Marginal Models for Categorical Data

The original version of this manuscript was published in 1997 by Tilburg University Press, Tilburg, The Netherlands. This text is essentially the same as the original save some minor corrections and the addition of some new references. The author welcomes any comments and suggestions (email: W.P.Bergsma@kub.nl).

Marginal Models for Categorical Data

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Brabant, op gezag
van de rector magnificus, Prof.dr. L.F.W. de Klerk,
in het openbaar te verdedigen ten overstaan van een
door het college van decanen aangewezen commissie
in de aula van de Universiteit
op vrijdag 6 juni 1997 om 14.15 uur

door

Wicher Pieter Bergsma

geboren op 3 oktober 1966 te Den Haag

Tilburg University Press 1997

Promotor: Prof.dr. J. A. P. Hagenaars
Copromotor: Dr. M. A. Croon

ISBN 90-361-9837-3

Acknowledgements

Many persons have helped and motivated me in writing this book. In the first place I would like to mention Jacques Hagenaars and Marcel Croon. They ploughed through the earlier versions of the text and greatly helped in transforming them into the final version by suggesting research directions and by correcting many of my mistakes. Several others had a direct influence on the contents of this book. In particular, I would like to thank Jeroen Vermunt, Joe Lang, Tamas Rudas, Emmanuel Aris, Henk Hardonk, and John Hoogendijk for interesting and motivational discussions, and Ton Heinen for invaluable help in using LaTeX. I am grateful to other colleagues and several students who helped to make the past four years, during which I have been writing this book at Tilburg University, greatly enjoyable. I also wish to thank my family and those friends not mentioned above for their interest in my work and for providing much needed distractions outside of working hours. Finally, I am grateful for the financial support I received from the Netherlands Organization for Scientific Research (NWO), the Work and Organization Research Centre (WORC), TEFOS, and the Methodology Department of the Faculty of Social and Behavioral Sciences at Tilburg University.

Contents

1 Introduction **1**
- 1.1 Modelling marginal distributions . . . 1
- 1.2 Outline of the subsequent chapters . . . 5
- 1.3 Notation . . . 5

2 Loglinear models **7**
- 2.1 Sampling distributions . . . 8
- 2.2 The loglinear model . . . 9
- 2.3 Modelling relations between variables . . . 12
  - 2.3.1 Hierarchical loglinear models . . . 12
  - 2.3.2 Models for variables with ordered categories . . . 14
- 2.4 Maximum likelihood estimation . . . 16
  - 2.4.1 The likelihood equations . . . 17
  - 2.4.2 Iterative proportional fitting . . . 19
  - 2.4.3 Newton-Raphson . . . 21
  - 2.4.4 Comparison of Newton-Raphson and iterative proportional fitting . . . 22
- 2.5 Assessing model goodness-of-fit . . . 23
  - 2.5.1 Chi-squared tests . . . 23
  - 2.5.2 Exact tests . . . 24
  - 2.5.3 Conditional tests . . . 25
  - 2.5.4 Analysis of residuals . . . 26
- 2.6 Asymptotic behaviour of MLEs . . . 27
  - 2.6.1 The delta method . . . 27
  - 2.6.2 The asymptotic distribution of MLEs . . . 28
  - 2.6.3 The average precision of MLEs . . . 31

3 Marginal homogeneity models **35**
3.1 Introduction and notation 36
3.2 Maximum likelihood estimation 38
3.2.1 The likelihood equations 38
3.2.2 A minimization method 41
3.2.3 A maximization method 42
3.2.4 A saddle point method 44
3.2.5 Comparison of the methods 45
3.3 Alternatives to ML estimation 46
3.3.1 Minimizing the discrimination information 46
3.3.2 Minimizing Neyman's statistic 48
3.4 Assessing model goodness-of-fit 49
3.4.1 Chi-squared test statistics 49
3.4.2 A conditional test for MH using standard loglinear models 51
3.4.3 Asymptotic behaviour of MLEs 52
3.5 Example: Unaided distance vision 53

4 Marginal models **57**
4.1 Definition of marginal models 59
4.2 Applying marginal models 61
4.2.1 Modelling the joint distribution 62
4.2.2 Marginal homogeneity models 63
4.2.3 Simultaneous modelling 65
4.3 Definition and representation of measures 66
4.3.1 A recursive "exp-log" notation for representing measures 66
4.3.2 Homogeneity of measures 68
4.3.3 Marginal probabilities 69
4.3.4 The difference of proportions 70
4.3.5 Measures of association 71
4.3.6 Measures of agreement 80
4.4 Example: modelling political survey data 81

5 Fitting and testing **89**
5.1 Maximum likelihood estimation 89
5.1.1 Estimation given Poisson sampling 91
5.1.2 Estimation given sampling constraints 93

5.1.3 Remarks on maximum likelihood estimation 93
5.2 Uniqueness of MLEs for marginal models 95
5.2.1 Definition of regular models 96
5.2.2 Examples of regular models 97
5.2.3 Examples of non-regular models with multiple local maxima 101
5.3 Alternatives to maximum likelihood 103
5.3.1 Weighted least squares 105
5.3.2 Generalized estimating equations 107
5.4 Assessing model goodness-of-fit 108
5.4.1 Chi-squared tests 108
5.4.2 Partitioning chi-squared statistics 109
5.4.3 Adjusted residuals 110
5.5 Asymptotic behaviour of MLEs 111
5.6 Conclusion 113

6 Future research **115**

A Maximum likelihood theory **117**
A.1 Aitchison and Silvey's method 117
A.2 Estimation of parameters 119
A.3 Parameter orthogonality 122

B Uniqueness of MLEs **125**
B.1 A theorem about uniqueness of MLEs 125
B.2 Proof of uniqueness of MLEs 126
B.3 A suggestion for an algorithm 135

C Results from matrix algebra **137**

D Homogeneous functions **139**

E A Mathematica program **141**

Bibliography **146**

References **147**

Summary in Dutch **157**

Chapter 1

Introduction

1.1 Modelling marginal distributions

The motivation for this book stems from the apparent lack of a general and flexible methodology for testing hypotheses about relations among correlated categorical marginal distributions (Hagenaars, 1992; Laird, 1991). A basic example of a question concerning marginal distributions is the following. Consider a two-wave panel study. Suppose a researcher wants to know whether or not the marginal distributions of a categorical characteristic, e.g., party preference, have remained the same over time. Since the observations at time points 1 and 2 are correlated, a standard chi-squared test is not appropriate. In order to test the null hypothesis of no net change, the turnover table for party preference has to be set up and tested for equality of the marginal distributions. The model asserting equality of correlated marginal distributions is known as the *marginal homogeneity* (MH) model. In this book, extensions of the MH model which are useful for testing whether there are certain specific relations among correlated marginal distributions are discussed.

It is important to note that, in general, the MH model is not equivalent to a standard loglinear model. In the literature on social mobility, there has been some misunderstanding about this. It was thought that so-called structural mobility, which is mobility implied by changes in marginal distributions, could be modelled using simple restrictions on first-order loglinear parameters. Sobel, Hout, and Duncan (1985) pointed out that this is generally not appropriate, essentially because the logarithm of a

sum of terms is usually not identical to the sum of the logarithms of those terms.

Since MH and loglinear models are distinct, separate methods must be used for testing the different types of models. MH tests have received considerable attention. The literature goes back to 1947, when McNemar (1947) presented a test for 2×2 contingency tables, known as McNemar's test. Stuart (1955) and Bhapkar (1966) each described a more general quadratic test for square tables, both again named after their creators. Madansky (1963), Gokhale (1973), and Bishop, Fienberg, and Holland (1975) presented methods for finding maximum likelihood estimates.

Applications of the marginal homogeneity model are not restricted to panel studies. In social mobility research, the occupations of fathers and sons may be tested for homogeneity (Sobel, 1988). Alternatively, Stuart (1955) tested whether the distributions of the quality of the left and right eyes of subjects were identical. In medical studies, the condition of patients may be tested for homogeneity before and after treatment.

The MH model can also be used for three or higher-dimensional contingency tables. For instance, consider a three-way table ABC. It may be tested whether the distributions of the one-dimensional marginals A, B, and C is identical. Alternatively, it is possible to test whether the distributions AB and BC are identical. In a three-wave panel study, where A, B, and C represent measurements on a variable at time points 1 to 3, $AB = BC$ corresponds to the hypothesis that turnover from time point 1 to time point 2 is the same as turnover from time point 2 to time point 3. Finally, it is possible to test whether $AB = BC = AC$. For multi-way tables, there are many other possible variations of the model (Bishop et al., 1975, Chapter 8).

