing observed data to the fitted data individually.

For cell (i, j) in a two-way contingency table, the Pearson standardized residual is defined by

$$e_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}}.$$

The Pearson residual is also related to the Pearson chi-squared test statistics by

$$\chi^2 = \sum_{i,j} e_{ij}^2.$$

When a log-linear model holds, the residuals $\{e_{ij}\}$ have an approximate normal distribution with mean 0. Then, by checking whether the Pearson residuals are larger than about ±2 that is standard normal percentage points, we detect the presence of the data that are influential on the fit of a log-linear model.

6.3 Inference for log-linear models using XploRe

To make inferences for log-linear models, we use the functions in the glm library. The library is available by

```
library("glm")
```

6.3.1 Estimation of the parameter vector λ

The parameter vector λ can be estimated by using the glmest function

```
glmest("polog", x, n)
```

where polog is a Poisson distribution with logarithm link function, x is the design matrix and n is the cell frequencies for contingency tables. Executing

```
lambda = glmest("polog", x, n)
```

the estimates of λ are assigned to the variable lambda. Then lambda contains the following output:

b : the estimated parameter vector λ

bv : the estimated covariance of b

stat : several statistics

The expected frequencies m are also computed from

```
m = exp(x*lambda.b)
```

Moreover the glmest function and other functions in glm library can be also specified several options by defining opt with glmopt. For the optional parameters, refer to Härdle et al. (1999) or Help function in XploRe.

6.3.2 Computing statistics for the log-linear models

A number of statistical characteristics can be computed using the `glmstat` function. Then statistical characteristics can be obtained from

```
stat = glmstat("polog", x, m, lambda.b, lambda.bv)
```

and are stored in the output `stat`:

`df` : degrees of freedom

`deviance` : the deviance of the estimated model

`pearson` : the Pearson statistic

`loglik` : the log-likelihood of the estimated model, using the estimated dispersion parameter

`aic, bic` : Akaike's AIC and Schwarz' BIC criterion, respectively

`r2, adr2` : the (pseudo) coefficient of determination and its adjusted version, respectively

`it` : the number of iterations needed

`ret` : the return code

`nr` : the number of replicated observation in `x`, if they were searched for.

6.3.3 Model comparison and selection

The computation of the likelihood ratio test statistic for comparing nested models can be performed by the `glmlrtest` function:

```
{lr, pvalue} = glmlrtest(loglik2, df2, loglik1, df1)
```

where `loglik1` and `loglik2` are the log-likelihoods for the log-linear models M_1 and M_2. Note that M_2 must be the nested model in M_1. These values are obtained from the `glmstat` function. The augments `df1` and `df2` are *df*s for each model. Executing the above call, the test statistic `lr` and the p-value `pvalue` are yielded.

Moreover, to select the best model automatically, XploRe has the `glmselect`, `glmforward` and `glmbackward` functions. The `glmselect` function performs a complete search model selection, the `glmforward` and `glmbackward` functions do the forward and backward search model selections, respectively. The syntax of these functions is the same as `glmest`. Note that best models found by these functions are not always hierarchical log-linear models. Then we repeat to compute the likelihood ratio statistics for comparing the nested models and finally find the best model. When the parameters λ_0, $\{\lambda_i^A\}$ and $\{\lambda_j^B\}$ are contained in all models, the optional parameter `fix` that specifies them is described as follows:

```
opt = glmopt("fix", 1|2|3)
```

To search the best model by using the backward search model selection, we type

```
select = glmbackward("polog", x, m, opt)
```

Then the output list `select` consists of five components:

`best` : the five best models

`bestcrit` : a list containing `bestcrit.aic` and `bestcrit.bic`, the Akaike and Schwarz criteria for the five best models

`bestord` : the best models of each order

`beatordcrit` : like `bestcrit`, but for the best model for each order

`bestfit` : containing `bestfit.b`, `bestfit.bv` and `bestfit.stat`, the estimation results for the best model

6.4 Numerical analysis of contingency tables

6.4.1 Testing independence

Chi-squared test

The data in Table 6.2 are a cross-sectional study of malignant melanoma taken from Roberts et al. (1981) and treated in Dobson (2001). For the two-way table, we are interested in whether there exists a association between Tumor type and Site.

