Chapter 1

The Analysis of Contingency Tables: Log-Linear and Graphical Models

G Tutz - preliminary

Contingency tables or cross-classified data come in various forms, differing in dimensions, distributional assumptions, and margins. In general, they may be seen as a structured way of representing count data. They were already used to represent data in binary and multinomial regression problems when explanatory variables were categorical (Chapter **??** and **??**). Also count data with categorical explanatory variables (Chapter **??**) may be given in the form of contingency tables.

In this chapter log-linear models are presented that may be seen as regression models or association models depending on the underlying distribution. Three types of distributions are considered, the Poisson distribution, the multinomial, and the product-multinomial distribution. When the underlying distribution is a Poisson distribution one usually considers regression problems as in Chapter **??**. When the underlying distribution is multinomial or product-multinomial one usually has more structure in the multinomial response than is considered in the regression problems in Chapter **??** and **??**. In these chapters the response is assumed to be multinomial without further structuring whereas in the present chapter the multinomial response arises from the consideration of several response variables which together form a contingency table. Then one wants to analyse the association between these variables. Log-linear models provide a common tool to investigate the association structure in terms of independence or conditional independence between variables.

Several examples of contingency tables have already been given in previous chapters. Two more examples are the following.

Example 1.1: Birth Data

In a survey study several variables have been collected that are linked to the birth process (see also Boulesteix (2006)). Table 1.1 shows the data for the variables gender of the child (G, 1:male, 2:female), if membranes did rupture before the beginning of labour (M, 1:yes, 0:no), if Cesarean section has been applied (C, 1:yes, 0:no) and if birth has been induced (I, 1:yes, 0:no). The association between the four variables is unknown and shall be investigated. □

			Induced	
			0	1
Gender	Membranes	Cesarean		
1	0	0	177	45
		1	37	18
	1	0	104	16
		1	9	7
2	0	0	137	53
		1	24	12
	1	0	74	15
		1	8	2

TABLE 1.1: Contingency table for birth data with variables gender (G), membranes (M), Cesarean section (C) and induced birth (I)

Example 1.2: Leukoplakia

Table 1.2 shows data from a study on leukoplakia, which is a clinical term used to describe patches of keratosis visible as adherent white patches on the membranes of the oral cavity. It shows the alcohol intake in grams of alcohol, smoking habits and presence of leukoplakia. The objective is the analysis of the association between disease and risk factors (data are taken from Hamerle and Tutz (1980)). □

1.1 Types of Contingency Tables

In particular three types of contingency tables and the corresponding scientific questions will be studied. The *first type* of contingency table occurs if cell counts are Poisson-distributed given the configuration of cells. Then the counts itself represent the response and the categorical variables that determine the cells are the explanatory variables. For example, in Table **??** in Chapter **??** the number of firms with insolvency problems may be considered as the response while year and month represent the explanatory variables. The total number of insolvent firms is not fixed beforehand and is itself a realization of a random variable.

In the *second type* of contingency table a fixed number of subjects is observed and the cell counts represent the multivariate response given the total number of

		Leukoplakia (A)	
		yes	no
Alcohol	Smoker		
no	yes	26	10
	no	8	8
$(0g, 40g]$	yes	38	8
	no	43	24
$(40g, 80g]$	yes	4	1
	no	14	17
$> 80g$	yes	1	0
	no	3	7

TABLE 1.2: Contingency table for oral leukoplakia

observations. The common assumption is that cell counts have multinomial distribution. Contingency tables of this types occur if a fixed number of individuals is cross-classified with respect to variables like gender and preference for distinct political parties (see Table **??** in Chapter **??**. The analysis for this type of contingency table may focus on the association between gender and preference for parties. Alternatively, one might be interested in modelling the preference as the response given gender as explanatory variable.

The *third type* of contingency tables is found for example in clinical trials. Table **??** shows cross-classified data which have been collected by randomly allocating patients to one of two groups, a treatment group and a group where a placebo is given. After ten days of treatment the pain occuring during movement of the knee is assessed on a five point scale. The natural response in this example is the level of pain given the treatment group. The number of people in the two groups is fixed while the counts themselves are random variables. The level of pain, given the treatment group, is a multivariate response, usually modelled by a multinomial distribution.

In general, two-way $(I \times J)$-contingency tables with I rows and J columns may be described by

$$\begin{array}{ll} X_{ij} & = \text{counts in cell} \quad (i,j), \\ X_A \in \{1,\ldots,I\} & \quad \text{representing the rows,} \\ X_B \in \{1,\ldots,J\} & \quad \text{representing the columns.} \end{array}$$

The observed contingency table has the form

$$\begin{array}{cc|cccc|c} & & & & X_B & & \\ & & 1 & 2 & \ldots & J & \\ \hline & 1 & X_{11} & X_{12} & \ldots & X_{1J} & X_{1+} \\ & 2 & X_{21} & \ddots & & \vdots & \vdots \\ X_A & \vdots & \vdots & & \ddots & \vdots & \\ & I & X_{I1} & \ldots & & X_{IJ} & X_{I+} \\ \hline & & X_{+1} & \ldots & & X_{+J} & \end{array}$$

where $X_{i+} = \sum_{j=1}^{J} X_{ij}, X_{+j} = \sum_{i=1} X_{ij}$ denote the marginal counts. The subscript " + " denotes the sum over that index.

The three types of contingency tables may be distinguished by the distribution that is assumed.

Type 1: Poisson distribution (Number of insolvent firms)

It is assumed that $X_{11}, \ldots, X_{IJ}$ are independent Poisson-distributed random variables, $X_{ij} \sim P(\lambda_{ij})$. The total number of counts $X_{11} + \cdots + X_{IJ}$ as well as the marginal counts are random variables. The natural model considers the counts as response and X_A and X_B as explanatory variables.

Type 2: Multinomial distribution (Gender and preference for political parties)

For a fixed number of subjects one observes independently the response tupel (X_A, X_B) with possible outcomes $\{(1,1), \ldots, (I,J)\}$. The cells of the table represent the IJ possible outcomes. The resulting cell counts follow a multinomial distribution $(X_{11}, \ldots, X_{IJ}) \sim M(n, (\pi_{11}, \ldots, \pi_{IJ}))$ where $\pi_{ij} = P(X_A = i, X_B = j)$ denotes the probability of a response in cell (i,j). The probabilities $\{\pi_{ij}\}$, or in vector form $\boldsymbol{\pi}^T = (\pi_{11}, \ldots \pi_{IJ})$, represent the joint distribution of X_A and X_B.

Type 3: Product-multinomial distribution (Treatment and pain)

In contrast to type 2 now one set of marginal counts is fixed. In the treatment and pain example the row tables are fixed by $n_1 = X_{1+}, n_2 = X_{2+}$. Given the treatment group one observes for each individual the response $X_B \in \{1, \ldots, J\}$. The cell counts for given treatment group i follow a multinomial distribution

$$(X_{i1}, \ldots, X_{iJ}) \sim M(n_i, (\pi_{i1}, \ldots \pi_{iJ})),$$

where π_{ij} now denotes the *conditional probability* $\pi_{ij} = P(X_B = j | X_A = i)$. The cell counts of the total contingency table follow a *product-multinomial* distribution. The natural modelling is to consider X_B as the response variable and X_A as the explanatory variable. This modelling approach follows directly from the design of the study. Of course, if the column totals are fixed, the natural response is X_A with X_B as the explanatory variable.

Models and Types of Contingency Tables

The three types of contingency tables differ by the way how data are collected. When considering typical scientific questions, to be investigated by the analysis of contingency tables, one finds a hierarchy within the types of tables. While the Poisson distribution is the most general, allowing for various types of analysis, the product-multinomial contingency table is the most restrictive. The hierarchy is due to the possible transformations of distributions by conditioning.

Poisson and Multinomial Distribution

Let $X_{ij}, i = 1, \ldots, I, j = 1, \ldots, J$ follow independent Poisson distributions, $X_{ij} \sim P(\lambda_{ij})$. Then the *conditional distribution* of $(X_{11}, \ldots, X_{IJ})$ given $n = \sum_{ij} X_{ij}$ is multinomial. More concrete, one has

$$(X_{11}, \ldots, X_{IJ}) | \sum_{i,j} X_{ij} = n \sim M(n, (\frac{\lambda_{11}}{\lambda}, \ldots, \frac{\lambda_{IJ}}{\lambda})),$$

where $\lambda = \sum_{ij} \lambda_{ij}$. Therefore, given one has Poisson-distributed cell counts, by conditioning one obtains a structured multinomial distribution which is connected to the response tupel (X_A, X_B). Hence, by conditioning on n one may study the response (X_A, X_B), its marginal distribution as well as the association between (X_A, X_B).

Multinomial and Product-Multinomial Distribution

Let $(X_{11}, \ldots, X_{IJ})$ have multinomial distribution, $(X_{11}, \ldots, X_{IJ}) \sim M(n, (\pi_{11}, \ldots, \pi_{IJ}))$. By conditioning on the row margins $n_{i+} = \sum_j X_{ij}$ one obtains the product-multinomial distribution with probability mass function

$$f(x_{11}, \ldots, x_{IJ}) = \prod_{i=1}^{I} \frac{n_{i+}!}{x_{i1}! \ldots x_{iJ}!} \pi_{1|i}^{x_{i1}} \ldots \pi_{J|i}^{x_{IJ}},$$

where $\pi_{j|i} = \pi_{ij} / \sum_j \pi_{ij} = \pi_{ij} / \pi_{i+}$. Thus the cell counts of one row given n_{i+} have multinomial distribution,

$$(X_{i1}, \ldots, X_{iJ}) \sim M(n_{i+}, (\pi_{1|i}, \ldots, \pi_{J|i}))$$

and the I multinomials corresponding to rows are independent.

If the Poisson distribution generates the counts in the table one may consider the counts given X_A and X_B within a regression framework; or one may condition on the total sample size n and model the marginal distribution and the association of X_A and X_B based on the multinomial distribution. One may also go one step further and condition on the marginal counts of the rows (columns) and consider the regression model where X_B (X_A) is the response and X_A (X_B) the explanatory variable. If the multinomial distribution generates the contingency table one may consider (X_A, X_B) as response or choose one of the two variables by conditioning on the other one. In this sense the Poisson distribution contingency table is the most versatile. Since the Poisson, multinomial and product-multinomial distribution may be treated within a general framework, in the following $\mu_{ij} = E(X_{ij})$ is used rather than $n\pi_{ij}$ (or $n_i \pi_{ij}$) which would be more appropriate for the multinomial (or product-multinomial) distribution.

In Table 1.3 the types of distributions and the modelling approaches are summarized. Most modelling approaches are regression problem approaches. The association between X_A and X_B may also be considered as a limiting case of regression, namely without explanatory variables. It should be noted that there are also approaches to model the dependence of X_B on X_A if the column marginals are fixed. However, then special consideration is necessary (see Section?)....

Poisson distribution	Regression $(X_A, X_B) \rightarrow$ Counts
	Association between X_A and X_B (conditional on n)
	Regression $X_A \rightarrow X_B$ (conditional on X_{i+})
	Regression $X_B \rightarrow X_A$ (conditional on X_{+j})
Multinomial distribution	Association between X_A and X_B (conditional on n)
	Regression $X_A \rightarrow X_B$ (conditional on X_{i+})
	Regression $X_B \rightarrow X_A$ (conditional on X_{+j})
Product-multinomial distribution	
X_{i+} fixed	Regression $X_A \rightarrow X_B$
X_{+j} fixed	Regression $X_B \rightarrow A_A$

TABLE 1.3: Types of two-way contingency tables and modelling approaches.