The models described above are examples of models for marginal distributions. It should be noted that many familiar models for categorical data are, in fact, also marginal models. For instance, Markov chain models are essentially models for marginal distributions. Furthermore, in the modified path modelling approach, because of the assumed causal order of variables, successive marginal tables of increasing dimension are analyzed (Goodman, 1973a, 1973b). For instance, for an ordered set of variables A through D, the relationship between A and B is analyzed in table AB, the effects on C are analyzed using table ABC, and the effects on D are analyzed using $ABCD$. Some but not all modified path models can be analyzed using standard loglinear models (Croon, Bergsma, & Hage-

naars, 2000). Still other models, such as cumulative logit (McCullagh, 1980; Agresti, 1990, Chapter 9) and global odds ratio models (Semenya & Koch, 1980, p. 103–118; Agresti, 1984, Section 8.2; Dale, 1986), are models for sums of frequencies. Since these sums are correlated, the latter two types of models yield the same type of problems as models for marginal distributions.

The MH models described above pertain to testing complete equality of various marginal distributions. Hypotheses of complete equality are rather strong, however. Alternatively, one can test a weaker form of MH, in particular, the equality of specific aspects of marginal distributions. For instance, for univariate marginal distributions, it may be interesting to test whether their means are identical. Other models can be used when marginal distributions are themselves "joint". For several bivariate marginal tables, it can be tested whether the association is the same in each table.

Suppose, for example, that in a panel study, party preference and preference for prime minister have been measured at two points in time. It may be interesting to test whether the association between the two variables is the same at both points. Since association can be measured using odds ratios, a test for homogeneity of association can be done by testing whether the odds ratios are identical. Again, it should be stressed that no standard chi-squared tests can be used since the odds ratios at the two points in time are correlated. The corresponding model is a loglinear model for marginal frequencies, a type that has received considerable attention, using basically three different testing approaches. These are: 1) Weighted least squares (Grizzle, Starmer, & Koch, 1969; Landis & Koch, 1979), 2) Generalized estimating equations (Liang, Zeger, & Qaqish, 1992; Diggle, Liang, & Zeger, 1994), and 3) Maximum likelihood (Haber, 1985; Haber & Brown, 1986; Agresti & Lang, 1993; Fitzmaurice & Laird, 1993; Lang & Agresti, 1994; Molenberghs & Lesaffre, 1994; Glonek & McCullagh, 1995; Becker, 1994). A review of the different methods is provided by Fitzmaurice, Laird, and Rotnitzky (1993).

When testing the homogeneity of certain aspects of marginal distributions, it is often possible to use different types of measures for the aspect to be measured. Above, odds ratios were mentioned as a possible candidate for measuring association. Of course, there are many other possibilities, such as global odds ratios (Clayton, 1974), gamma, or Kendall's tau (Goodman & Kruskal, 1979). It should be noted that odds ratios pro-

vide a set of measures describing association in a contingency table, while gamma or Kendall's tau yields only a single number.

Other aspects of marginal distributions besides association or the means can be tested for homogeneity, provided one or more appropriate measures summarizing this aspect are available. For instance, agreement can be measured using the agreement measure kappa. Agresti (1990, p. 367–370) showed how it can be modelled using "diagonal" odds ratios.

Equality of aspects of marginal distributions are not the only features that can be modelled. When more than two marginal tables are of interest, a regression model may be used to test whether the distributions of the tables are related in a specific way. In a panel study consisting of more than two waves, it is possible to test whether there is a linear increase over time in the association between two variables. For example, it is possible to test whether association between party preference and preference for prime minister increases as the election date approaches.

The models described above lead to complex methods for testing goodness-of-fit, partly because of the correlations between marginal distributions and partly because of the mathematical complexity of the various measures. Generally, the correlations between marginal distributions also depend on higher order moments. The simplest way to test a model for correlated marginal distributions is the weighted least squares (WLS) or GSK procedure (Grizzle et al., 1969; Kritzer, 1977). It requires no iterative methods and therefore the computational complexity is very low compared to, for instance, maximum likelihood (ML). However, WLS is very sensitive to sparse data, and apparantly because of this, other methods of testing various marginal models have been sought. ML methods were considered too difficult so alternatives were devised, such as the quasi-likelihood (Wedderburn, 1974) and the generalized estimating equations approaches (Liang & Zeger, 1986). A disadvantage of these methods is that they do not yield statistical models in the proper sense since they are not based on a probability distribution (Lindsey, 1993, Section 2.9). Maximum likelihood is preferred by many statisticians (see, for instance, the discussion of the paper by Liang et al., 1992). This book emphasizes maximum likelihood estimation. A general specification for marginal distribution models is presented and a general method for fitting and testing them is given.

1.2 Outline of the subsequent chapters

Chapter 2 discusses loglinear modelling and serves as an introduction to the basic concepts used in this book. Some important loglinear models are presented, and it is shown how they can be fitted using the maximum likelihood method. Testing goodness-of-fit and the asymptotic behaviour of maximum likelihood estimates of various parameters is discussed. In the third chapter, it is shown how the marginal homogeneity model and, more generally, models inducing linear constraints on expected frequencies can be tested. These models are mainly applied to test equality of various marginal distributions, and are relatively easy to analyze. The aim of the third chapter is to give an overview of the most important literature on the marginal homogeneity model.

Chapters 4 and 5 form the core of the book. In Chapter 4, a general class of models, referred to as marginal models, is presented. Marginal models generalize both the loglinear models of Chapter 2 and the marginal homogeneity models of Chapter 3. They can be used to test whether there are specific relationships between marginal distributions, but also provide means of modelling joint distributions and various simultaneous models. The marginal model specification is discussed in Chapter 4; methods for testing and fitting these models are presented in Chapter 5. Emphasis is on maximum likelihood estimation, though a brief description is given of the generalized estimating equations approach and WLS. A modified maximum likelihood fitting algorithm based on work by Aitchison and Silvey (1958; 1960), Haber (1985), Lang and Agresti (1994), and Lang (1996a) is presented. It is shown that, for an important class of loglinear models for marginal frequencies, the likelihood function is uniquely maximized subject to the model constraints. This greatly eases maximum likelihood estimation. An overview of some unresolved problems relating to marginal modelling is presented in Chapter 6.

1.3 Notation

Vectors and matrices are always represented in boldtype, where matrices are capitalized, and vectors are written using small letters. Scalars are in normal letters. Matrix notation will be used frequently throughout the book. It may take some time to get used to this type of notation,

but I think it pays to make the effort, since many derivations of various formulas are simplified considerably by using matrix algebra. In general, a diagonal matrix with the elements of a vector $\mathbf{q}$ on the main diagonal is denoted as $\mathbf{D}_{\mathbf{q}}$. Functions of vectors, such as $\log \mathbf{q}$ or $\exp(\mathbf{q})$, are generally applied elementwise. A vector of ones is denoted as "$\mathbf{1}$", a vector of zeroes as "$\mathbf{0}$", and a matrix of zeroes using the slightly bigger symbol "$\mathbf{0}$" (since it will be clear from the context whether the vector or the matrix of zeroes is denoted, a microscope will not be needed to distinguish the latter two symbols).

The derivative of a vector of functions $\mathbf{f}$ with respect to a vector of variables $\mathbf{x}$ can be taken in two ways, namely, by the following matrix and its transpose:

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}'} \quad \text{or} \quad \frac{\partial \mathbf{f}'}{\partial \mathbf{x}}$$

On the left, one has a matrix with (i, j)th element $\partial f_i / \partial x_j$, on the right is a matrix with (i, j)th element $\partial f_j / \partial x_i$.

The expected frequencies in a contingency table are denoted by the vector $\mathbf{m}$, and the observed frequencies by the vector $\mathbf{n}$. Expected and observed marginal frequencies are denoted by the vectors $\boldsymbol{\mu}$ and $\mathbf{y}$, respectively. Maximum likelihood estimates (MLEs) are denoted using a "hat", i.e., the MLEs of $\mathbf{m}$ are written as $\hat{\mathbf{m}}$. The asymptotic covariance matrix of a vector, say $\hat{\mathbf{m}}$, is denoted as $\boldsymbol{\Sigma}(\hat{\mathbf{m}})$.

It should be noted that for better readability of this book the mathematical notation used is not entirely rigorous. In particular, depending on the context certain symbols may sometimes represent a variable and sometimes the population value of a parameter. This is in line with what is often done in books on applied statistics.