Table 6.2: Contingency table with tumor type and site

Tumor Type (i)	Site (j)	Cell frequency
Hutchinson's melanotic freckle (H)	Head & neck (HN)	22
	Trunk (T)	2
	Extremities (E)	10
Superficial spreading melanoma (S)	Head & neck (HN)	16
	Trunk (T)	54
	Extremities (E)	115
Nodular (N)	Head & neck (HN)	19
	Trunk (T)	33
	Extremities (E)	73
Indeterminate (I)	Head & neck (HN)	11
	Trunk (T)	17
	Extremities (E)	28

Table 6.3: Expected frequencies for the independence model

Tumor Type	Site	Cell frequency
Hutchinson's melanotic freckle	Head & neck	5.78
	Trunk	9.01
	Extremities	19.21
Superficial spreading melanoma	Head & neck	31.45
	Trunk	49.03
	Extremities	104.52
Nodular	Head & neck	21.25
	Trunk	33.13
	Extremities	70.62
Indeterminate	Head & neck	9.52
	Trunk	14.84
	Extremities	31.64

Let $m = \{m_{ij}\}$ be the expected frequencies for the contingency table. The log-linear model that Tumor type is independent of Site is expressed by

$$\log m_{ij} = \lambda_0 + \lambda_i^{Type} + \lambda_j^{Site}, \tag{6.7}$$

for all i and j. From the constraints

$$\lambda_H^{Type} + \lambda_S^{Type} + \lambda_N^{Type} + \lambda_I^{Type} = \lambda_{HN}^{Site} + \lambda_T^{Site} + \lambda_E^{Site} = 0,$$

the parameter vector λ for the independence model is identified by

$$\lambda = [\lambda_0, \lambda_H^{Type}, \lambda_S^{Type}, \lambda_N^{Type}, \lambda_{HN}^{Site}, \lambda_T^{Site}].$$

To find the expected frequencies, we estimate the MLEs $\hat{\lambda}$ using the following statements:

```
library("glm")
n=#(22,2,10,16,54,115,19,33,73,11,17,28)
x=read("design.dat")
lambda = glmest("polog", x, n)
```

where `design.dat` is specified by

```
1 -1 -1 -1 -1 -1
1 -1 -1 -1  1  0
1 -1 -1 -1  0  1
1  1  0  0 -1 -1
1  1  0  0  1  0
1  1  0  0  0  1
1  0  1  0 -1 -1
1  0  1  0  1  0
1  0  1  0  0  1
1  0  0  1 -1 -1
1  0  0  1  1  0
1  0  0  1  0  1
```

The expected frequencies shown in Table 6.3 can be obtained by

```
m = exp(x*lambda.b)
```

and are compared to the data in Table 6.2 by using χ^2. The value of χ^2 is computed from

```
      lambda.stat
```

or

```
      stat = glmstat("polog", x, m, lambda.b, lambda.bv)
      stat.pearson
```

Then χ^2 of 65.8 is very significant compared to the chi-square distribution with 6 *df* and indicates that the independence model does not fit to the data. We can conclude that there exists the association between Tumor type and Site.

Note that the function `crosstable` provides the chi-squared statistic for testing independence for two-way contingency tables.

Model residuals

Table 6.4: Pearson residuals for the independence model

Tumor Type	Site	Residual
Hutchinson's melanotic freckle	Head & neck	6.75
	Trunk	2.34
	Extremities	-2.10
Superficial spreading melanoma	Head & neck	-2.76
	Trunk	0.71
	Extremities	1.03
Nodular	Head & neck	-0.49
	Trunk	-0.02
	Extremities	0.28
Indeterminate	Head & neck	0.48
	Trunk	0.56
	Extremities	-0.65

Table 6.4 shows the Pearson standardized residuals for the fit of the independence model. The values are easily computed from

```
      e = (n-m)/sqrt(m)
```

We can see that the residual for Hutchinson's melanotic freckle and Head & neck reflects the overall poor fit, because the value of $6.75^2 = 45.56$ is related to $\chi^2 = 65.8$.

6.4.2 Model comparison

Chi-squared test

The data in Table 6.5 summarize to a survey the Wright State University school of Medicine and the United Health Services in Dayton, Ohio. The analysis for the contingency table is given in Agresti (2002). For the three-way table, we search the best model by using the likelihood ratio tests.