1.2 Log-Linear Models for Two-Way Tables

Consider an $(I \times J)$-contingency table $\{X_{ij}\}$. Let $\mu_{ij} = E(X_{ij})$ denote the mean, where $\mu_{ij} = n\pi_{ij} = nP(X_A = i, X_B = j)$ for the multinomial distribution, $\mu_{ij} = n_{i+}P(X_B = j|X_A = i)$ if one conditions on $n_{i+} = \sum_j X_{ij}$, and $\mu_{ij} = n_{+j}P(X_A = i|X_B = j)$ if one conditions on $n_{+j} = \sum_i X_{ij}$. The general log-linear model for two-way tables has the form

$$\log(\mu_{ij}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{AB(ij)} \tag{1.1}$$

or equivalently

$$\mu_{ij} = e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}} e^{\lambda_{AB(ij)}}.$$

Since model (1.1) contains too many parameters, identifiability requires constraints on the parameters. Two sets of constraints are in common use, the symmetrical constraints and constraints that use a baseline parameter.

Symmetrical constraints:

$$\sum_{i=1}^{I} \lambda_{A(i)} = \sum_{j=1}^{J} \lambda_{B(j)} = \sum_{i=1}^{I} \lambda_{AB(ij)} = \sum_{j=1}^{J} \lambda_{AB(ij)} = 0 \quad \text{for all } i, j.$$

Baseline parameters set to zero:

$$\lambda_{A(I)} = \lambda_{B(J)} = \lambda_{AB(iJ)} = \lambda_{AB(Ij)} = 0 \quad \text{for all } i, j.$$

The symmetrical constraints are identical to the constraints used in analysis-of-variance (ANOVA). In ANOVA the dependence of a response variable on categorical variables, called factors, is studied. In particular one is often interested in

interaction effects. There is a strong similarity between ANOVA and Poisson contingency tables where the counts represent the response and the categorical variables X_A and X_B form the design. The main difference is that ANOVA models assume normal distribution for the response whereas in log-linear models for count data the response is integer-valued.

These sets of constraints are closely related to the coding of dummy variables. Symmetrical constraints refer to effect coding whereas the choice of baseline parameters is equivalent to choosing a reference category in dummy coding (see Section **??**). Model (1.1) may be also written with dummy variables yielding

$$\begin{aligned}\log(\mu_{ij}) = \lambda_0 &+ \lambda_{A(1)}x_{A(1)} + \cdots + \lambda_{A(I-1)}x_{A(I-1)} \\ &+ \lambda_{B(1)}x_{B(1)} + \cdots + \lambda_{B(J-1)}x_{B(J-1)} \\ &+ \lambda_{AB(1,1)}x_{A(1)}x_{B(1)} + \cdots + \lambda_{AB(I-1,J-1)}x_{A(I-1)}x_{B(J-1)},\end{aligned}$$

where $x_{A(1)}, \ldots$ are dummy variables coding $A = i$ and $x_{B(1)}, \ldots$ are dummy variables coding $B = j$. This form is usually too clumsy and will be avoided. However, it is easily seen that effect coding of dummy variables is equivalent to the symmetric constraints and choosing $(X_A = I, X_B = J)$ as reference categories in dummy coding is equivalent to using baseline parameters. One should keep in mind that baseline parameters which refer to reference categories may be chosen arbitrarily, different software uses different constraints.

The sets of constraints given above apply for Poisson distribution tables. For multinomial and product-multinomial tables additional constraints are needed to ascertain that $\sum_{ij} X_{ij} = n$ (multinomial) and $\sum_j X_{ij} = n_{i+}$ (product-multinomial, fixed row sums) holds.

Additional constraint for multinomial tables:

$$\sum_{i,j} e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}} e^{\lambda_{AB(ij)}} = n.$$

Additional constraints for product-multinomial tables:

$$\sum_{j=1}^{J} e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}} e^{\lambda_{AB(ij)}} = n_{i+}, \quad i = 1, \ldots, I \quad (\text{for } n_{i+} \text{ fixed}),$$

$$\sum_{i=1}^{I} e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}} e^{\lambda_{AB(ij)}} = n_{+j}, \quad j = 1, \ldots, J \quad (\text{for } n_{+j} \text{ fixed}).$$

Model (1.1) is the most general model for two-way contingency tables, the so-called *saturated model*. It is saturated since it represents only a reparameterization of the means $\{\mu_{ij}\}$, any set of means $\{\mu_{ij}\}$ $(\mu_{ij} > 0)$ may be represented by parameters $\lambda_B, \lambda_{A(i)}, \lambda_{B(j)}, \lambda_{AB(ij)}, i = 1, \ldots, I, j = 1, \ldots, J$.

Log-linear Model for Two-Way Tables

$$\log(\mu_{ij}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{AB(i,j)}$$

Constraints:

$$\sum_{i=1}^{I} \lambda_{A(i)} = \sum_{j=1}^{J} \lambda_{B(j)} = \sum_{i=1}^{I} \lambda_{AB(ij)} = \sum_{j=1}^{J} \lambda_{AB(ij)} = 0$$

or

$$\lambda_{A(I)} = \lambda_{B(J)} = \lambda_{AB(iJ)} = \lambda_{AB(Ij)} = 0$$

TABLE 1.4: Log-linear model for two-way tables

Consequently not much insight is gained by considering the saturated log-linear model. The most important submodel is the *log-linear model of independence*

$$\log(\mu_{ij}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)}, \tag{1.2}$$

where it is assumed that $\lambda_{AB(ij)} = 0$. This is no longer a saturated model since it implies severe restrictions on the underlying regression or association structure. The restriction has different meanings, depending on the distribution of the cell counts X_{ij}. For the Poisson distribution it simply means that there is no interaction effect of variables X_A and X_B when effecting on the cell counts. For the multinomial model it is helpful to consider the multiplicative form of (1.2).

$$\mu_{ij} = nP(X_A = i, X_B = j) = e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}}. \tag{1.3}$$

That means that the probability $P(X_A = i, X_B = j)$ may be written in a multiplicative form with factors depending only on X_A or X_B. Taking constraints into account, it is easily derived that (1.3) is equivalent to assuming that X_A and X_B are independent random variables, or equivalently that $P(X_A = i, X_B = j) = P(X_A = i)P(X_B = j)$ holds. That property gives the model its name.

For the product-multinomial table (row marginals n_{i+} fixed) one has

$$\mu_{ij} = n_{i+}P(X_B = j|X_A = i) = e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}}.$$

With the constraint $\sum_j e^{\lambda_0} e^{\lambda_{A(i)}} e^{\lambda_{B(j)}} = n_{i+}$ one obtains

$$P(X_B = j|X_A = i) = e^{\lambda_{B(j)}} / \sum_r e^{\lambda_{B(r)}},$$

which means that the response X_B does not depend on variable X_A. Thus the model postulates that the response probabilities are identical across rows

$$P(X_B = j|X_A = 1) = \ldots = P(X_B = j|X_A = I),$$

which means *homogeneity* across rows. Considering it is a regression model with X_B as response and X_A as explanatory variables, it means that X_A has no effect upon X_B. The interpretation of the models is summarized in the following.

- Poisson distribution: No interaction effect of X_A and X_B on counts.
- Multinomial distribution: X_A and X_B are independent.
- Product-multinomial distribution: Response X_B does not depend on X_A (fixed row marginals), response X_A does not depend on X_B (fixed column marginals).

Tests for the nullhypothesis $H_0 : \mu_{AB(ij)} = 0$ *for all* i, j have different interpretation. If H_0 is not rejected, that means for Poisson distribution tables, that the interaction term is not significant. For multinomial distribution tables the test is equivalent to testing the independence between X_A and X_B. If X_A and X_B are random variables and data have been collected as Poisson counts, by conditioning on X_A and X_B the interpretation as a test for independence also holds for Poisson tables (by conditioning on n). Of course, in applications where X_A and X_B refer to experimental conditions that interpretation is useless. Consider Example **??** in Chapter **??** where the counts of cases of encephalitis are modeled depending on country and time. These explanatory variables are experimental conditions rather than random variables and it is futile to try to investigate the independence of these conditions.

Parameters and Odds Ratio

The parameters of the saturated model (1.1) with symmetric constraints are easily computed as

$$\lambda = \frac{1}{IJ}\sum_{i,j}\log(\mu_{ij}), \quad \lambda_{A(i)} = \frac{1}{J}\sum_{j}\log(\mu_{ij}) - \lambda,$$
$$\lambda_{B(j)} = \frac{1}{I}\sum_{i}\log(\mu_{ij}) - \lambda, \quad \lambda_{AB(ij)} = \log(\mu_{ij}) - \lambda - \lambda_{A(i)} - \lambda_{B(j)}.$$

The parameters $\lambda_{A(i)}, \lambda_{B(j)}$ are the *main effects*, $\lambda_{AB(ij)}$ is a *two-factor interaction*.

For multinomial and product-multinomial distribution an independent measure of association which is strongly linked to two-factor interactions is the odds ratio. For the simple (2×2)-contingency table the odds ratio has the form

$$\gamma = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{P(X_A = 1, X_B = 1)/P(X_A = 1, X_B = 2)}{P(X_A = 2, X_B = 1)/P(X_A = 2, X_B = 2)}$$
$$= \frac{P(X_B = 1|X_A = 1)/P(X_B = 2|X_A = 1)}{P(X_B = 1|X_A = 2)/P(X_B = 2|X_A = 2)}$$

By using $\mu_{ij} = n\pi_{ij}$ (multinomial distribution) or $\mu_{ij} = n_{i+}\pi_{ij}$ (product-multinomial, fixed rows) one obtains for the log-linear model with symmetrical constraints

$$\log(\gamma) = 4\lambda_{AB(11)},$$

and for the model with the last category set to zero $\log(\gamma) = \lambda_{AB(11)}$. Thus γ is a direct function of the two-factor interaction. The connection to independence is immediately seen: $\lambda_{AB(11)} = 0$ is equivalent to $\gamma = 1$ which means independence of variables X_A and X_B (multinomial distribution) or homogeneity (product-multinomial distribution).

In the general case of $(I \times J)$-contingency tables one considers the odds ratio formed by the (2×2)-subtable built from rows $\{i_1, i_2\}$ and columns $\{j_1, j_2\}$ with cells $\{(i_1, j_1), (i_1, j_2), (i_2, j_1), (i_2, j_2)\}$. The corresponding odds ratio

$$\gamma_{(i_1 i_2)(j_1 j_2)} = \frac{\pi_{i_1 j_1}/\pi_{i_1 j_2}}{\pi_{i_2 j_1}/\pi_{i_2 j_2}}$$

may be expressed in two-factor interactions by

$$\log(\gamma_{(i_1 i_2)(j_1 j_2)}) = \lambda_{AB(i_1 j_1)} + \lambda_{AB(i_2, j_2)} - \lambda_{AB(i_2 j_1)} - \lambda_{AB(i_1 j_2)}.$$

1.3 Log-linear Models for Three-Way Tables

Three-way tables are characterized by three categorical variables, $X_A \in \{1, \ldots, I\}$, $X_B \in \{1, \ldots, J\}$ and $X_C \in \{1, \ldots, K\}$ which refer to rows, columns and layers of the table. Let $\{X_{ijk}\}$ denote the collection of cell counts where

$$\begin{aligned} X_{ijk} \quad & \text{denotes the counts in cell} \quad (i, j, k), \\ & \text{i.e. the number of observations with} \quad X_A = i, X_B = j, X_C = k. \end{aligned}$$

The general form of three-way tables is given in Table 1.5. Throughout the section the convention is used that the subscript " $+$ " denotes the sum over that index, for example $X_{ij+} = \sum_k X_{ijk}$. The types of contingency tables are in principle the same as for two-way tables. However, now there are more variants of conditioning.