Chapter 2

Loglinear models

In the past three decades, the loglinear model has gained wide popularity in categorical data analysis. Several textbooks present a comprehensive overview of the field (Haberman, 1974, 1978, 1978; Bishop et al., 1975; Fienberg, 1980; Hagenaars, 1990; Agresti, 1990, 1996). The aim of this chapter is to introduce the basic concepts of loglinear modelling that are necessary for an understanding of the main chapters of this book.

Different sampling schemes, in particular Poisson and (product) multinomial, are described in Section 2.1. These are presented before the loglinear model itself because the precise form of the loglinear model is dependent on the sampling scheme. A basic outline of the loglinear model and the notation used is given in Section 2.2. Two important types of models are explained in Section 2.3, namely, hierarchical loglinear models and models for variables with ordered categories. In Section 2.4, maximum likelihood methods for loglinear models are presented. Included are proofs of existence and uniqueness of maximum likelihood estimates (MLEs). Two algorithms for finding MLEs are described: iterative proportional fitting and Newton-Raphson. Testing goodness-of-fit is discussed in Section 2.5. chi-squared test statistics and exact testing methods are given. Furthermore, attention is paid to conditional testing of a hypothesis against a directed alternative. To analyse why there is lack of fit of a model, the analysis of cell residuals is also explained. The chapter ends with Section 2.6, which deals with the asymptotic behaviour of MLEs, and is the most technical part of the chapter. General methods for deriving asymptotic distributions of estimators are explained.

2.1 Sampling distributions

Commonly used sampling distributions for categorical data are Poisson and (product) multinomial. A Poisson distribution implies that the events to be counted occur randomly over time or space, and that the outcomes in disjunct periods are independent. A Poisson distribution becomes a multinomial one when the sample size is fixed a priori.

For Poisson sampling, the probability of observing the counts $\mathbf{n} = (n_1, n_2, \ldots, n_r)'$ with expected values $\mathbf{m} = (m_1, m_2, \ldots, m_r)'$ is

$$\prod_i \frac{m_i^{n_i} e^{-m_i}}{n_i!}. \tag{2.1}$$

Multinomial sampling is obtained when the total number of observations n_+ in (2.1) is fixed by design, i.e., when

$$m_+ = n_+, \tag{2.2}$$

where the "+" in the subscript represents a summation over the index values, i.e., $m_+ = \sum_i m_i$. In vector notation, (2.2) can be written as

$$\mathbf{1}'\mathbf{m} = \mathbf{1}'\mathbf{n},$$

where $\mathbf{1}$ is a $r \times 1$ vector of 1's. For two multinomial samples with expected frequencies m_{1j} and m_{2j} respectively, one has

$$m_{1+} = n_{1+} \tag{2.3}$$
$$m_{2+} = n_{2+}. \tag{2.4}$$

Here, the frequencies can be assembled in a $2 \times r$ contingency table with fixed row totals. With I samples and J categories per sample, the expected frequencies can be put in a rectangular contingency table with I rows and J columns. For such a table, the row totals are fixed by design.

In matrix notation, with $\mathbf{m} = (m_{11}, \ldots, m_{1r}, m_{21}, \ldots, m_{2r})'$ and $\mathbf{n} = (n_{11}, \ldots, n_{1r}, n_{21}, \ldots, n_{2r})'$, the equations (2.3) and (2.4) can be written as

$$\begin{pmatrix} \mathbf{1}' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}' \end{pmatrix} \mathbf{m} = \begin{pmatrix} \mathbf{1}' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{1}' \end{pmatrix} \mathbf{n}.$$

In general, when linear combinations of cells are fixed by design, these restrictions can be represented, using the appropriate matrix $\mathbf{W}$, by the formula

$$\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}. \tag{2.5}$$

Though the restrictions will usually represent a (product) multinomial design, the notation allows for more generality.

2.2 The loglinear model

Loglinear modelling is used to model the relations between categorical variables. A loglinear model defines a *multiplicative* structure on the expected cell frequencies of a contingency table. The expected frequencies are regarded as the product of a number of parameters which can be interpreted in useful ways.

Below, following Lang (1996b) and Aitchison and Silvey (1960), two approaches to the loglinear model specification will be given. First is the more common *freedom equation* approach, where the cell frequencies are written as the product of a number of freedom parameters, or, equivalently, where the log cell frequencies are written as the sum of freedom parameters. The adjective "freedom" is used because the more freedom parameters are used, the less restricted the expected cell frequencies are. That is, for the saturated loglinear model, the expected cell frequencies are unrestricted by the model, and the number of independent freedom parameters is maximal, i.e., equal to the number of cells. The second approach to specifying a loglinear model is the *constraint equation* approach, where the loglinear model is written in terms of multiplicative constraints on the expected frequencies, or, equivalently, in terms of linear constraints on the log expected frequencies. This approach is less common, but will be shown to be useful for certain calculations.

In general, when a certain loglinear model holds, the expected frequencies m_i $(i = 1, \ldots, r)$ of a contingency table can be written in the form

$$\log m_i = \sum_{j=1}^{b} x_{ij}\beta_j, \qquad \forall i \tag{2.6}$$

where the x_{ij} are constants, and the β_j are unknown parameters. In matrix notation, with $\mathbf{m} = (m_1, \ldots, m_r)'$ denoting the vector of expected frequencies, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_b)'$ the vector of loglinear parameters, and $\mathbf{X}$ the $r \times b$ matrix with elements x_{ij}, the expected frequencies given a loglinear model can be written as

$$\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta}, \tag{2.7}$$

where the logarithm is taken elementwise. The matrix $\mathbf{X}$ is called the *design matrix*.

For Poisson sampling, a model is defined to be loglinear if it can be defined by constraint (2.7) with the restriction on $\mathbf{X}$ that

$$\text{range}(\mathbf{X}) \supset \text{range}(\mathbf{1}), \tag{2.8}$$

where the range of a matrix (or vector) is defined as the vector space spanned by its columns. The restriction (2.8) means that there is a vector $\mathbf{w}$ such that $\mathbf{X}'\mathbf{w} = \mathbf{1}$. For instance, the restriction is satisfied if the vector $\mathbf{1}$ is a column of $\mathbf{X}$. When the expected frequencies are subject to the sampling restrictions (2.5), the restriction that will be imposed on the design matrix $\mathbf{X}$ is

$$\text{range}(\mathbf{X}) \supset \text{range}(\mathbf{W}, \mathbf{1}), \tag{2.9}$$

where $\text{range}(\mathbf{W}, \mathbf{1})$ is the space spanned by the columns of $\mathbf{W}$ and the vector $\mathbf{1}$. These restrictions ensure that the same model is obtained whether the model equations are specified for the expected cell frequencies or for the expected cell probabilities $\boldsymbol{\pi}$, i.e., that for every $\boldsymbol{\beta}$, a $\boldsymbol{\beta}^*$ can be found such that

$$\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta} \qquad \Leftrightarrow \qquad \log \boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}^*. \tag{2.10}$$

The restrictions (2.8) and (2.9) do not seem to exclude many models of practical interest. They are used because (2.10) is a convenient property, ensuring that the same maximum likelihood fitting methods can be used under all sampling restrictions. The definition of loglinear models is slightly nonstandard, because usually only the requirement (2.8) is made rather than (2.9). For the more standard definition, see Lauritzen (1996, p. 72).

An alternative way of specifying a loglinear model is by using constraints on the expected frequencies, rather than by parameterizing them. This is the constraint equation approach. For instance, independence in a 2×2 table with expected frequencies $(m_{11}, m_{12}, m_{21}, m_{22})$ can be defined by requiring the odds ratio to equal one, i.e., by imposing the constraint equation

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

Taking logarithms on both sides, one gets

$$\log m_{11} - \log m_{12} - \log m_{21} + \log m_{22} = 0.$$

In matrix notation, the latter equation can be written as

$$\begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \log \begin{pmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \end{pmatrix} = 0,$$

where the logarithm of the vector is taken elementwise. More generally, any loglinear model can be written in the form

$$\mathbf{C}' \log \mathbf{m} = \mathbf{0}, \tag{2.11}$$

where $\mathbf{C}$ is a matrix of constants. Restriction (2.9) on $\mathbf{X}$ is equivalent to the restriction on $\mathbf{C}$ that

$$\mathbf{C}'\mathbf{W} = \mathbf{0}. \tag{2.12}$$

For instance, for multinomial sampling, $\mathbf{W} = \mathbf{1}$, and (2.12) is equivalent to the requirement that $\mathbf{C}'$ is a contrast matrix, i.e., that the rows sum to zero.