Table 6.5: Alcohol, cigarette and marijuana use for high school seniors

Alcohol use (A)	Cigarette use (C)	Marijuana use (M)	Cell frequency
Yes	Yes	Yes	911
		No	538
	No	Yes	44
		No	456
No	Yes	Yes	3
		No	43
	No	Yes	2
		No	279

Table 6.6 shows the expected frequencies for log-linear models of no three-factor interaction and conditional independence models. The expected frequencies for each model are computed by using `glmest`.

The expected frequencies for the log-linear model $[AC][AM][CM]$ are found using the following statements:

```
library("glm")
n=#(911, 538, 44, 456, 3, 43, 2, 279)
x=read("design.dat")
lambda = glmest("polog", x, n)
m = exp(x*lambda.b)
```

Table 6.6: Expected frequencies for log-linear models applied to Table 6.5

			Log-linear model			
A	C	M	$[AC][AM]$	$[AM][CM]$	$[AC][CM]$	$[AC][AM][CM]$
Yes	Yes	Yes	710.00	909.24	885.88	910.38
		No	175.64	438.84	133.84	538.62
	No	Yes	131.05	45.76	123.91	44.62
		No	2005.80	555.16	470.55	455.38
No	Yes	Yes	5.50	4.76	28.12	3.62
		No	24.23	142.16	75.22	42.38
	No	Yes	1.02	0.24	3.93	1.38
		No	276.70	179.84	264.45	279.62

Then, under the constraints with the λ terms, the parameter vector λ is identified by

$$\lambda = [\lambda_0, \lambda^A_{Yes}, \lambda^C_{Yes}, \lambda^M_{Yes}, \lambda^{AC}_{Yes,Yes}, \lambda^{AM}_{Yes,Yes}, \lambda^{CM}_{Yes,Yes}]^T,$$

and the design matrix `x` is specified by

```
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
```

To compute the expected frequencies of the nested models in the log-linear model $[AC][AM][CM]$, we delete the columns of `x` corresponding to λ setting to zero in these models and then execute the above statements. For example, deleting the seventh column of `x`, we can obtain the expected frequencies of the log-linear model $[AC][AM]$. The command

```
x[,1|2|3|4|5|6]
```

produces the design matrix for the log-linear model $[AC][AM]$.

Table 6.7 shows results of the likelihood ratio and Pearson chi-squared tests for log-linear models. The statements to compute the values of G^2 for the saturated model M_1 and a log-linear model M_2 are

```
stat1 = glmstat("polog", x1, n, lambda1.b, lambda1.bv)
df1 = rows(lambda1.b)
stat2 = glmstat("polog", x2, n, lambda2.b, lambda2.bv)
df2 = rows(lambda2.b)
{lr,pvalue}=glmlrtest(stat2.loglik, df2, stat1.loglik, df1)
lr
pvalue
```

where the design matrix x1 for the saturated model is specified by

```
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
```

The value of χ^2 is also computed from

```
lambda = glmest("polog",x,n)
lambda.stat
```

Then the values of G^2 and χ^2 or p-value indicate that the model $[AC][AM][CM]$ fits well to the data.

Model residuals

To examine lack of fit to the data, we analyze the residuals for each log-linear model. Table 6.8 shows the Pearson standardized residuals for the log-linear models. All residuals for the log-linear model $[AC][AM][CM]$ are very small and demonstrate that the model well fits to the data. On the other hand, the residuals for conditional independence models indicate poorly fit to the data. In particular, the extremely large residuals of -34.604 for the model $[AC][AM]$ and of 34.935 for the model $[AC][CM]$ cause the lack of fit to the data.

Table 6.7: Goodness of fit tests for log-linear models

Log-linear model	G^2	χ^2	Degrees of freedom	p-value
$[AC][AM][CM]$	0.4	0.4	1	0.53
$[AC][AM]$	497.4	443.8	2	0.00
$[AM][CM]$	187.8	177.6	2	0.00
$[AC][CM]$	92.0	80.8	2	0.00

Table 6.8: The Pearson standardized residuals for log-linear models

			Log-linear model			
A	C	M	$[AC][AM]$	$[AM][CM]$	$[AC][CM]$	$[AC][AM][CM]$
Yes	Yes	Yes	7.543	0.058	0.844	0.020
		No	27.342	4.734	34.935	-0.027
	No	Yes	-7.604	-0.260	-7.179	-0.092
		No	-34.604	-4.209	-0.671	0.029
No	Yes	Yes	-1.077	-0.807	-4.737	-0.324
		No	3.813	-8.317	-3.715	0.095
	No	Yes	0.969	3.596	-0.975	0.524
		No	0.138	7.394	0.895	-0.037