Type 1: Poisson distribution

It is assumed that the X_{ijk} are independent Poisson-distributed random variables, $X_{ijk} \sim P(\lambda_{ijk})$. The total number of counts $n = \sum_{ijk} X_{ijk}$ as well as marginal counts are random variables. The natural model considers the counts as response and X_A, X_B and X_C as explanatory variables, which might refer to experimental conditions or random variables.

Type 2: Multinomial distribution

For a fixed number of subjects one observes the tupel (X_A, X_B, X_C) with possible outcomes $\{(1, 1, 1), \ldots, (I, J, K)\}$. The count in cell (i, j, k) is the number of

		X_C				
X_A	X_B	1	2	$\cdots$	K	
1	1	X_{111}	X_{112}	$\cdots$	X_{11K}	X_{11+}
	2	X_{121}	X_{122}			$\vdots$
	$\vdots$	$\vdots$				
	J	X_{1J1}	$\cdots$		X_{1JK}	X_{1J+}
2	1	X_{211}	X_{212}	$\cdots$	X_{21K}	X_{21+}
	2	X_{221}	X_{222}			$\vdots$
	$\vdots$	$\vdots$				
	J	X_{2J1}	$\cdots$		X_{2JK}	X_{2J+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
I	1	X_{I11}	X_{I12}	$\cdots$	X_{I1K}	X_{I1+}
	2	X_{I21}	X_{I22}			$\vdots$
	$\vdots$	$\vdots$				
	J	X_{IJ1}	$\cdots$		X_{IJK}	X_{IJ+}

TABLE 1.5: General form of three-way tables

observations with $X_A = i, X_B = j, X_C = k$. The counts $\{X_{ijk}\}$ follow a multinomial distribution $M(n, \{\pi_{ijk}\})$ where $\pi_{ijk} = P(X_A = i, X_B = j, X_C = k)$. For three and higher dimensional tables the notation $\{X_{ijk}\}$ and $\{\pi_{ijk}\}$ is preferred over the representation as vectors.

Type 3: Product-multinomial distribution

There are several variants of the product-multinomial distribution. Either one of the variables (X_A or X_B or X_C) is a design variable, meaning that corresponding marginals are fixed or two of them are design variables, meaning that two-dimensional margins are fixed. Let us consider as an example of the first variant the table that results from design variable X_A. That means $n_{i++} = X_{i++}$ is fixed and

$$(X_{i11}, \ldots, X_{iJK}) \sim M(n_{i++}, (\pi_{i11}, \ldots, \pi_{iJK})), \tag{1.4}$$

where $\pi_{ijk} = P(X_B = j, X_C = k | X_A = i)$. An example of the second variant (two design variables) is obtained by letting X_A and X_B be design variables, i. e. $n_{ij+} = X_{ij+}$ is fixed and

$$(X_{ij1}, \ldots, X_{ijK}) \sim M(n_{ij+}, (\pi_{ij1}, \ldots, \pi_{ijK})), \tag{1.5}$$

where $\pi_{ijk} = P(X_C = k | X_A = i, X_B = j)$. Hence, only X_C is a response variable, the number of observations for $(X_A, X_B) = (i, j)$ are given.

It should be noted that there is again a hierarchy among distributions. If $\{X_{ijk}\}$ have Poisson distribution, conditioning on $n = \Sigma_{ijk} X_{ijk}$ yields the multinomial distribution $\{X_{ijk}\} \sim M(n, \{\pi_{ijk}\})$ where $\pi_{ijk} = \lambda_{ijk}/\Sigma_{ijk}\lambda_{ijk}$. Further conditioning on $n_{i++} = \Sigma_{jk} X_{ijk}$ yields the product of the multinomial distributions (1.4). If in addition one conditions on $n_{ij+} = \Sigma_k X_{ijk}$ one obtains the product of distributions (1.5).

Log-linear Model for Three-Way Tables

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{AB(ij)} + \lambda_{AC(ik)} + \lambda_{BC(jk)} + \lambda_{ABC(ijk)}.$$

Constraints:

$$\sum_i \lambda_{A(i)} = \sum_j \lambda_{B(j)} = \sum_k \lambda_{C(k)} = 0,$$
$$\sum_i \lambda_{AB(ij)} = \sum_j \lambda_{AB(ij)} = \sum_i \lambda_{AC(ik)} = \sum_k \lambda_{AC(ik)} = \sum_j \lambda_{BC(jk)} = \sum_k \lambda_{BC(jk)} = 0,$$
$$\sum_i \lambda_{ABC(ijk)} = \sum_j \lambda_{ABC(ijk)} = \sum_k \lambda_{ABC(ijk)} = 0,$$

or

$$\lambda_{A(I)} = \lambda_{B(J)} = \lambda_{C(K)} = 0,$$
$$\lambda_{AB(i,J)} = \lambda_{AB(I,j)} = \lambda_{AC(iK)} = \lambda_{AC(Ik)} = \lambda_{BC(jK)} = \lambda_{BC(Jk)} = 0,$$
$$\lambda_{ABC(Ijk)} = \lambda_{ABC(iJk)} = \lambda_{ABC(ijK)} = 0.$$

TABLE 1.6: Log-linear model for three-way tables with constraints

Let in general $\mu_{ijk} = \mathrm{E}(X_{ijk})$ denote the mean of cell counts. Then the general form of the three-dimensional log-linear model is

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{AB(ij)} + \lambda_{AC(ik)} + \lambda_{BC(jk)} + \lambda_{ABC(ijk)}.$$

For necessary constraints see Table 1.6 where two sets of constraints are given, the set of symmetric constraints corresponding to ANOVA models and the set of constraints based on reference categories. For the multinomial and the product-

multinomial model additional constraints are needed, which are easily derived from the restrictions $n = \sum_{ijk} X_{ijk}$, etc. The model has three types of parameters, the three-factor interactions $\lambda_{ABC(ijk)}$, the two-factor interactions $\lambda_{AB(ij)}$, $\lambda_{AC(ik)}$ and $\lambda_{BC(jk)}$ and the main effects $\lambda_{A(i)}$, $\lambda_{B(j)}$, $\lambda_{C(k)}$. The general model is saturated, that means it has as many parameters as means μ_{ijk} and consequently every data set (without empty cells) yields perfect fit by setting $\hat{\mu}_{ijk} = X_{ijk}$.

More interesting models are derived from the general model by omitting groups of parameters corresponding to interaction terms. The attractive feature of log-linear models is that most of the resulting models have an interpretation in terms of independence or conditional independence. In general, categorical variables X_A, X_B, X_C are *independent*, if

$$P(X_A = i, X_B = j, X_C = k) = P(X_A = i)P(X_B = j)P(X_C = k)$$

holds for all i, j, k. *Conditional independence* of X_A and X_B given X_C (in short $X_A \perp X_B | X_C$) holds if for all i, j, k

$$P(X_A = i, X_B = j | X_C = k) = P(X_A = i | X_C = k)P(X_B = j | X_C = k).$$

Hierarchical Models

The interesting class of models that may be interpreted in terms of (conditional) independence are the *hierarchical models*. A model is called hierarchical if the model includes all lower-order terms composed from variables contained in a higher-order term. For example, if a model contains $\lambda_{BC(jk)}$ it also contains the marginals $\lambda_{B(j)}$ and $\lambda_{C(k)}$. Hierarchical models may be abbreviated by giving the terms of highest order. For example the symbol AB/AC denotes the model containing $\lambda_{AB}, \lambda_{AC}, \lambda_A, \lambda_B, \lambda_C$ (and a constant term). Further examples are given in Table 1.7. The notation is very similar to the Wilkinson-Rogers notation (see...) which for the model AB/AC is $A * B + A * C$. The latter form is itself shorthand for the extended Wilkinson-Rogers notation $A.B + A.C + A + B + C$.

Graphical Models

Most of the hierarchical log-linear models for three-way tables are also *graphical models* which are considered in more detail in Section 1.5. The basic concept is only sketched here. If a graph is drawn by linking variables for which the two-factor interaction is contained in the model one obtains a simple graph. If in the resulting graph there is no connection between groups of variables, these groups of variables are independent. If two variables are connected only by edges through the third variable, the two variables are conditionally independent given the third variables. For examples, see Table 1.7, for a more concise definition of graphical models see Section 1.5.

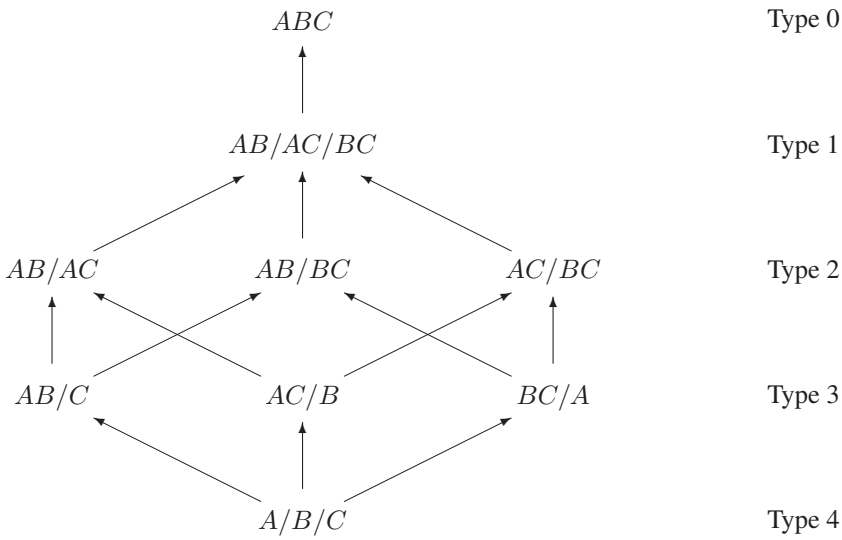


FIGURE 1.1: Hierarchy of three-dimensional log-linear models

1.4 Specific Log-Linear Models

In the following the types of hierarchical models for three-way tables are considered under the assumption that X_A, X_B, X_C represent random variables (i.e. multinomial contingency tables).

Type 0: Saturated Model

The saturated model is given by

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{AB(i,j)} + \lambda_{AC(ik)} + \lambda_{BC(jk)} + \lambda_{ABC(ijk)}.$$

It represents a reparameterization of the means $\{\mu_{ijk}\}$ without implying any additional structure (except $\mu_{ijk} > 0$).

Type 1: No three-factor interaction

The model

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{AB(ij)} + \lambda_{AC(ik)} + \lambda_{BC(jk)}$$

contains only two-factor interactions and is denoted by $AB/AC/BC$. Since the three-factor interaction is omitted the model has to imply restrictions on the underlying probabilities. In order to see what the model implies it is useful to look at the conditional association of two variables given a specific level of the third variable.