Equation (2.11) is called a *constraint equation*, while equation (2.7) is called a *freedom equation*. It is important to see that, provided all expected frequencies are greater than zero, both formulations (2.7) and (2.11), subject to restrictions (2.9) and (2.12) respectively, describe the same class of models. The formulations are equivalent if and only if $\mathbf{C}$ is the orthogonal complement of $\mathbf{X}$, i.e., if and only if $\mathbf{C}'\mathbf{X} = \mathbf{0}$ and the rows of $\mathbf{C}$ and $\mathbf{X}$ together span the whole vector space of the expected frequencies.

2.3 Modelling relations between variables

2.3.1 Hierarchical loglinear models

Consider a frequency table formed by three categorical variables, say A, B, and C, which have I, J, and K categories respectively. Let m_{ijk} be the expected frequency in cell (i, j, k) in frequency table ABC. Then the saturated loglinear model for the three-way table ABC is given by the parameterization

$$m_{ijk} = \tau \times \tau_i^A \times \tau_j^B \times \tau_k^C \times \tau_{ij}^{AB} \times \tau_{jk}^{BC} \times \tau_{ik}^{AC} \times \tau_{ijk}^{ABC}, \qquad (2.13)$$

where all parameters are constrained to be positive. Taking the logarithm on both sides, (2.13) is equivalent to the loglinear form

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}, \qquad (2.14)$$

where

$$\begin{array}{ccccc} \lambda = \log \tau & \lambda_i^A = \log \tau_i^A & \lambda_j^B = \log \tau_j^B & \lambda_k^C = \log \tau_k^C & \\ \lambda_{ij}^{AB} = \log \tau_{ij}^{AB} & \lambda_{jk}^{BC} = \log \tau_{jk}^{BC} & \lambda_{ik}^{AC} = \log \tau_{ik}^{AC} & \lambda_{ijk}^{ABC} = \log \tau_{ijk}^{ABC} & . \end{array}$$

The saturated model does not restrict the expected frequencies in any way, and any other loglinear model is a special case of the saturated model.

Equations (2.13) and (2.14) contain too many parameters to be identifiable. Given the expected frequencies m_{ijk}, there is no unique solution for the λ and τ parameters. Restricting attention to formula (2.14), identification of the parameters can be achieved by imposing additional constraints, such as the following:

$$\begin{aligned} &\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_k \lambda_k^C = 0 \\ &\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = \sum_j \lambda_{jk}^{BC} = \sum_k \lambda_{jk}^{BC} = \sum_i \lambda_{ik}^{AC} = \sum_k \lambda_{ik}^{AC} = 0 \\ &\sum_i \lambda_{ijk}^{ABC} = \sum_j \lambda_{ijk}^{ABC} = \sum_k \lambda_{ijk}^{ABC} = 0 \end{aligned}$$

for all i, j, and k. This method of identifying the parameters is called *effect* coding. An alternative method of identification is *dummy* coding,

where certain parameters are set to zero, e.g.,

$$\lambda_I^A = \lambda_J^B = \lambda_K^C = 0$$
$$\lambda_{iJ}^{AB} = \lambda_{Ij}^{AB} = \lambda_{jK}^{BC} = \lambda_{Jk}^{BC} = \lambda_{iK}^{AC} = \lambda_{Ik}^{AC} = 0$$
$$\lambda_{ijK}^{ABC} = \lambda_{iJk}^{ABC} = \lambda_{Ijk}^{ABC} = 0$$

for all i, j, and k.

Effect coding is used in most applications of loglinear modelling. Using this parameterization, the following interpretation can be given to the parameters (Alba, 1987). The parameter λ is called the *intercept*. It is equal to the mean of the log expected frequencies, i.e.,

$$\lambda = \frac{1}{IJK} \sum \log m_{ijk}.$$

Equivalently, $\tau = \exp(\lambda)$ is equal to the geometric mean of the expected frequencies. The intercept ensures that the loglinear model is the same whether specified for probabilities or frequencies. The one-variable parameters λ_i^A, λ_j^B, and λ_k^C are equal to the average deviation from λ of the log expected frequencies in category i of A, j of B, and k of C respectively. Normally, these parameters are included in a loglinear model. The parameters λ_{ij}^{AB}, λ_{jk}^{BC}, and λ_{ik}^{AC} reflect the strength of the association between A and B, B and C, and A and C respectively, given the level of the third variable. Finally, the λ_{ijk}^{ABC} parameters indicate how much the conditional two-variable effects differ from one another within the categories of the third variable. The latter parameters are also called the three-variable, or three-factor interaction effects.

Loglinear models other than the saturated model are obtained by imposing additional linear constraints on the loglinear parameters. The no-three-factor interaction model can be obtained by setting the term λ_{ijk}^{ABC} to zero for all (i, j, k), which yields

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC}.$$

For this model, the association between any two variables is the same whatever the level of the third variable. To model complete statistical independence between A, B, and C, all two-variable effect parameters should additionally be set to zero. The following parameterization is obtained

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C.$$

In general, the hierarchical no k-factor interaction model is defined as the model for which all k-factor and higher-order effects are set to zero.

Models of the type described above are examples of *hierarchical loglinear models*. A loglinear model is called hierarchical if, when a certain λ parameter is set to zero, all effects of the same or a higher order which include all the letters of the superscript of this λ parameter are also set to zero. For instance, if in equation (2.14) the two-factor interaction parameters λ_{ij}^{AB} are set to zero, then all the higher-order effects λ_{ijk}^{ABC} (i.e., those parameters which include AB in the superscript) must also be set to zero.

Given the sampling restrictions (2.5), the requirement (2.9) implies that not all parameters can be set to zero. If certain marginals are fixed by design, the requirement implies that the parameters pertaining to the marginal frequencies that are fixed by design must be included in the model. For instance, if, for a three-way table, the sampling design is such that $m_{ij+} = n_{ij+}$, then the parameters λ, λ_i^A, λ_j^B, and λ_{ij}^{AB} in (2.14) must be included in the model. For a single multinomial sample, the intercept λ must be included.

2.3.2 Models for variables with ordered categories

In many situations, variables with ordered categories are used. The models described in the previous section treat the variables as nominal and do not exploit this ordering. In some situations, the no-three-factor interaction model fits the data well, while the complete independence model is too restrictive and yields a bad fit. If variables have ordered categories, one can try fitting a model which is more parsimonious than no-three-factor interaction, but less restrictive than complete independence, and which takes account of the ordering of the categories of the variables. Below, two types of loglinear models, developed by Haberman (1979, Chapter 6), will be discussed. For other loglinear and non-loglinear models for modelling the association, see Goodman (1984) or Clogg and Shihadeh (1994).

The linear-by-linear association model

Consider an $I \times J$ table with expected frequencies m_{ij}, pertaining to variables A and B. The saturated model is

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}.$$

For the *linear-by-linear* model, fixed ordered scores are assigned to the row and column categories. Assigning fixed values u_i^A to the row and u_j^B to the column categories, the interaction parameter λ_{ij}^{AB} is decomposed as $\beta^{AB} u_i^A u_j^B$, where the β^{AB}-parameter reflects the strength of association. The following formula is obtained:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \beta^{AB} u_i^A u_j^B. \tag{2.15}$$

Variables A and B are independent if $\beta^{AB} = 0$. With monotone category orderings,

$$u_1^A \leq u_2^A \leq \cdots \leq u_I^A$$
$$u_1^B \leq u_2^B \leq \cdots \leq u_J^B$$

should be taken. Identification of the parameters can be achieved by requiring $\sum \lambda_i^A = \sum \lambda_j^B = 0$. The model has one parameter more than the independence model, so the number of degrees of freedom is one less than for the independence model, i.e., $(I-1)(J-1)-1$. Let ζ_{ij} be the local odds ratios, defined as

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

for $i = 1, \ldots, I-1$ and $j = 1, \ldots, J-1$. It follows from (2.15) that

$$\log \zeta_{ij} = \beta^{AB}(u_i^A - u_{i+1}^A)(u_j^B - u_{j+1}^B). \tag{2.16}$$

It can be seen that the association, measured in terms of the log odds ratios, is a linear function of the differences of both the successive row and the successive column scores, hence the name "linear-by-linear". With equally spaced row and column scores, the log odds ratios are equal to β^{AB} everywhere.

The model can be generalized in a straightforward manner when there are more than two variables. For a three-way table, when the model of no-three-factor interaction holds, with one β parameter for the association between each pair of variables,

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \beta^{AB} u_i^A u_j^B + \beta^{AC} v_i^A v_k^C + \beta^{BC} w_j^B w_k^C$$

can be used. One may take $u_i^A = v_i^A$, $u_j^B = w_j^B$, and $v_k^C = w_k^C$.