Test for partial associations

Moreover, it is possible to compare nested log-linear models by testing partial associations. We test to compare the model $[AC][AM]$ with the model $[AC][AM][CM]$. Then the test examines whether there exists a partial association between Alcohol use and Cigarette use, that is,

$$\lambda_{11}^{CM} = \lambda_{12}^{CM} = \lambda_{21}^{CM} = \lambda_{22}^{CM} = 0.$$

Each log-linear model is expressed by

$$\begin{array}{ll} M_1: & \log m_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}, \\ M_2: & \log m_{ijk} = \lambda_0 + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ik}^{AC} + \lambda_{jk}^{AM}. \end{array}$$

From Table 6.7,

$$\triangle G^2 = 497.4 - 0.4 = 497.0$$

and $df = 2-1 = 1$, so that $\triangle G^2$ provides strong evidence of a partial association between Cigarette use and Marijuana use. We can also test for partial associations by comparing the models $[AM][CM]$ and $[AC][CM]$ with the model $[AC][AM][CM]$.

Search for the best model

Next we illustrate the best model search using the `glmbackward` function. Using

```
opt = glmopt("fix", 1|2|3|4)
```

we specify to contain λ_0, $\{\lambda_i^A\}$, $\{\lambda_j^C\}$ and $\{\lambda_k^M\}$ in the model. To choose a best model, we execute

```
select=glmbackward("polog", x, n, opt)
select.best
```

where `x` is the design matrix for the saturated model. Then `select.best` displays the five best models for the data:

```
Contents of best
[1,]        1        1        1        1        1
[2,]        2        2        2        2        2
[3,]        3        3        3        3        3
[4,]        4        4        4        4        4
[5,]        5        5        0        5        5
[6,]        6        6        6        6        0
[7,]        7        7        7        0        7
[8,]        0        8        8        8        8
```

In the above output, each row corresponds to the parameter vector λ in the saturated log-linear model as follows:

row	λ term
1	λ_0
2	$\{\lambda^{A}_{Yes}\}$
3	$\{\lambda^{C}_{Yes}\}$
4	$\{\lambda^{M}_{Yes}\}$
5	$\{\lambda^{AC}_{Yes,Yes}\}$
6	$\{\lambda^{AM}_{Yes,Yes}\}$
7	$\{\lambda^{CM}_{Yes,Yes}\}$
8	$\{\lambda^{ACM}_{Yes,Yes,Yes}\}$

Those components that are not contained in a log-linear model are indicated by zero. The first column shows the no three-interaction model, since the row `8` is zero. The second column represents the saturated model. The last three columns are not the hierarchical models. Therefore the model $[AC][AM][CM]$ is also selected as the best model. The output `select.bestfit` includes all estimation results with the best model.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis* 2nd edition, John Wiley & Sons, New York.

Bishop, Y.M., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis*, M.I.T. Press, Cambridge, MA.

Dobson, A.J. (2001). *An Introduction to Generalized Linear Models* 2nd edition, Chapman & Hall, London.

Everitt, B.S. (1977). *The analysis of Contingency Tables*, Chapman & Hall, London.

Härdle, W., Klinke, S. and Müller, M. (1999). *XploRe—Learning Guide*, Springer-Verlag, Berlin.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* 2nd edition, Chapman & Hall, London.

Roberts, G., Martyn, A.L., Dobson, A.J. and McCarthy, W.H. (1981). Tumour thickness and histological type in malignant melanoma in New South Wales, Australia, 1970-76, *Pathology* **13**: 763-770.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis* 2nd edition, John Wiley & Sons, New York.

Bishop, Y.M., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis*, M.I.T. Press, Cambridge, MA.

Dobson, A.J. (2001). *An Introduction to Generalized Linear Models* 2nd edition, Chapman & Hall, London.

Everitt, B.S. (1977). *The analysis of Contingency Tables*, Chapman & Hall, London.

Härdle, W., Klinke, S. and Müller, M. (1999). *XploRe—Learning Guide*, Springer-Verlag, Berlin.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* 2nd edition, Chapman & Hall, London.

Roberts, G., Martyn, A.L., Dobson, A.J. and McCarthy, W.H. (1981). Tumour thickness and histological type in malignant melanoma in New South Wales, Australia, 1970-76, *Pathology* **13**: 763-770.