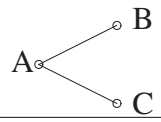
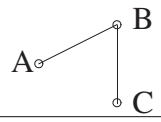
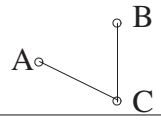
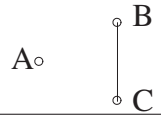
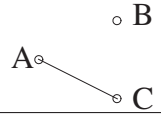
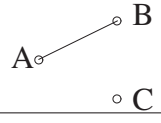
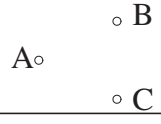
	Log-linear Model	Regressors of Logit-Model (with response X_C)
AB/AC X_B, X_C conditionally independent, given X_A		$1, x_A$
AB/BC X_A, X_C conditionally independent, given X_B		$1, x_B$
AC/BC X_A, X_B conditionally independent, given X_C		$1, x_A, x_B$
A/BC X_A independent of (X_B, X_C)		$1, x_B$
AC/B (X_A, X_C) independent of X_B		$1, x_A$
AB/C (X_A, X_B) independent of X_C		1
$A/B/C$ X_A, X_B, X_C are dependent		1

TABLE 1.7: Graphical models for three-way tables

Let us consider the odds ratios of X_A and X_B given $X_C = k$ and for simplicity assume that all variables are binary. Then the conditional association measured by the odds ratio has the form

$$\gamma(X_A, X_B|X_C = k) = \frac{P(X_A = 1, X_B = 1|X_C = k)/P(X_A = 2, X_B = 1|X_C = k)}{P(X_A = 1, X_B = 2|X_C = k)/P(X_A = 2, X_B = 2|X_C = k)}$$

and is built from the (2×2)-table formed by X_A and X_B for fixed level $X_C = k$.

By using $\mu_{ijk} = n\pi_{ijk}$ and

$$\gamma(X_A, X_B|X_C = k) = \frac{\pi_{11k}/\pi_{21k}}{\pi_{12k}/\pi_{22k}} = \frac{\mu_{11k}/\mu_{21k}}{\mu_{12k}/\mu_{22k}}$$

one obtains for the model without three-factor interactions that all terms depending on k cancel out and therefore $\gamma(X_A, X_B|X_C = k)$ *does not depend on* k. That means that the conditional association between X_A and X_B given $X_C = k$ does not depend on the level k. Whatever the conditional association between these two variables is, strong or weak or not present, it is the same for all levels of X_C, thus X_C does not modify the association between X_A and X_B. The same holds if X_A and X_B have more than two categories where conditional association is measured by odds ratios of (2×2)- subtables built form the total table. Moreover, since the model is symmetric in the variables it is also implied that the conditional association between X_A and X_C given $X_B = j$ does not depend on j and the conditional association between X_B and X_C given $X_A = i$ does not depend on i.

It should be noted that the model without the three-factor interaction does *not* imply that two variables are independent of the third variable. There might be a strong dependence between $\{X_A, X_B\}$ and X_C although the conditional association of X_A and X_B given $X_C = k$ does not depend on the level of X_C. The model is somewhat special since it is the only log-linear model for three-way tables that is not a graphical model and therefore cannot be represented by a simple graph. It is also the only model that can not be interpreted in terms of independence or conditional independence of variables.

Type 2: Only two two-factor interactions contained

A model of this type is the model AC/BC given by

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{AC(ik)} + \lambda_{BC(jk)}.$$

If the model holds the variables X_A and X_B are *conditionally independent*, given X_C, or more formally

$$P(X_A = i, X_B = j|X_C = k) = P(X_A = i|X_C = k)P(X_B = j|X_C = k).$$

This may be easily derived by using that the model is equivalent to postulating that $\mu_{ijk} = \mu_{i+k}\mu_{+jk}/\mu_{++k}$. It means that conditionally the variables X_A and X_B are not associated. However, that does not mean that there is no marginal association between X_A and X_B. X_A and X_B may be strongly associated when X_C is ignored. The model is a graphical model with the graph given in Table 1.7. The graph contains edges between X_A and X_C as well as between X_B and X_C but not between X_A and X_B. It illustrates that X_A and X_B have some connection through the common variable X_C. And that is exactly the meaning of the graph: given X_C the variables X_A and X_B are independent since the connection between X_A and X_B is only through X_C.

The other two models of this type are AB/AC and AB/BC (shown in Table 1.7). The first postulates that X_B and X_C are conditionally independent given X_A and the latter postulates that X_A and X_C are conditionally independent given X_B.

Type 3: Only one two-factor interaction contained

A model of this type is the model A/BC given by

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{BC(jk)},$$

which contains only main effects and one two-factor interaction. By simple derivation one obtains that the model postulates that X_A is *jointly independent* of X_B and X_C. That means the groups of variables $\{A\}$ and $\{B, C\}$ are independent, or more formally

$$P(X_A = i, X_B = j, X_C = k) = P(X_A = i)P(X_B = j, X_C = k).$$

The model implies stronger restrictions on the underlying probability structure than the model AC/BC since now in addition the two-factor interaction λ_{AC} is omitted. The corresponding graph in Table 1.7 is very suggestive. There is no edge between the variable X_A and the two variables X_B, X_C, the two groups of variables are well separated, corresponding to the interpretation of the model that X_A and X_B, X_C are independent.

Type 4: Main effects model

The model has the form

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)}.$$

The model represents independence of variables X_A, X_B, X_C

$$P(X_A = i, X_B = j, X_C = k) = P(X_A = i)P(X_B = j)P(X_C = k),$$

meaning in particular that all the variables are mutually independent.

Figure 1.1 shows the hierarchy of log-linear models. It is obvious that the model AB/BC is a submodel of $AB/BC/AC$ since the latter is less restrictive than the former. But not any two models are nested. For example there is no hierarchy between the models AB/AC and AB/BC. The possible models form a lattice with semi-ordering.

Table 1.8 shows what restrictions are implied by omitting interaction terms. For example, the model AB/AC implies that $\mu_{ijk} = \mu_{ij+}\mu_{i+k}/\mu_{i++}$ holds. When the sampling is multinomial, it is easily derived what that means for conditional probabilities and therefore for the interpretation of the assumed association structure. The corresponding graphs are given in Table 1.7.

Models for Product-Multinomial Contingency Tables

While all models in Figure 1.1 apply for multinomial contingency tables, not all models may be built for product-multinomial contingency tables. The cause is that marginal sums which are fixed by design have to be fitted by the model. In general, if margins are fixed by design, the corresponding interaction term has to be contained

AB/AC	$\lambda_{ABC} = \lambda_{BC} = 0$	$\mu_{ijk} = \frac{\mu_{ij+}\mu_{i+k}}{\mu_{i++}}$ $P(X_B, X_C\|X_A) = P(X_B\|X_A)P(X_C\|X_A)$
AB/BC	$\lambda_{ABC} = \lambda_{AC} = 0$	$\mu_{ijk} = \frac{\mu_{+jk}\mu_{ij+}}{\mu_{+j+}}$ $P(X_A, X_C\|X_B) = P(X_A\|X_B)P(X_C\|X_B)$
AC/BC	$\lambda_{ABC} = \lambda_{AB} = 0$	$\mu_{ijk} = \frac{\mu_{i+k}\mu_{+jk}}{\mu_{++k}}$ $P(X_A, X_B\|X_C) = P(X_A\|X_C)P(X_B\|X_C)$
A/BC	$\lambda_{ABC} = \lambda_{AB} = \lambda_{AC} = 0$	$\mu_{ijk} = \frac{\mu_{i++}\mu_{+jk}}{\mu_{+++}}$ $P(X_A, X_B, X_C) = P(X_A)P(X_B, X_C)$
AC/B	$\lambda_{ABC} = \lambda_{AB} = \lambda_{BC} = 0$	$\mu_{ijk} = \frac{\mu_{i+k}\mu_{+j+}}{\mu_{+++}}$ $P(X_A, X_B, X_C) = P(X_A, X_C)P(X_B)$
AB/C	$\lambda_{ABC} = \lambda_{AC} = \lambda_{BC} = 0$	$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}}$ $P(X_A, X_B, X_C) = P(X_A, X_B)P(X_C)$
A/B/C	$\lambda_{ABC} = \lambda_{AB} = \lambda_{AC} = 0$ $\lambda_{BC} = 0$	$\mu_{ijk} = \frac{\mu_{i++}\mu_{+j+}\mu_{++k}}{\mu^2_{+++}}$ $P(X_A, X_B, X_C) = P(X_A)P(X_B)P(X_C)$

TABLE 1.8: Graphical models and interpretation for three-way-tables

in the model. For example, if the two dimensional margins $X_{ij+} = \sum_k X_{ijk}$ are fixed by design, the model has to contain the interaction λ_{AB}. The model AC/BC is not a valid model since it fits the margins X_{i+k} and X_{+jk} but does not contain λ_{AB} (see also Lang, 1996a , Bishop, Fienberg, and Holland, 1975 , Agresti, 2002).

1.5 Log-Linear and Graphical Models for Higher Dimensions

Log-linear models for higher dimensions than three have basically the same structure, but the number of possible interaction terms and the number of possible models increases. For example in four-way tables a four-factor interaction term can be contained. It is helpful that for hierarchical models the same notation applies as in lower dimensional models. An example of a four-way model is ABC/AD which is given by

$$\begin{aligned}\log(\mu_{ijkl}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} + \lambda_{D(l)} \\ + \lambda_{AB(ij)} + \lambda_{AC(ik)} + \lambda_{BC(jk)} + \lambda_{AD(il)} + \lambda_{ABC(ijk)}.\end{aligned}$$

The model contains only one three-factor interaction and only four two-factor interaction but all main effects. For the interpretation of higher dimensional tables the representation as graphical models is a helpful tool.

FIGURE 1.2: Graphs for log-linear model AB/AC (left) and ABC (right)

Graphical Models

In order to obtain models which have simple interpretation in terms of conditional interpretation it is useful to restrict consideration to subclasses of log-linear models. We already made the restriction to hierarchical models. A log-linear model is hierarchical if the model includes all lower-order terms composed from variables contained in a higher order term. A further restriction is that to graphical models.

> A log-linear model is *graphical* if whenever the model contains all two-factor interactions generated by a higher-order interaction, the model also contains the higher-order interaction.

In three-way tables there is only one log-linear model that is not graphical, namely the model AB/AC/BC. That model contains all two-factors interactions λ_{AB}, λ_{AC}, λ_{BC} which are generated as marginal parameters of the three-factor interaction λ_{ABC}, but λ_{ABC} itself is not contained in the model.

A graphical model has a graphical representation which makes it easy to see what types of conditional independence structure is implied. The representation is based on mathematical graph theory, outlined for example in Whittaker (1990). and Lauritzen (1996) In general a *graph* consists of two sets, the sets of *vertices*, K, and the set of *edges*, E. The set of edges consists of pairs of elements from $K, E \subset K \times K$. In graphical log-linear models the vertices correspond to variables and edges correspond to pairs of variables. Therefore, we will set K to $K = \{A, B, C, \dots\}$ and an element from E has the form (A, C). In undirected graphs, the type of graph that is considered here, if (A, B) is in E, also (B, A) is in E and the edge or line between A and B is undirected. A *chain* between vertices A and C is determined by a sequence of distinct vertices $V_1, \dots, V_m, V_i \in K$. The chain is given by the sequence of edges $[AV_1/V_1V_2/\dots/V_mC]$ for which (V_i, V_{i+1}), as well as $(AV_1), (V_mC)$ are in E. That means a chain represents a sequence of variables leading from one variable to another within the graph. Although A and C may be identical it is not allowed that a vertex between A and C is included more than once. Therefore circles are avoided. The left graph in Fig.1.2 contains for example the chains $[BA/AC]$, $[CA/AB]$, $[AB]$, the right graph contains the chain $[AB/BC/CA]$.