The row-effects and column-effects models

Here, only $I \times J$ tables will be considered. If either the row or the column effect scores are unknown, they can, instead of using fixed values, be estimated. Substituting μ_i for $\beta^{AB} u_i^A$ in (2.15), the loglinear model

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \mu_i u_j^B$$

is obtained The u_j^B are fixed constants, and the μ_i are parameters called *row effects*, hence the name "row-effects model." Analogously, the column-effects model can be defined. Identification can be achieved by requiring $\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_i \mu_i = 0$. The number of degrees of freedom is $(I-1)(J-2)$, which is $I-1$ less than for the independence model.

2.4 Maximum likelihood estimation

For loglinear models, the maximum likelihood (ML) method has some appealing properties. The log likelihood is a concave function of the parameters resulting in a unique maximum. An elegant and intuitive estimation algorithm (iterative proportional fitting) also exists. Furthermore, estimates are identical for Poisson and multinomial sampling.

First, the likelihood equations will be derived, and existence and uniqueness of maximum likelihood estimators (MLEs) will be proven. Then two estimation algorithms will be described and compared, namely, iterative proportional fitting (IPF), which also has an appealing interpretation, and Newton-Raphson (N-R). A comparison of the two will be given in Section 2.4.4.

2.4.1 The likelihood equations

The likelihood equations for Poisson sampling are easiest to derive, so this will be done first. Then it will be shown that, if there are sampling constraints $\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}$, the resulting likelihood equations are the same as the Poisson equations. Thus, no separate fitting methods are required for any of the other sampling distributions described in Section 2.1.

For Poisson sampling, the probability of observing the frequencies $\mathbf{n} = (n_1, n_2, \ldots, n_r)'$ given the expected frequencies $\mathbf{m} = (m_1, m_2, \ldots, m_r)'$ is given by (2.1). The MLEs are defined as those expected frequencies $\mathbf{m}$, which maximize this probability. In practice, it is usually easier to maximize the *kernel* of the logarithm of (2.1), which is

$$\mathcal{L} = \sum_i (n_i \log m_i - m_i), \tag{2.17}$$

i.e., the logarithm of (2.1) is taken and the terms not dependent on any of the m_i are left out. In order to maximize $\mathcal{L}$ subject to the loglinear model $\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta}$, $\mathcal{L}$ can be differentiated with respect to the model parameters $\boldsymbol{\beta}$ and the result equated to zero. This yields the following equations:

$$k_i = \frac{\partial \mathcal{L}}{\beta_j} = \sum_i x_{ij} n_i - \sum_i x_{ij} m_i = 0.$$

In matrix notation,

$$\mathbf{k} = \mathbf{X}'\mathbf{n} - \mathbf{X}'\mathbf{m} = \mathbf{0} \tag{2.18}$$

is derived. Together with (2.6), these equations form the *likelihood equations*, and the MLEs are their solutions (Haberman, 1974).

A nice property of ML estimation is that not all the information from all cells is needed to calculate the MLEs: the statistics formed by the elements of $\mathbf{X}'\mathbf{n}$ are sufficient. These are called the *sufficient statistics* for the MLEs. For hierarchical loglinear models, the sufficient statistics are certain marginal frequencies. For example, for the independence model for a three-way table, the one-dimensional marginal frequencies n_{i++}, n_{+j+}, and n_{++k} are sufficient statistics. For the no-three-factor interaction model for three-way tables, the two-dimensional marginal frequencies n_{ij+}, n_{i+k}, and n_{+jk} are sufficient statistics. In general, when there are v

variables, the sufficient statistics for the hierarchical no-f-factor interaction model, with $f \leq v$, are the $(f-1)$-dimensional marginal frequencies.

It will be shown next that, given a Poisson sampling scheme and provided $n_i > 0$ for all i, an MLE $\hat{\boldsymbol{\beta}}$ exists and is unique (Birch, 1963; Haberman, 1974). It is assumed here that $\mathbf{X}$ has independent columns. The matrix of second derivatives $\mathbf{K}$ of $\mathcal{L}$ is

$$\mathbf{K} = \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\mathbf{X}'\mathbf{D}_{\mathbf{m}}\mathbf{X}, \tag{2.19}$$

where $\mathbf{D}_{\mathbf{m}}$ denotes a diagonal matrix with the elements of $\mathbf{m}$ on the main diagonal. Matrix $\mathbf{K}$ is negative definite if all $m_i > 0$ and if the columns of $\mathbf{X}$ are independent. The former is the case for all finite $\boldsymbol{\beta}$, and the latter was assumed, so $\mathcal{L}$ is a concave function of $\boldsymbol{\beta}$ on the whole parameter space. It can be noted that $\mathcal{L} \to -\infty$ as $\log m_i \to \pm\infty$. This means that $\mathcal{L}$ is maximized in the interior of the parameter space. As $\mathcal{L}$ is also concave, $\mathcal{L}$ has a unique maximum.

If, for some h, the observed frequency $n_h = 0$, the definition of the loglinear model can be extended so that the MLEs $\hat{m}_i$ exist and are unique, though some parameter estimates may tend to plus or minus infinity (Lauritzen, 1996, p. 73). Even if the sufficient statistics for the model are all strictly positive, it may still be the case that some parameter estimates do not exist (Haberman, 1974). Positivity of sufficient statistics only guarantees existence of parameter estimates if the model is decomposable (Glonek, Darroch, & Speed, 1988). Decomposable models also have closed form solutions to the likelihood equations.

The MLEs given Poisson sampling can be shown to satisfy the multinomial sampling constraints. Consider the sampling scheme $\mathcal{S}$ defined by $\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}$ and the loglinear model defined by $\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta}$. From basic linear algebra, requirement (2.9) implies the existence of a matrix $\mathbf{L}$ such that $\mathbf{W} = \mathbf{X}\mathbf{L}$. Thus, if the likelihood equations (2.18) are satisfied, it follows that

$$\mathbf{W}'\hat{\mathbf{m}} = \mathbf{L}'\mathbf{X}'\hat{\mathbf{m}} = \mathbf{L}'\mathbf{X}'\mathbf{n} = \mathbf{W}'\mathbf{n}.$$

It can be seen that the MLEs for Poisson sampling automatically satisfy the sampling constraints implied by $\mathcal{S}$. Thus, the existence and uniqueness results that were given above also apply for sampling scheme $\mathcal{S}$.

2.4.2 Iterative proportional fitting

A simple and elegant algorithm for calculating MLEs is the *iterative proportional fitting* (IPF) procedure. It works by successively scaling the cell frequencies to match the successive estimates of their sufficient statistics. Starting values must satisfy the model. The method is illustrated by an example. The no-three-factor interaction model for three-way tables has sufficient statistics n_{ij+}, n_{+jk}, and n_{i+k} for all i, j, and k. Since starting values may not contain effects not included in the model, they may be set to one:

$$m_{ijk}^{(0)} = 1.$$

The first cycle of the estimation process is

$$\begin{aligned} m_{ijk}^{(1)} &= m_{ijk}^{(0)} \frac{n_{ij+}}{m_{ij+}^{(0)}} \\ m_{ijk}^{(2)} &= m_{ijk}^{(1)} \frac{n_{+jk}}{m_{+jk}^{(1)}} \\ m_{ijk}^{(3)} &= m_{ijk}^{(2)} \frac{n_{i+k}}{m_{i+k}^{(2)}}. \end{aligned}$$

The second cycle is of the same form as the first but uses the updated estimates:

$$\begin{aligned} m_{ijk}^{(4)} &= m_{ijk}^{(3)} \frac{n_{ij+}}{m_{ij+}^{(3)}} \\ m_{ijk}^{(5)} &= m_{ijk}^{(4)} \frac{n_{+jk}}{m_{+jk}^{(4)}} \\ m_{ijk}^{(6)} &= m_{ijk}^{(5)} \frac{n_{i+k}}{m_{i+k}^{(5)}}. \end{aligned}$$

The algorithm can be terminated at iteration h when the hth estimates are close to satisfying the likelihood equations

$$m_{ij+} = n_{ij+} \qquad m_{+jk} = n_{+jk} \qquad m_{i+k} = n_{i+k}.$$

In the literature it is sometimes suggested to to stop the iterations when the difference between successive values of the log likelihood is small. As

was noted by Bishop, Fienberg, and Holland (1975), this criterion does not guarantee that expected cell frequencies are accurate. For this reason, they proposed to stop the iterations when differences in successive values of expected cell frequencies are small. However, the latter criterion still does not guarantee that the algorithm is close to convergence, since, even if cell values change little per iteration, many small changes can add up to a large change. A better convergence criterion which can be used is that, for the hth estimates,

$$\sum_{i,j,k} \left[(m_{ij+}^{(h)} - n_{ij+})^2 + (m_{+jk}^{(h)} - n_{+jk})^2 + (m_{i+k}^{(h)} - n_{i+k})^2 \right] < \epsilon^2$$

for a sufficiently small value of ϵ (say $\epsilon = 10^{-10}$). This guarantees that the sufficient statistics are close to being reproduced.