Chains are important for the interpretation of the model. The left graph in Fig. 1.2 corresponds to the model AB/AC, which implies that X_B and X_C are condi-

tionally independent given A. If one looks at paths that connect B and C it is seen that any paths connecting B and C involves A. This property of the graph may be read as conditional independence of X_B and X_C given X_A.

For the correspondence of graphical log-linear models and graphs it is helpful to consider the largest sets of vertices that include all possible edges between them. A set of vertices for which all the vertices are connected by edges is called *complete*. The corresponding vertices form a complete *subgraph*. A complete set that is not contained in any other complete set is called a *maximal complete set* or a *clique*. The cliques determine the graphical linear model and correspond directly to the notation defining the model. For example the model AB/AC has the cliques $\{AB\}$ and $\{AC\}$. The saturated model ABC which contains all possible edges has the maximal complete set or clique $\{ABC\}$. An example of a higher dimensional model is the model ABC/AD, which is a graphical model. The model has the cliques $\{ABC\}$, $\{AD\}$ (see Fig. 1.3 for the graph). Fig 1.3 also shows the graphs for the saturated model $ABCD$ and the model $ABC/ABD/DE$.

The strength of graphing log-linear models becomes obvious in higher dimensional tables. The basic tool for the interpretation of graphical log-linear models is a result by Darroch, Lauritzen, and Speed (1980) :

> Let the sets F_0, F_1, F_2 denote disjoint subsets of the variables in a graphical log-linear model. The factors in F_1 are conditionally independent of the factors in F_2 given F_0 if and only if every chain between a factor in F_1 and a factor in F_2 involves at least one factor in F_0. Then F_0 is said to separate the subgraphs formed by F_1 and F_2.

For the model AB/AC (see graph in Fig. 1.2) one may consider $F_1 = \{B\}, F_2 = \{C\}$ and $F_0 = \{A\}$. The conditional independence of X_B and X_C given X_A, in short $X_B \perp X_C|X_A$, follows directly from the result of Darroch, Lauritzen, and Speed (1980) . For the model $ABC/ABD/DE$ (see graph in Fig. 1.3) one may build several subsets of variables. By considering $F_1 = \{A, B, C\}, F_2 = \{E\}$ and $F_0 = \{D\}$ one obtains that $\{X_A, X_B, X_C\}$ are conditionally independent of X_E given X_D, $\{X_A, X_B, X_C\} \perp X_E|X_D$. It is said that X_D separates the subgraphs formed by $\{X_A, X_B, X_C\}$ and $\{X_E\}$. By considering $F_1 = \{B, C\}, F_2 = \{E\}, F_0 = \{A, D\}$ one obtains that $\{X_B, X_C\}$ are conditionally independent of X_E given $\{X_A, X_D\}$. It is seen that for higher dimensional models usually several independence structures are involved when considering a graphical log-linear model.

Marginal independence occurs if there are no chains in the graph that connect two groups of variables. The graph corresponding to model AB/C (see graph in Table 1.7) contains no chain between $\{A, B\}$ and $\{C\}$. The implication is that the variables $\{X_A, X_B\}$ are independent of X_C, in short $\{X_A, X_B\} \perp X_C$. A model may also imply certain marginal independence structures. For example the model $AB/BC/CD$ implies conditional independence relations $X_A \perp X_D|\{X_B, X_C\}$ and $X_A \perp \{X_C, X_D\}|X_B$, which include all four variables, but also $X_A \perp X_C|X_B$, which concerns the marginal distribution of X_A, X_B, X_C.

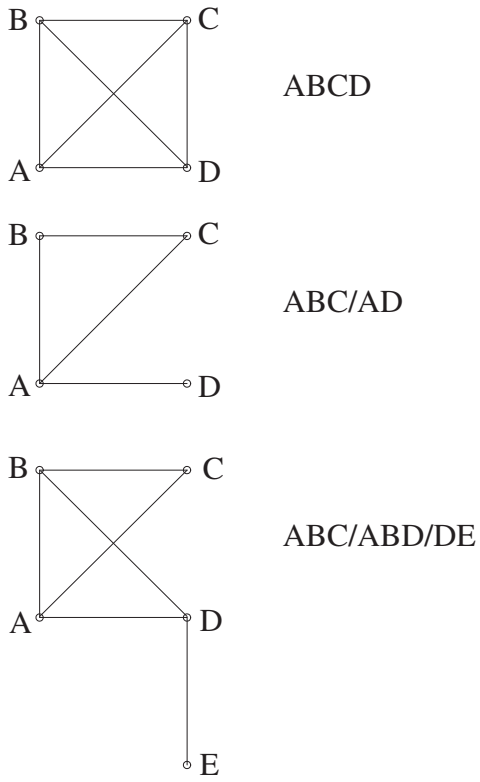


FIGURE 1.3: Graphs for log-linear models in multi-way tables

As was already seen in three-way tables not all log-linear models are graphical. That raises the question how to interpret a log-linear model that is not graphical. Fortunately, any log-linear model can be embedded in a graphical model. For the interpretation one uses the smallest graphical model that contains the specific model. Since the specific model is a submodel of that graphical model, all the (conditional) independence structure of the larger model also has to hold for the specific model. That strategy does not always work satisfactorily. The smallest graphical model that contains the three-way table model $AB/AC/BC$ is the saturated model which implies no independence structure. But the model $AB/AC/BC$ also has no simple interpretation in terms of conditional independence, although exclusion of the three-factor interaction restricts the association structure between variables.

1.6 Collapsibility

In general association in marginal tables differs from association structures found in the full table. For example, X_A and X_B can be conditionally independent given

$X_C = k$, even if variables X_A and X_B are marginally dependent. Marginal dependence means that the association is considered in the marginal table obtained from collapsing over the categories of the other variables, i.e. the other variables are ignored (see also Exercise 1.5).

The question arises under which conditions is it possible to infer on the association structure from marginal tables. Let us consider a three-way table with means μ_{ijk}. The marginal association between binary factors X_A and X_B, measured in odds ratios, is determined by

$$\frac{\mu_{11+}/\mu_{12+}}{\mu_{21+}/\mu_{22+}},$$

while the conditional association between X_A and X_B given $X_C = k$ is determined by

$$\frac{\mu_{11k}/\mu_{12k}}{\mu_{21k}/\mu_{22k}}.$$

One can show that the association is the same if model AB/AC holds (compare Exercise 1.12).

In general the association is unchanged if groups of variables are separated:

> Let the sets F_1, F_2, F_0 denote disjoint subsets of the variables in a graphical log-linear model. If every chain between a factor in F_1 and a factor in F_2 involves at least one factor in F_0 the association among the factors in F_1 and F_0 can be examined in the marginal table obtained from collapsing over the factors in F_2. In the same way the association among the factors in F_2 and F_0 can be examined in the marginal table obtained from collapsing over the factors in F_1

(compare Darroch, Lauritzen, and Speed (1980) and Bishop, Fienberg, and Holland (2007)) . Therefore, if F_0 separates the subgraphs formed by F_1 and F_2 one can collapse over F_2 (or F_1, respectively). In the model AB/AC, considered previously, the association between X_A and X_B as well as the association between X_A and X_C can be examined from the corresponding marginal tables.

1.7 Log-Linear Models and the Logit Model

The log-linear models for contingency tables may be represented as logit models. Let us consider a three-way table with categorical variables X_A, X_B, X_C. The most general log-linear model is the saturated model

$$\begin{aligned}\log(\mu_{ijk}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{C(k)} \\ + \lambda_{AB(ij)} + \lambda_{AC(ik)} + \lambda_{BC(jk)} + \lambda_{ABC(ijk)}.\end{aligned}$$

For multinomial tables, for which $\mu_{ijk} = n\pi_{ijk}$, one obtains the logit model with reference category $(X_A = I, X_B = J, X_C = K)$

$$\log\left(\frac{\pi_{ijk}}{\mu_{IJK}}\right) = \gamma_{A(i)} + \gamma_{B(j)} + \gamma_{C(k)} + \gamma_{AB(ij)} + \gamma_{AC(ik)} + \gamma_{BC(jk)} + \gamma_{ABC(ijk)},$$

where the γ-parameters are obtained as differences, for example $\gamma_{A(i)} = \lambda_{A(i)} - \lambda_{A(I)}$, $\gamma_{ABC(ijk)} = \lambda_{ABC(ijk)} - \lambda_{ABC(IJK)}$. The parametrization of the model, which uses reference categories for the variables X_A, X_B, X_C, reflects that a structured multinomial distribution is given. In contrast to the simple multinomial distributions considered in Chapter **??**, the distribution of the response is determined by three separate variables that structure the multinomial distribution.

Logit Models with Selected Response Variables

Consider now that X_C is chosen as response variable. Then one obtains for multinomial tables

$$\log\left(\frac{\mu_{ijr}}{\mu_{ijK}}\right) = \log\left(\frac{P(X_A = i, X_B = j, X_C = k)}{P(X_A = i, X_B = j, X_C = K)}\right) = \log\left(\frac{\pi_{r|ij}}{\pi_{K|ij}}\right),$$

where $\pi_{r|ij} = P(X_C = r|X_A = i, X_B = j)$. By using the saturated model, which always holds, one obtains from easy derivation the multinomial logit model

$$\log\left(\frac{P(X_C = r|X_A = i, X_B = j)}{P(X_C = K|X_A = i, X_B = j)}\right) = \gamma_{0r} + \gamma_{A(i),r} + \gamma_{B(j),r} + \gamma_{AB(ij),r},$$

where

$$\gamma_{0r} = \lambda_{C(r)} - \lambda_{C(K)}, \quad \gamma_{A(i),r} = \lambda_{AC(ir)} - \lambda_{AC(iK)},$$
$$\gamma_{B(j),r} = \lambda_{BC(jr)} - \lambda_{BC(jK)}, \quad \gamma_{AB(ij),r} = \lambda_{ABC(ijr)} - \lambda_{ABC(ijK)}.$$

An alternative form of the model, which uses dummy variables is

$$\log\left(\frac{\pi_{r|ij})}{\pi_{K|ij})}\right) = \gamma_{0r} + \gamma_{A(1),r}x_{A(1)} + \cdots + \gamma_{B(1),r}x_{B(1)} + \quad \cdots +$$
$$\gamma_{AB(11),r}x_{A(1)}x_{B(1)} + \cdots + \gamma_{AB(I-1,J-1),r}x_{A(I-1)}x_{B(J-1)}.$$

The constraints on the λ parameters and therefore the type of coding of dummy variables carry over to the γ parameters. For example, the constraint $\sum_i \lambda_{AC(ik)} = 0$ transforms into $\sum_i \gamma_{A(i),r} = 0$.

In summary, by choosing one variable as response variable the log-linear model of association between X_A, X_B, X_C turns into a regression model. If the log-linear model is a submodel of the saturated model some γ terms are not contained in the corresponding logit model. For example, by assuming the log-linear model AB/AC (meaning that X_B and X_C are conditionally independent) one obtains the logit model

$$\log\left(\frac{P(X_C = r|X_{A=i}, X_B = j)}{P(X_C = K|X_A = i, X_B = j)}\right) = \gamma_{0r} + \gamma_{A(i),r},$$

which contains only the explanatory variable X_A. Since X_B and X_C are conditionally independent given X_A it is quite natural that X_B does not effect upon X_C since

it is associated with X_C only through X_A. In Table 1.7 the explanatory variables of logit models with response X_C are given together with the underlying log-linear model. It is seen that model AB/BC as well as model A/BC yield a logit model with X_B as the only explanatory variable. Model AB/BC is weaker than A/BC. Since the effect of variable X_A on X_C is already omitted if model AB/BC holds, it is naturally omitted if an even stronger model holds.