An attractive property of IPF is that, if direct estimates for the cell frequencies exist and if the dimension of the table is less than seven, the algorithm finds them after just one full cycle (Haberman, 1974, p. 197). For instance, it can be verified that, for the independence model for a three-way table, the MLEs

$$\hat{m}_{ijk} = \frac{1}{n_{+++}} n_{i++} n_{+j+} n_{++k}$$

are obtained after one full cycle.

Csiszár (1975) showed that IPF can be viewed as a series of I-projections, and gave an elegant information-theoretic proof of convergence. A different algorithm called *generalized iterative scaling* (GIS) was developed by Darroch and Ratcliff (1972). For hierarchical log-linear models they showed it can be reduced to IPF. The relation between IPF and GIS was clarified by Csiszár (1989)

General rules for the number of cycles needed to reach satisfactory convergence are difficult to give. However, the rate of convergence of IPF is first-order, which means that the convergence satisfies

$$|\hat{\beta}_i^{(k+1)} - \hat{\beta}_i| \leq c\, |\hat{\beta}_i^{(k)} - \hat{\beta}_i| \quad \text{for some } c > 0.$$

This implies that the number of iterations needed to obtain an extra digit of accuracy is bounded. Thus, if the dth digit has been calculated, one needs, at most, a certain number of iterations, say u, to calculate the $(d+1)$th digit. However, u is unknown and is possibly large.

2.4.3 Newton-Raphson

To maximize $\mathcal{L}$ with respect to $\boldsymbol{\beta}$ using N-R, a sequence of estimates is calculated using the "updating" function $\mathbf{v}$, defined as

$$\mathbf{v}(\boldsymbol{\beta}, \mathit{step}) = \boldsymbol{\beta} - \mathit{step}\,\mathbf{K}^{-1}\mathbf{k}. \tag{2.20}$$

Here $\mathbf{k}$ and $\mathbf{K}$ are defined by (2.18) and (2.19) respectively, and *step* is a step size. With $\mathbf{m} = \exp(\mathbf{X}\boldsymbol{\beta})$, (2.20) becomes

$$\mathbf{v}(\boldsymbol{\beta}, \mathit{step}) = \boldsymbol{\beta} + \mathit{step}\,(\mathbf{X}'\mathbf{D}_\mathbf{m}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{n} - \mathbf{m}). \tag{2.21}$$

Note that (2.21) can be rewritten as

$$\begin{aligned} \mathbf{v}(\boldsymbol{\beta}, \mathit{step}) &= (\mathbf{X}'\mathbf{D}_\mathbf{m}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\mathbf{m}\left(\mathbf{X}\boldsymbol{\beta} + \mathit{step}\mathbf{D}_\mathbf{m}^{-1}(\mathbf{n} - \mathbf{m})\right) \\ &= (\mathbf{X}'\mathbf{D}_\mathbf{m}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\mathbf{m}\left(\log\mathbf{m} + \mathit{step}\mathbf{D}_\mathbf{m}^{-1}(\mathbf{n} - \mathbf{m})\right). \end{aligned} \tag{2.22}$$

The following algorithm can now be used. As starting values $\boldsymbol{\beta}^{(0)}$, one can take (2.22) with $\mathbf{m}$ substituted by $\mathbf{n}$ and $\mathit{step} = 1$. Then, for $k = 0, 1, \ldots$,

$$\begin{aligned} \boldsymbol{\beta}^{(0)} &= (\mathbf{X}'\mathbf{D}_\mathbf{n}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\mathbf{n}\log\mathbf{n} \\ \boldsymbol{\beta}^{(k+1)} &= \mathbf{v}(\boldsymbol{\beta}^{(k)}, \mathit{step}^{(k)}). \end{aligned} \tag{2.23}$$

If some observed frequencies are zero, a small constant should be added (say 10^{-6}) for calculating the initial estimate $\boldsymbol{\beta}^{(0)}$. At iteration k, $\mathit{step}^{(k)}$ should be chosen such that the value of the likelihood function (2.17) increases. An appropriate method is to start with $\mathit{step}^{(k)} = 1$, and, if necessary, keep halving its value until the likelihood function evaluated at $\mathbf{v}(\boldsymbol{\beta}^{(k)}, \mathit{step}^{(k)})$ is greater than at $\boldsymbol{\beta}^{(k)}$. Various other methods for choosing a step size can be used. See, e.g., Dennis and Schnabel (1983).

The algorithm can be terminated when the likelihood equation $\mathbf{X}'\mathbf{m} = \mathbf{X}'\mathbf{n}$ is close to being satisfied. As noted in the previous section on IPF, it is not always appropriate to stop the iterative process when there is little change in the successive values of the log likelihood function. That criterion does not guarantee that the algorithm is close to convergence, since, even if the likelihood changes little per iteration, there can still be a large change over many iterations. A different criterion, is based on an appropriate measure for the distance of $\mathbf{m}$ from satisfying the likelihood equation $\mathbf{X}'\mathbf{m} = \mathbf{X}'\mathbf{n}$. The following quadratic form can be used:

$$e(\mathbf{m}) = (\mathbf{n} - \mathbf{m})'\mathbf{X}(\mathbf{X}'\mathbf{D}_\mathbf{m}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{n} - \mathbf{m}).$$

If, at iteration k, $e(\mathbf{m}^{(k)})$ is small enough, say less than 10^{-10}, the iterative process can be terminated. Generally, not many iterations are needed before this criterion is satisfied. In fact, the speed of convergence of N-R is second-order, meaning that the convergence satisfies

$$|\hat{\beta}_i^{(k+1)} - \hat{\beta}_i| \leq c\,|\hat{\beta}_i^{(k)} - \hat{\beta}_i|^2 \quad \text{for some } c > 0.$$

This implies that the number of iterations needed to double the number of accurate digits is bounded. Thus, if the dth digit has been calculated, at most a certain number of iterations, say u, are needed to calculate the next d digits. However, u is unknown and is possibly large.

The first estimate $\boldsymbol{\beta}^{(0)}$, as given by (2.23), is the weighted least squares (WLS) estimate of $\boldsymbol{\beta}$ (Grizzle et al., 1969), which has been used instead of the MLE. The advantage of WLS is that computing time is saved. However, WLS parameter estimates are very sensitive to sparse data. For loglinear models, it seems that WLS is highly inferior to maximum likelihood, and is not to be recommended, except when every cell in the table has large counts, say at least 5 to 10. In such a case, maximum likelihood and WLS give similar results. In fact, N-R can be shown to consist of a sequence of WLS estimates, and for this reason it is sometimes also called *iteratively reweighted least squares* (IRLS). For a further discussion of the relationship between IRLS and ML estimation, see Green (1984), Jennrich and Moore (1975), and Jørgenson (1984).

2.4.4 Comparison of Newton-Raphson and iterative proportional fitting

In general, IPF is faster per iteration than N-R, because with the latter, a matrix is calculated and inverted for every iteration. For decomposable models, IPF uses only one iteration and is therefore recommended. For non-decomposable models, N-R generally needs fewer iterations than IPF. Unfortunately, no general recommendation can be given on which method to use. The relative speeds also depend strongly on how efficiently the algorithms are programmed.

2.5 Assessing model goodness-of-fit

2.5.1 Chi-squared tests

The goodness-of-fit of a postulated loglinear model can be assessed by comparing the observed frequencies, $\mathbf{n}$, with estimated expected frequencies, $\hat{\mathbf{m}}$. For loglinear models, two test statistics are commonly used, namely, the likelihood ratio test

$$G^2 = 2\sum_i n_i \log \frac{n_i}{\hat{m}_i}$$

and Pearson's chi-squared test

$$X^2 = \sum_i \frac{(n_i - \hat{m}_i)^2}{\hat{m}_i}.$$

If the postulated model is true, these test statistics have an asymptotic chi-squared distribution. Using the freedom equation representation of a loglinear model, the number of degrees of freedom (*df*) is equal to the number of cells minus the number of linearly independent columns of the design matrix $\mathbf{X}$. When using constraint equations, *df* is equal to the number of functionally independent constraints, or the rank of matrix $\mathbf{C}$ in (2.11).

One problem is that a sufficiently large number of observations is needed in order to obtain a good approximation to the chi-squared distribution. Larntz (1978), Koehler and Larntz (1980), and Koehler (1986), showed that X^2 can be used with smaller sample sizes and sparser tables than G^2. When n/r, with r the number of cells, is less than 5, they showed that G^2 gives a bad approximation to the chi-squared distribution. Though it is difficult to give general rules, Agresti and Yang (1987) gave simulation results where X^2 tends to do well when n/r exceeds 1.

The Wald statistic (Wald, 1943) can be used for testing goodness-of-fit even without estimated frequencies. This test is also asymptotically chi-squared distributed when the null hypothesis is true, but, in general, converges more slowly to this distribution. To test a hypothesis $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, the Wald statistic is

$$W^2 = \mathbf{h}(\mathbf{n})'\tilde{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{n}))^{-1}\mathbf{h}(\mathbf{n}),$$

where $\tilde{\mathbf{\Sigma}}(\mathbf{h}(\mathbf{n}))$ is the sample value of the estimated covariance matrix of $\mathbf{h}(\mathbf{n})$. When applying this test to the loglinear model, the constraint equation formulation (2.11) should be used. With $\mathbf{h} = \mathbf{C}' \log \mathbf{m}$,

$$W^2 = (\log \mathbf{n})' \mathbf{C} (\mathbf{C}' \mathbf{D}_{\mathbf{n}}^{-1} \mathbf{C})^{-1} \mathbf{C}' (\log \mathbf{n})$$

can be derived. Using result 2 in Appendix A.3, this expression can be rewritten using the design matrix $\mathbf{X}$ instead of $\mathbf{C}$ as

$$W^2 = (\log \mathbf{n})' (\mathbf{D}_{\mathbf{n}} - \mathbf{X} (\mathbf{X}' \mathbf{D}_{\mathbf{n}} \mathbf{X})^{-1} \mathbf{X}') (\log \mathbf{n}).$$

One problem with using W^2 for testing loglinear models is that it is very sensitive to zero observed cells. For this reason, W^2 is not recommended unless every cell has a large number of observations. In such cases W^2, G^2, and X^2 tend to have similar values if the model is true.

For testing independence in a 2×2 table with observed frequencies (n_1, n_2, n_3, n_4), W^2 is

$$W^2 = \left(\frac{\text{sample log odds ratio}}{\text{sample variance}} \right)^2 = n_+ \frac{\left(\log \frac{n_1 n_4}{n_2 n_3} \right)^2}{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}$$

with df=1. If one of the observed frequencies approaches zero, the statistic also approaches zero. Thus, in such a case it cannot be used.

2.5.2 Exact tests

In the previous section, large sample "chi-square" statistics were described for testing goodness-of-fit. As the sample size increases, these statistics have a distribution that is more nearly chi-square. If the sample size is small, hypotheses can be tested using *exact* distributions rather than large sample approximations.

A well-known test using an exact distribution is Fisher's exact test for independence in a 2×2 table (Fisher, 1934). By conditioning on the observed marginal distribution, the frequencies follow a hypergeometric distribution. The exact p-value is defined as the sum of hypergeometric probabilities for outcomes at least as favourable to the alternative hypothesis as the observed outcome. For a more extensive description of Fisher's exact test, see, for example, Agresti (1990, p. 59–66). A drawback of the test, and of exact conditional tests for loglinear models in general, is that

they tend to be conservative and therefore are not powerful (Berkson, 1978).

Testing loglinear hypotheses for larger tables can be done by conditioning on the sufficient statistics for the model, so that a parameter-free distribution is obtained. The distance of the observed frequencies, $\mathbf{n}$, from H_0 should be measured using some meaningful statistic, such as G^2 or X^2. The exact p-value is then defined as the probability of the set of tables for which the test statistic is at least as great as the observed one, where the probability is calculated using the exact conditional distribution rather than the large-sample chi-square distribution. For a comprehensive review of exact tests for contingency tables, with comments by various authors, see Agresti (1992). One problem with exact tests is that the evaluation of p-values can be a computationally expensive procedure. This is especially problematic for large tables.

2.5.3 Conditional tests

The chi-squared and exact tests described in the previous two sections are in fact conditional tests against the alternative hypothesis that the saturated model holds. More generally, a conditional test of a model $[\omega_2]$ against the alternative model $[\omega_1]$ can be performed when $[\omega_2]$ strictly implies $[\omega_1]$. It is said that such models are nested. The conditional likelihood-ratio test can be carried out by subtracting the value of G^2 for $[\omega_1]$ from the value of G^2 for $[\omega_2]$. With $\hat{m}_{1i}$ and $\hat{m}_{2i}$ the MLEs for models $[\omega_1]$ and $[\omega_2]$ respectively, and with $[\omega_0]$ the saturated model, the following property holds:

$$G^2(\omega_2|\omega_1) = G^2(\omega_2|\omega_0) - G^2(\omega_1|\omega_0) = 2\sum \hat{m}_{1i} \log(\hat{m}_{1i}/\hat{m}_{2i}) \quad (2.24)$$

(Simon, 1973). If $[\omega_2]$ is true, this test statistic has a chi-squared distribution, with *df* equal to *df* of the first test minus *df* for the second test. For X^2, the difference between the statistics for the separate models is not of Pearson form and not even necessarily nonnegative. A more appropriate definition of the conditional Pearson statistic may be

$$X^2(\omega_2|\omega_1) = \sum \frac{(\hat{m}_{1i} - \hat{m}_{2i})^2}{\hat{m}_{1i}}. \quad (2.25)$$

Both $G^2(\omega_2|\omega_1)$ and $X^2(\omega_2|\omega_1)$ depend on the data only through the sufficient statistics for $[\omega_1]$. Under the hypothesis that $[\omega_1]$ holds, both

statistics have identical large-sample behaviour, and tend to have similar values, even for fairly sparse tables (Haberman, 1977).

A test against an unsaturated alternative can have several advantages. First, provided the alternative is true, such a test is more powerful because it is based on fewer degrees of freedom (Agresti, 1990, p. 97–100). Second, the test depends only on the data through the sufficient statistics for the alternative hypothesis, and will therefore be more nearly chi-squared distributed than a test against the saturated model if the hypothesis is true.

2.5.4 Analysis of residuals

In the previous two sections, goodness-of-fit statistics have been described which can be used to ascertain whether a given model fits the data. If the model does not fit well, insight can be gained in the reasons for this by analyzing cell residuals, which are measures for the deviation of observed from fitted cell values.

For cell i, the raw residual $n_i - \hat{m}_i$ depends strongly on the size of $\hat{m}_i$, and is therefore of limited use. A measure which adjusts for the size of $\hat{m}_i$ is the *standardized residual*, which is defined as

$$e_i = \frac{n_i - \hat{m}_i}{\hat{m}_i^{1/2}}.$$

The e_i are related to the Pearson statistic by $\sum e_i^2 = X^2$. As a measure of the deviation of a fitted value for model $[\omega_2]$ to a fitted value for a simpler model $[\omega_1]$ (i.e., $[\omega_2]$ implies $[\omega_1]$), one can define a conditional standardized residual. With $\hat{m}_{1i}$ and $\hat{m}_{2i}$ the fitted values for models $[\omega_1]$ and $[\omega_2]$ respectively, the following definition can be used:

$$e_i(\omega_2|\omega_1) = \frac{\hat{m}_{1i} - \hat{m}_{2i}}{\hat{m}_{2i}^{1/2}}. \tag{2.26}$$

The conditional residuals are related to the conditional X^2 statistic (2.25) by $\sum e_i(\omega_2|\omega_1)^2 = X^2(\omega_2|\omega_1)$.

One drawback of standardized residuals is that their variance is less than 1, so that a comparison with the standard normal distribution is not appropriate (Sobel, 1995, p. 298). The *adjusted residual* is defined as the raw residual $n_i - \hat{m}_i$ divided by its standard error (Haberman, 1974).

As its mean is 0 and variance is 1, it is more appropriate for comparison with the standard normal than the standardized residual. Denoting the adjusted residuals by r_i, the definition is

$$r_i = \frac{n_i - \hat{m}_i}{\sigma(n_i - \hat{m}_i)}. \tag{2.27}$$

Analogous to (2.26), conditional adjusted residuals can be defined as

$$r_i(\omega_2|\omega_1) = \frac{\hat{m}_{1i} - \hat{m}_{2i}}{\sigma(\hat{m}_{1i} - \hat{m}_{2i})}.$$

Formulae for the variances of the raw residuals will be given in section 2.6.2. For further material on residuals and their relative performance, see Pierce and Schafer (1986).