1.8 Inference for Log-linear Models

Log-linear models may be embedded into the framework of generalized linear models. For all three sampling schemes, Poisson distribution, multinomial distribution and product-multinomial distribution, the response distribution is in the exponential family. The log-linear model has the form assumed in GLMs where the mean is linked to the linear predictor by a transformation function. Thus maximum likelihood estimation and testing is based on the methods developed in Chapter **??** and Chapter **??**. An advantage of log-linear models is that maximum likelihood estimates are sometimes easier to compute and that sufficient statistics have a simple form. In the following the results are shortly stretched.

1.8.1 Maximum Likelihood Estimates and Minimal Sufficient Statistics

For simplicity the Poisson distribution is considered for three-way models. Let all the parameters be collected in one parameter vector $\boldsymbol{\lambda}$. From the likelihood function

$$L(\boldsymbol{\lambda}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\mu_{ijk}^{x_{ijk}}}{x_{ijk}!} e^{-\mu_{ijk}}$$

one obtains the log-likelihood

$$l(\boldsymbol{\lambda}) = \sum_{i,j,k} x_{ijk} \log(\mu_{ijk}) - \sum_{i,j,k} \mu_{ijk} - \sum_{i,j,k} \log(x_{ijk}!).$$

With μ_{ijk} parameterized as the saturated log-linear model one obtains by rearranging terms (and omitting constants)

$$\begin{aligned} l(\boldsymbol{\lambda}) &= n\lambda_0 + \sum_i x_{i++}\lambda_{A(i)} + \sum_j x_{+j+}\lambda_{B(j)} + \sum_k x_{++k}\lambda_{C(k)} \\ &+ \sum_{i,j} x_{ij+}\lambda_{AB(ij)} + \sum_{i,k} x_{i+k}\lambda_{AC(ik)} + \sum_{j,k} x_{+jk}\lambda_{BC(jk)} \\ &+ \sum_{i,j,k} x_{ijk}\lambda_{ABC(ijk)} - \sum_{i,j,k} \exp(\lambda_0 + \lambda_{A(i)} + ... + \lambda_{ABC(ijk)}). \end{aligned}$$

The form of the log-likelihood remains the same when non-saturated models are considered. For example, if $\lambda_{ABC} = 0$ the term $\sum_{i,j,k} x_{ijk}\lambda_{ABC(ijk)}$ is omitted.

Since the Poisson is an exponential family distribution the factors on parameters represent sufficient statistics which contain all the information about parameters. That means that for non-saturated models the parameter estimates are determined by marginal sums. For example, the likelihood of the independence model $A/B/C$ contains only the marginal sums $x_{i++}, x_{+j+}, x_{++k}$. It is noteworthy that the sufficient statistics which are even minimal statistics correspond directly to the symbol for the model. Table 1.9 gives the sufficient statistics for the various types of log-linear models for three-way tables.

As usual maximum likelihood estimates are obtained by setting the derivations of the log-likelihood equal to zero. The derivative for one of the parameters, say $\lambda_{AB(ij)}$, is given by

$$\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda_{AB(ij)}} = x_{ij+} - \sum_k \exp(\lambda_0 + \lambda_{A(i)} + \dots) = x_{ij+} - \mu_{ij+}.$$

From $\partial l(\boldsymbol{\lambda})/\partial \lambda_{AB(ij)} = 0$ one obtains immediately $x_{ij+} = \hat{\mu}_{ij+}$. Hence, computation of maximum likelihood estimates reduces to solving the equations which equal the sufficient statistics to their expected values. For example for the independence model one has to solve the system of equations

$$x_{i++} = \hat{\mu}_{i++}, \quad x_{+j+} = \hat{\mu}_{+j+}, \quad x_{++k} = \hat{\mu}_{++k}, \tag{1.6}$$

$i = 1, \dots, I, j = 1, \dots J, k = 1, \dots, K$.

If the log-linear model is represented in the general vector form

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\lambda}$$

with $\boldsymbol{\mu}$ containing all the expected cell counts and $\boldsymbol{X}$ denoting the corresponding design matrix, the likelihood equations, equating sufficient statistics to expected values, have the form

$$\boldsymbol{X}^T\boldsymbol{x} = \boldsymbol{X}^T\boldsymbol{\mu},$$

where $\boldsymbol{x}$ is the vector of cell counts (for three way tables $\boldsymbol{x}^T = (x_{111}, x_{112}, \dots, x_{IJK})$).

ABC	$\{x_{ijk}\}$
AB/AC/BC	$\{x_{ij+}\}, \{x_{i+k}\}\{x_{+jk}\}$
AB/AC	$\{x_{ij+}\}, \{x_{i+k}\}$
A/BC	$\{x_{i++}\}, \{x_{+jk}\}$
A/B/C	$\{x_{i++}\}, \{x_{+j+}\}, \{x_{++k}\}$

TABLE 1.9: Log-linear models and sufficient statistics for three-way tables

Solving these equations can be very easy. For example, the solution of (1.6) is directly given by

$$\hat{\mu}_{ijk} = n\frac{x_{i++}}{n}\frac{x_{+j+}}{n}\frac{x_{++k}}{n}.$$

The form mimics $\mu_{ijk} = n\pi_{i++}\pi_{+j+}\pi_{++k}$. For all log-linear models for three-way tables except $AB/AC/BC$ direct estimates are available. A general class of

models for which direct estimates exist are decomposable models. A model is called decomposable if it is graphical and chordal, where chordal means that every closed chain $[AV_1/V_1V_2/\ldots/V_mA]$ (for which starting point end end point are identical) and which involves at least four distinct edges has a shortcut. A simple example is the model $AB/BC/CD/AD$ which is represented by a rectangle. It is not decomposable since the chain $[AB/BC/CD/DA]$ has no shortcut. It becomes decomposable by adding one more edge, AC or BD, yielding the model ABC/ACD or ABD/BCD, respectively. A more extensive treatment of direct estimates was given for example by Bishop, Fienberg, and Holland (1975) .

If no direct estimates are available, iterative procedures as in GLMs can be used. An alternative, rather stable procedure that is still used for log-linear models is *iterative proportional fitting*, also called *Deming -Stephan algorithm* (Deming and Stephan (1940)). It iteratively fits the marginals, which for the example of the independence model are given in 1.6. It works for direct estimates as well as for models, for which no direct estimates exist.

For graphical models the density can always be represented in the form

$$f(\{x_{ijk}\}) = \frac{1}{z_0}\prod_{C_l}\phi_{C_l}(x_{C_l}),$$

where the sum is over the cliques, z_0 is a normalizing constant, and $\phi_{C_l}(x_{C_l})$ are so-called clique potentials depending on observations x_{C_l} within the subgraph formed by C_l. The clique potentials have not to be density functions but contain the dependencies in C_l. Therefore, estimation is based on marginals that are determined by the cliques (for general algorithms based on the decomposition see for example Lauritzen (1996)).

For Poisson sampling the Fisher matrix $\boldsymbol{F}(\hat{\boldsymbol{\lambda}})$ has the simple form $\boldsymbol{X}^T diag(\boldsymbol{\mu})\boldsymbol{X}$ yielding the approximation

$$\mathrm{cov}(\hat{\boldsymbol{\lambda}}) \approx (\boldsymbol{X}^T \,\mathrm{diag}(\hat{\boldsymbol{\mu}})\boldsymbol{X})^{-1}.$$

For multionomial sampling one has to separate the intercept, which is fixed by the sample size. For the corresponding model $\log(\boldsymbol{\mu}) = \lambda_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\gamma}$ the Fisher matrix is $\boldsymbol{X}^T(diag(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{X}$ yielding the approximation

$$\mathrm{cov}(\hat{\boldsymbol{\gamma}}) \approx (\boldsymbol{X}^T(diag(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)\boldsymbol{X})^{-1}.$$

ML estimates for both sampling distributions can be computed within a closed framework. Let $\mu = \sum_i \mu_i = \sum_i \exp(\lambda_0 + \boldsymbol{x}_i^T\boldsymbol{\gamma})$ denote the total expected cell counts and

$$\pi_i = \frac{\mu_i}{\sum_j \mu_j} = \frac{\exp(\lambda_0 + \boldsymbol{x}_i^T\boldsymbol{\gamma})}{\sum_j \exp(\lambda_0 + \boldsymbol{x}_j^T\boldsymbol{\gamma})} = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\gamma})}{\sum_j \exp(\boldsymbol{x}_j^T\boldsymbol{\gamma})}$$

the relative expected cell counts, which do not depend on λ_0. With $x = \sum_i x_i$ representing the total count one obtains the log-likelihood for the Poisson distribution

$$l(\lambda_0, \boldsymbol{\gamma}) = \sum_i x_i \log(\mu_i) - \sum_i \mu_i = \sum_i x_i(\lambda_0 + \boldsymbol{x}_i^T\boldsymbol{\gamma}) - \mu = x\lambda_0 + \sum_i x_i(\boldsymbol{x}_i^T\boldsymbol{\gamma}) - \mu.$$

By including $x\log(\mu) - x\log(\mu)$ one obtains the additive decomposition

$$l(\lambda_0, \boldsymbol{\gamma}) = \{\sum_i x_i(\boldsymbol{x}_i^T\boldsymbol{\gamma}) - x\log(\sum_i \exp(\boldsymbol{x}_i^T\boldsymbol{\gamma}))\} + \{x\log(\mu) - \mu\}.$$

The first term is the log-likelihood of a multinomial distribution $(x_1, x_2, \dots) \sim M(x, (\pi_1, \pi_2, \dots))$, the term $x\log(\mu) - \mu$ is the log-likelihood of a Poison distribution $x \sim P(\mu)$. Therefore, maximization of the first term, which does not include the intercept, yields estimates $\hat{\boldsymbol{\gamma}}$ for multinomial sampling, conditional on the number of cell counts x. Maximization of the Poisson log-likelihood yields $\hat{\mu} = x$, which determines the estimate of λ_0, since $\mu = c\exp(\lambda_0)$, where $c = \sum_j \exp(\boldsymbol{x}_j^T\boldsymbol{\gamma})$ is just a scaling constant determined by maximization of the first term.

A similar decomposition of the Poisson log-likelihood holds for the product-multinomial distribution. Computation as well as inference can be based on the same likelihood with conditioning arguments. For details see Palmgren (1981) , Lang (1996b) .

1.8.2 Testing and Goodness-of-fit

Let the cell counts be given by the vector $\boldsymbol{x}^T = (x_1, \dots, x_N)$ where N is the number of cells and only a single index is used for denoting the cell. The vector $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_N)$ denotes the corresponding fitted means. For models with an intercept the deviance has the form

$$D = 2\sum_{i=1}^{N} x_i \log(\frac{x_i}{\hat{\mu}_i}). \tag{1.7}$$

When considering goodness-of-fit an alternative is Pearson's χ^2

$$\chi_P^2 = \sum_{i=1}^{N} \frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

For fixed N, both statistics have approximate χ^2-distribution if the assumed model holds and means μ_i are large. The degrees of freedom are $N - p$ where p is the number of estimated parameters.
The degrees of freedom are computed from the general rule

Number of parameters in the saturated model

– Number of parameters in the assumed model.

More concise, the number of parameters is the number of linearly independent parameters. For example, the restriction $\sum_i \lambda_{A(i)} = 0$ implies that the effective number of parameters $\lambda_{A(i)}, i = 1, \dots, I$, is $I-1$ since $\lambda_{A(I)} = -\lambda_{A(1)} - \cdots - \lambda_{A(I-1)}$.

Let us consider an example for three-way tables. The saturated model has IJK parameters (corresponding to the cells) for Poisson data, but $IJK-1$ parameters for

multinomial data, since the restriction $\sum_{ijk} \mu_{ijk} = 1$ applies. For the independence model the number of parameters is determined by the $I-1$ parameters $\lambda_{A(i)}$, the $J-1$ parameters $\lambda_{B(j)}$ and the $K-1$ parameters $\lambda_{C(k)}$. For Poisson data one has an additional intercept yielding the difference

$$df = IJK - (1 + I - 1 + J - 1 + K - 1) = IJK - I - J - K + 2.$$

For multinomial data, the restriction $\sum_{ijk} \mu_{ijk} = 1$ applies (reducing the number of parameters by 1) and one obtains

$$df = \{IJK - 1\} - \{I - 1 + J - 1 + K - 1\} = IJK - I - J - K + 2,$$

which is the same as for Poisson data. In general, the degrees of freedom of the approximate χ^2-distribution are the same for the sampling schemes. For obtaining asymptotically an χ^2- distribution one has to assume $\sum_i \mu_i \to \infty$ with μ_i/μ_j being constant for Poisson data and $n \to \infty$ for multinomial data and product-multinomial data, where in the latter case a constant ratio between n and the sampled subpopulation is assumed (for a derivation of the asymptotic distribution see for example Christensen (1997) , Section 2.3).

The analysis of deviance as given in Section **??** provides test statistics for the comparison of models. Models are compared by the difference in deviances. If $\tilde{M}$ is a submodel of M one considers

$$D(\tilde{M}|M) = D(\tilde{M}) - D(M). \tag{1.8}$$

The deviance (1.7) may be seen as the difference between the fitted model and the saturated model since the deviance of the saturated model, which has perfect fit, is zero.

A hierarchical submodel is always determined by assuming that part of the parameters equals zero. For example, the model AB/C assumes that

$$H_0 : \lambda_{AC(ik)} = \lambda_{BC(jk)} = \lambda_{ABC(ijk)} = 0 \quad \text{for all } i, j, k.$$

The deviance for the model AB/C may also be seen as a test statistics of null hypothesis H_0. When using the difference (1.8) one implicitly tests that the parameters that are contained in M but not in $\tilde{M}$ are zero, given model M holds.

Example 1.3: Birth Data

In the birth data example (Example 1.1) the variables are gender of the child (G, 1:male, 2:female), if membranes did rupture before the beginning of labour (M, 1:yes, 0:no), if Cesarean section has been applied (C, 1:yes, 0:no) and if the birth has been induced (I, 1:yes, 0:no). The search for an adequate model is started by fitting of models which contain all interaction terms of a specific order. Let $M([m])$ denote the model which contains all m-factor interactions. For example $M([1])$ denotes the main effect model $G/M/C/I$. From Table 1.10 it is seen that $M([3])$, $M([2])$ fit well but $M([1])$ should be rejected. Thus one considers models between $M([2])$ and $M([1])$. Starting from $M([2])$ reduced models are obtained by omitting one of the six two-factor interactions at a time. For example $M([2])\backslash GM$ denotes the

model which contains all two factor interactions except GM. The difference of deviances, e.g. for model $M([2])$ and model $M([2])\backslash GM$ is an indicator of the relevance of the interaction GM. It is seen that the interactions MC, CI and MI should not be omitted. The model $G/MC/MI/CI$ shows satisfying fit while further reduction by omitting G is inappropriate.

The model $G/MC/MI/CI$ is not a graphical model. The smallest graphical model that contains $G/MC/MI/CI$ is the model G/CMI which is shown in Figure 1.4. It means that I, C, M are interacting but are independent of gender. The gender of the child seems not to be connected to the variables membranes, Cesarean section and induced birth. □

model	dev.	df	differences	diff-df	diff-dev	p-value
M([4])	0	0				
M([3])	0.834	1	M([3])-M([4])	1	0.834	0.361
M([2])	4.765	5	M([2])-M([3])	4	3.931	0.415
M([1])	28.915	11	M([1])-M([2])	6	24.150	0.000
M([2]\GM)	5.244	6	M([2]\GM) - M([2])	1	0.478	0.489
M([2]\MC)	9.965	6	M([2]\MC) - M([2])	1	5.200	0.023
M([2]\CI)	12.167	6	M([2]\CI) - M([2])	1	7.402	0.007
M([2]\GI)	6.971	6	M([2]\GI) - M([2])	1	2.206	0.137
M([2]\GC)	6.566	6	M([2]\GC) - M([2])	1	1.801	0.180
M([2]\MI)	10.100	6	M([2]\MI) - M([2])	1	5.334	0.021
M(G/MC/MI/CI)	8.910	8	M(G/CI/MI/CI)-M([2])	3	4.145	0.246
M(MC/MI/CI)	19.428	9	M(CI/MI/CI)-M([2])	4	14.663	0.005

TABLE 1.10: Deviances and differences for log-linear models for birth data

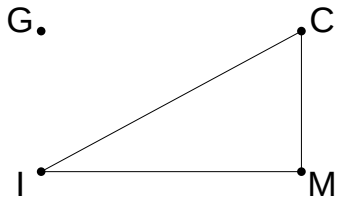


FIGURE 1.4: Graphical model for birth data

1.9 Model Selection and Regularization

Model selection is usually guided by the objective of the underlying study. If a specific association structure is to be investigated the analysis can be reduced to testing if certain interaction terms can be omitted, which is equivalent to testing the fit of

the correspondingly reduced model or, more general, a sequence of models. When no specific hypotheses are to be investigated, model selection aims at a compromise between two competing goals, sparsity and goodness-of-fit. One wants to find models that are close to the data but have an economic representation that allows simple interpretation.

Several model selection procedures for log-linear models were proposed. Some try to account for the selection error by using multiple testing strategies, others rely on screening procedures (for references see Section 1.10). More recently, regularization methods for the selection of log-linear and grahical models have been developed. The methods are particularly attractive for finding sparse solutions that fit the data well. In particular in bioinformatics the goal to identify relevant structure is very ambitious. With thousands of variables in genomics, it is to be seen if selection strategies are sufficiently reliable. However, the strategies are also useful when the number of variables is much smaller but too large for the fitting of all possible models.

A strategy that is strongly related to the regularization methods in Chapter **??** has been given by Dahinden, Parmigiani, Emerick, and Bühlmann (2007). Let $X_1, \ldots, X_p$ denote the factors where $X_j \in \{1, \ldots, k_j\}$ and $I = \{1, \ldots, p\}$ denote the index set of factors. By using subsets $A \subset I$ to define main and interaction terms the design matrix of the log-linear $\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\lambda}$ can be decomposed into

$$\boldsymbol{X} = [\boldsymbol{X}_{A_1} | \ldots | \boldsymbol{X}_{A_m}],$$

where $\boldsymbol{X}_{A_j}$ refers to a specific main or interaction term. For example $\boldsymbol{X}_{\{1,2\}}$ refers to the interaction terms of variables X_1, X_2. Correspondingly, let $\boldsymbol{\lambda}_{A_j}$ denote the vector of main or interaction parameters. The penalized log-likelihood, considered in Chapter **??**, has the form $l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{\lambda}{2} J(\boldsymbol{\beta})$, where $l(\boldsymbol{\beta})$ is the usual log-likelihood, $J(\boldsymbol{\beta})$ represents a penalty term and λ is a tuning parameter. Then the grouped lasso (Section **??**) can be applied by using the penalty

$$J(\boldsymbol{\lambda}) = \sum_{j=1}^{G} \sqrt{df_j} \|\boldsymbol{\lambda}_{A_j}\|_2,$$

where $\|\boldsymbol{\lambda}_{A_j}\|_2 = (\lambda^2_{A_j,1} + \cdots + \lambda^2_{A_j,df_j})^{1/2}$ is the L_2-norm of the parameters of the jth group of parameters, which comprises df_j parameters. The penalty encourages sparsity in the sense that either $\hat{\boldsymbol{\lambda}}_{A_j} = \mathbf{0}$ or $\lambda_{A_j,s} \neq 0$ for $s = 1, \ldots, df_j$. If one has one binary variable X_1 and a variable X_2 with three categories for example the interaction term comprises two parameters $\lambda_{12(11)}, \lambda_{12(12)}$ and the L_2-norm of the parameters is $(\lambda^2_{12(11)} + \lambda_{12(12)})^{1/2}$.

When using the grouped lasso the resulting model will in general be non-hierarchical. Of course, it is easy to fit the corresponding hierarchical model with all the necessary marginal effects included. However, if one single high order interaction term is selected the resulting model can be quite complex. Therefore, Dahinden, Parmigiani, Emerick, and Bühlmann (2007) proposed to start the selection procedure not

only from the full model but from all models $M([m])$, which contains all m factor interactions. Then the best model is selected.

A strategy that is quite common, is to start from $M([2])$, which contains all two-factor interactions. For binary variables $X_i \in \{0, 1\}$, the approach is usually based on the *Ising model*, which assumes that the joint probabilities are given by

$$P(X_1, \ldots, X_p) = \exp(\sum_{(j,k)\in E} \theta_{jk} X_j X_k - \phi(\boldsymbol{\theta})),$$

where the normalizing function $\boldsymbol{\theta}$ contains the parameters θ_{jk}, and the sum is over the edges E of a graphical model (see for example Ravikumar, Wainwright, and Lafferty (2009)). For technical reasons an artificial variable $X_0 = 1$ and edges between X_0 and the variables are included. For log-linear models the Ising model is equivalent to the multinomial model that contains all two factor interactions. For the conditional model, conditioned on the other variables, one obtains a main effect model, which in the parametrization of the Ising model is given by

$$P(X_j = 1 | X_1 = x_1, \ldots, X_p = x_p) = \frac{\exp(\sum_{(j,k)\in E} \theta_{jk} x_k)}{1 + \exp(\sum_{(j,k)\in E} \theta_{jk} x_k)}.$$

The model is equivalent to a main effect logit model with response variable X_j and explanatory variables X_k that are linked to X_j within the graph. If relevant two-factor interactions are identified it is straightforward to identify the corresponding graphical model. However, starting from a two-factor interaction model has the disadvantage that all higher interaction terms are neglected during the selection procedure. It might be more appropriate to enforce sparse modelling by administer stronger penalties on higher interaction terms or by strictly fitting hierarchical models within a boosting procedure.

Example 1.4: Birth Data

Let us consider again the birth data (Example 1.3). Figure 1.5 shows the coefficient build-ups for the fitting of a log-linear model with two-factor interactions where the two-factor interactions are penalized. The coefficient build-ups show the parameter estimates for varying degrees of smoothing λ; here they are plotted against $\|\boldsymbol{\beta}\|$. At the right end no penalty is exerted and the model that contains all two-factor interactions is fitted. The drawn lines show the two-factor interactions, the dashed lines represent the main effects. Since the main effects are not penalized they remain rather stable. The vertical lines in Figure 1.5) show the models that are selected by use of AIC and BIC . The stronger criterion, BIC, yields a model that contains only the strong interactions MC, MI, CI (Figure 1.4). The graphical model that contains these interactions is the same as found before, $G/MC/MI/CI$. If one uses AIC, in addition the rather weak interaction GI has to be included. As was to be expected BIC yields a sparser model.

□

Example 1.5: Leukoplakia

In Example 1.2 one wants to examine the association between occurrence of leukoplakia (L),

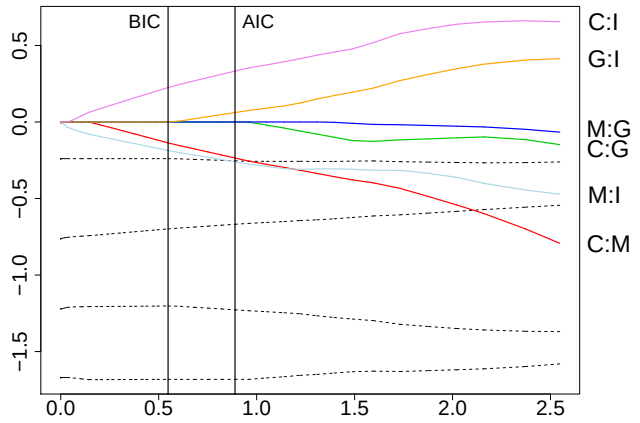


FIGURE 1.5: Coefficient build-ups for log-linear model with two-factor interactions (birth data)

alcohol intake in grams of alcohol (A) and smoking habits (S). Figure 1.6 shows the coefficient build-ups for the penalized fitting of a log-linear model with two-factor interactions. AIC as well as BIC (vertical line) suggest that the interaction between leucoplakia and alcohol intake is not needed. Leukoplakia and alcohol seem to be conditional independent given smoking habits (see also Exercise 1.10). □

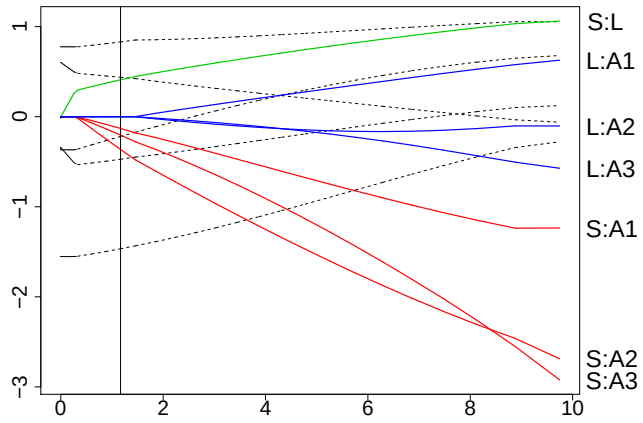


FIGURE 1.6: Coefficient build-ups for log-linear model with two-factor interactions, leucoplakia data, drawn lines show the two-factor interactions, dashed lines show the main effects

1.10 Further Reading

Surveys on Log-linear and Graphical Models. An early summary of log-linear models was given by Bishop, Fienberg, and Holland (1975) , also available as Bishop, Fienberg, and Holland (2007) . More on log-linear models is also found in Christensen (1997) . An applied treatment of graphical models is in the book of Whittaker (2008) , a more mathematical treatment is found in Lauritzen (1996).

Model Selection. Selection among models by multiple test procedures was considered by Aitkin (1979) , Aitkin (1980) . Alternative strategies including screening procedures were given by Brown (1976) , Benedetti and Brown (1978), Edwards and Havranek (1987) .

Ordinal Association. Ordinal association models, which use assigned scores, were considered by Goodman (1979), Haberman (1974), Goodman (1981), Goodman (1983), Goodman (1985), Agresti and Kezouh (1983) . An overview was given by Agresti (2009)

R packages. Log-linear models can be fitted by use of the R-function *loglin* from package stats which applies an iterative-proportional-fitting algorithm. Function *loglm* from package MASS provides a front-end to *loglin*, to allow log-linear models to be specified and fitted in a manner similar to that of other fitting functions, such as GLM.

1.11 Exercises

1.1 Consider the log-linear model for a (2×2)-contingency tables table

$$\log(\mu_{ij}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{AB(ij)}$$

with multinomial distribution and appropriate constraints.

(a) Derive the parameters $\lambda_{A(i)}, \lambda_{B(j)}, \lambda_{AB(ij)}$ as functions of odds and odds ratios for symmetrical constraints and when the last category parameter is set to zero.

(b) Show that $\lambda_{AB(ij)} = 0$ is equivalent to the independence of the variables X_A, X_B, which generate the rows and columns.

1.2 Consider the log-linear model for a (2×2)-contingency tables table

$$\log(\mu_{ij}) = \lambda_0 + \lambda_{A(i)} + \lambda_{B(j)} + \lambda_{AB(ij)}$$

which describes the distribution of product-multinomial sampling with fixed marginals $x_{i.}$.

(a) Specify appropriate constraints for the parameters.

(b) Show that $\lambda_{AB(ij)} = 0$ is equivalent to the homogeneity of response X_A across levels of X_B.

1.3 Show that the log-linear model AB/AC in a multinomial distribution $(I \times J \times K)$-contingency table is equivalent to assuming that the variables X_A and X_C are conditionally independent given X_C.

1.4 Show that for the log-linear model AB/AC ML estimates of means are given by $\hat{\mu}_{ijk} = x_{ij+}x_{i+k}/x_{i++}$. Use the estimation equations that have to hold.

1.5 Find a set of probabilities $\{\pi_{ijk}\}$ for three-way tables, where $\pi_{ijk} = P(X_A = i, X_B = j, X_C = k)$, such that X_A and X_B are conditionally independent given $X_C = k$ but the variables X_A and X_B are (marginally) dependent.

1.6 Compute the parameters of a three-way- contingency table as functions of the underlying means $\{\mu_{ijk}\}$ for symmetric side constraints.

1.7 Consider the log-linear models

$AB/AC/AD/DE$, $ABC/BCD/BDE/CDE$, $AB/BCE/CDE/AE$.

Are these models graphical? If they are draw the graph. If not give the smallest graphical model that includes the corresponding model and draw the graph.

1.8

(a) Interpret the model $AE/BC/CD/BD$.

(b) Give all the independence relations that are implied by the model $AB/BC/CD$.

1.9 The contingency table 1.11 shows data from a survey on the reading behaviour of women (Hamerle and Tutz (1980)). The cells are determined by working (yes/no), age in categories, education level (L1 to L4) and if the women is a regular reader of a specific journal. Find an appropriate log-linear model for the data.

1.10 Fit log-linear models for the leucoplakia data (Table 1.2) and select an appropriate model (compare to Example 1.5).

1.11 In contingency table 1.12 defendants in cases of multiple murders in Florida between 1976 and 1987 are classified with respect to death penalty, race of defendent and race of victim (see Agresti (2002) , Radelet and Pierce (1991)).

(a) Investigate the association between defendant's race and death penalty when victim's race is ignored (from the marginal table).

(b) Investigate the association between defendant's race and death penalty when victim's race is taken into account.

1.12 Consider the marginal association between binary factors X_A and X_B, measured in odds ratios,

$$\frac{\mu_{11+}/\mu_{12+}}{\mu_{21+}/\mu_{22+}}.$$

Show that the value is the same as for the conditional association between X_A and X_B given $X_C = k$,

$$\frac{\mu_{11k}/\mu_{12k}}{\mu_{21k}/\mu_{22k}}.$$

1.13 Consider the saturated log-linear model for three variables X, X_B, X_C and multinomial distribution. Derive the parameters of the logit model with with reference category $(X_A = I, X_B = J, X_C = K)$. What constraints hold for the parameters?

			Regular Reader	
			yes	no
Working(W)	Age (A)	Education (E)		
Yes	18-29	L1	1	14
		L2	32	49
		L3	20	34
		L4	8	3
	30-39	L1	9	23
		L2	31	57
		L3	11	26
		L4	5	7
	40-49	L1	1	33
		L2	12	50
		L3	5	11
		L4	1	7
No	18-29	L1	3	24
		L2	12	41
		L3	19	20
		L4	14	13
	30-39	L1	1	37
		L2	12	68
		L3	14	43
		L4	4	7
	40-49	L1	11	54
		L2	14	53
		L3	8	15
		L4	1	3

TABLE 1.11: Regular Reader of women's journal with employment, age, education

Victims's	Defendant's	Death Penalty	
Race	Race	Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

TABLE 1.12: Death Penalty Verdict by Defendant's Race and Victim's Race

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Agresti, A. (2009). *Analysis of Ordinal Categorical Data, 2nd Edition*. New York: Wiley.

Agresti, A. and A. Kezouh (1983). Association models for multi-dimensional cross-classifications of ordinal variables. *Comm. in Statist., Part A – Theory Meth. 12*, 1261–1276.

Aitkin, M. (1979). A simultaneous test procedure for contingency table models. *Journal of Applied Statistics 28*, 233–242.

Aitkin, M. (1980). A note on the selection of log-linear models. *Biometrics 36*, 173–178.

Benedetti, J. K. and M. B. Brown (1978). Strategies for the selection of loglinear models. *Biometrics 34*, 680–686.

Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Bishop, Y., S. Fienberg, and P. Holland (2007). *Discrete Multivariate Analysis*. New York: Springer.

Boulesteix, A.-L. (2006). Maximally selected chi-squared statistics for ordinal variables. *Biometrical Journal 48*, 451–462.

Brown, M. B. (1976). Screening effects in multidimensional contingency tables. *Journal of Applied Statistics 25*, 37–46.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. New York: Springer-Verlag.

Dahinden, C., G. Parmigiani, M. C. Emerick, and P. Bühlmann (2007). Penalized likelihood for sparse contingency tables with application to full-length cDNA libraries. *BMC Bioinformatics 8*, 476.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics 8(3)*, 522–539.

Deming, W. E. and F. F. Stephan (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.

Edwards, D. and T. Havranek (1987). A fast model selection procedure for large family of models. *Journal of the American Statistical Association 82*, 205–213.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classification having ordered categories. *Journal of the American Statistical Society 74*, 537–552.

Goodman, L. A. (1981). Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika 68*, 347–355.

Goodman, L. A. (1983). The analysis of dependence in cross-classification having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrika 39*, 149–160.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or ordered categories. *The Annals of Statistics 13*(1), 10–69.

Haberman, S. J. (1974). Loglinear models for frequency tables with ordered classifications. *Biometrics 30*, 589–600.

Hamerle, A. K. P. and G. Tutz (1980). Kategoriale Reaktionen in multifaktoriellen Versuchsplänen und mehrdimensionale Zusammenhangsanalysen. *Archiv für Psychologie*, 53–68.

Lang, J. (1996a). On the comparison of multinomial and Poisson log-linear models. *Journal of the Royal Statistical Society B*, 253–266.

Lang, J. B. (1996b). On the comparison of multinomial and Poisson log-linear models. *Journal of the Royal Statistical Society Ser. B. 58*, 253–266.

Lauritzen, S. (1996). *Graphical Models*. New York: Oxford University.

Palmgren, J. (1981). The Fisher information matrix for log-linear models arguing conditionally in the observed explanatory variables. *Biometrika 68*, 563–566.

Radelet, M. L. and G. L. Pierce (1991). . *Florida Law Rev. 43*, 1–34.

Ravikumar, P., M. Wainwright, and J. Lafferty (2009). High-dimensional graphical model selection using l_1-regularized logistic regression. *The Annals of Statistics*.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

Whittaker, J. (2008). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.