2.6 Asymptotic behaviour of MLEs

The asymptotic distribution of MLEs of several statistics, in particular the cell frequencies, the loglinear parameters, and cell residuals will be described, given one of the sampling distributions described in Section 2.1, and assuming that a certain loglinear model is true.

All MLEs that are described have an asymptotic normal distribution. First, the delta method, which can be used to derive the covariance matrix of a function of a normally distributed estimator, is explained. Then, expressions are derived for the asymptotic covariance matrices of MLEs. Finally, in subsection 2.6.3, it is shown how the average precision of estimators relates to the number of cells and the number of parameters of the model.

2.6.1 The delta method

If an estimator $\hat{\boldsymbol{\theta}}$ has an asymptotic multivariate normal distribution with expected value $\boldsymbol{\theta}$, then a differentiable function $\mathbf{g}(\hat{\boldsymbol{\theta}})$ of $\hat{\boldsymbol{\theta}}$ also has an asymptotic multivariate normal distribution, with expected value $\mathbf{g}(\boldsymbol{\theta})$. The delta method can be used to derive the covariance matrix of $\mathbf{g}(\hat{\boldsymbol{\theta}})$. This is done as follows. Suppose the covariance matrix of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{\Sigma}$. Let $\mathbf{g}(\boldsymbol{\theta})$ be a differentiable function of $\boldsymbol{\theta}$, and let

$$\mathbf{G} = \frac{\partial \mathbf{g}'}{\partial \boldsymbol{\theta}}.$$

The asymptotic covariance matrix of $\mathbf{g}(\boldsymbol{\theta})$ is

$$\boldsymbol{\Sigma}(\mathbf{g}(\hat{\boldsymbol{\theta}})) = \mathbf{G}'\boldsymbol{\Sigma}\mathbf{G}.$$

For further details on the delta method, see also Bishop, Fienberg, and Holland (1975, Chapter 14), or Agresti (1990, Chapter 12).

2.6.2 The asymptotic distribution of MLEs

The results in Appendix A are used to derive the asymptotic distribution of $\log \hat{\mathbf{m}}$, given that a sampling distribution described in Section 2.1 is used, and assuming that a certain loglinear model is true. Then, using the delta method, the asymptotic distribution of functions of $\log \hat{\mathbf{m}}$ is derived. Finally, the asymptotic distribution of residuals is presented. Generalizations of known results are given. In particular, the asymptotic distribution of MLEs given a broader class of sampling distributions than previously considered is derived, and the asymptotic distribution of conditional residuals is given.

The distribution of MLEs of the log expected frequencies

Let $\hat{\mathbf{m}}_s$ be the MLE of $\mathbf{m}$ given a certain sampling scheme $\mathcal{S}$ described in Section 2.1, let $\hat{\mathbf{m}}_m$ be the MLE of $\mathbf{m}$ given a certain loglinear model $\mathcal{M}$ assuming Poisson sampling, and let $\hat{\mathbf{m}}_{s+m}$ be the MLE of $\mathbf{m}$ given both sampling scheme $\mathcal{S}$ and loglinear model $\mathcal{M}$. Below, the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_s$ is derived first, then of $\log \hat{\mathbf{m}}_m$. Formula (A.18) is then used to derive the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_{s+m}$.

For a Poisson sampling distribution, the kernel of the log likelihood function is $\mathcal{L} = \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m}$. With $\boldsymbol{\theta} = \log \mathbf{m}$, let

$$\mathbf{B} = E\left(-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = \mathbf{D}_{\mathbf{m}}.$$

The covariance matrix of the MLE of $\boldsymbol{\theta}$ given Poisson sampling is $\mathbf{B}^{-1} = \mathbf{D}_{\mathbf{m}}^{-1}$. Consider the sampling restriction $\mathbf{h}_1 = \mathbf{W}'\mathbf{m} - \mathbf{W}'\mathbf{n} = \mathbf{0}$. Differentiating $\mathbf{h}_1$ with respect to $\boldsymbol{\theta}$ yields

$$\mathbf{H}_1 = \frac{\partial \mathbf{h}_1'}{\partial \boldsymbol{\theta}} = \mathbf{D}_{\mathbf{m}}\mathbf{W}.$$

Let $\hat{\boldsymbol{\theta}}_1$ be the MLE of $\boldsymbol{\theta}$ given sampling constraint $\mathbf{h}_1 = \mathbf{0}$. Using (A.5), the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_1 = \hat{\boldsymbol{\theta}}_1$ is

$$\begin{aligned} \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_1) &= \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}_1(\mathbf{H}_1'\mathbf{B}^{-1}\mathbf{H}_1)^{-1}\mathbf{H}_1'\mathbf{B}^{-1} \\ &= \mathbf{D}_{\mathbf{m}}^{-1} - \mathbf{W}(\mathbf{W}'\mathbf{D}_{\mathbf{m}}\mathbf{W})^{-1}\mathbf{W}'. \end{aligned}$$

It follows that, for any sampling scheme $\mathcal{S}$ presented in Section 2.1 (i.e., $\mathcal{S}$ is either Poisson sampling or Poisson sampling with constraint $\mathbf{h}_1 = \mathbf{0}$, the covariance matrix of the MLE $\log \hat{\mathbf{m}}_s$ given $\mathcal{S}$ is

$$\boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_s) = \mathbf{D}_{\mathbf{m}}^{-1} - \boldsymbol{\Lambda}(\mathcal{S}),$$

where $\boldsymbol{\Lambda}(\mathcal{S})$ is a nonnegative definite matrix which depends on $\mathcal{S}$, and which is defined as

$$\boldsymbol{\Lambda}(\mathcal{S}) = \begin{cases} \mathbf{0} & \mathcal{S} : \text{Poisson sampling} \\ \mathbf{W}(\mathbf{W}'\mathbf{D}_{\mathbf{m}}\mathbf{W})^{-1}\mathbf{W}' & \mathcal{S} : \mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n} \end{cases} . \tag{2.28}$$

In the case of multinomial sampling, $\mathbf{W} = \mathbf{1}$, and $\boldsymbol{\Lambda}(\mathcal{S})$ reduces to

$$\boldsymbol{\Lambda}(\mathcal{S}) = \frac{1}{n_+}\mathbf{1}\mathbf{1}',$$

with n_+ the sample size.

Next, consider the loglinear model defined by the constraint

$$\mathbf{h}_m = \mathbf{C}' \log \mathbf{m} = \mathbf{0}.$$

Note that the loglinear model is written using constraint equations instead of the more usual freedom equations. This allows formula (A.5) to be used. Let $\hat{\mathbf{m}}_m$ be the MLE of $\mathbf{m}$ given the constraint $\mathbf{h}_m = \mathbf{0}$. With

$$\mathbf{H}_m = \frac{\partial \mathbf{h}_m'}{\partial \boldsymbol{\theta}} = \mathbf{C},$$

the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_m$ is

$$\begin{aligned} \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_m) &= \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}_m(\mathbf{H}_m'\mathbf{B}^{-1}\mathbf{H}_m)^{-1}\mathbf{H}_m'\mathbf{B}^{-1} \\ &= \mathbf{D}_{\mathbf{m}}^{-1} - \mathbf{D}_{\mathbf{m}}^{-1}\mathbf{C}(\mathbf{C}'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{D}_{\mathbf{m}}^{-1}. \end{aligned} \tag{2.29}$$

Let $\mathbf{X}$ be the orthogonal complement of $\mathbf{C}$, i.e., $\mathbf{X}$ is the design matrix of the loglinear model. Using Result 2, formula 2.29 can be rewritten in terms of $\mathbf{X}$:

$$\boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_m) = \mathbf{X}(\mathbf{X}'\mathbf{D}_{\mathbf{m}}\mathbf{X})^{-1}\mathbf{X}'.$$

It follows from assumption (2.12) that $\mathbf{H}_m'\mathbf{B}^{-1}\mathbf{H}_1 = \mathbf{C}'\mathbf{W} = \mathbf{0}$. Thus, from definition (A.17), $\mathbf{C}'\log\mathbf{m}$ and $\mathbf{W}'\mathbf{m}$ are orthogonal parameters. Therefore, equation (A.18) can be used to find the covariance matrix of $\log\hat{\mathbf{m}}_{s+m}$, the MLE of $\log\mathbf{m}$ given sampling scheme $\mathcal{S}$ and the constraint $\mathbf{h}_m(\mathbf{m}) = \mathbf{0}$. One obtains

$$\begin{aligned}\boldsymbol{\Sigma}(\log\hat{\mathbf{m}}_{s+m}) &= \boldsymbol{\Sigma}(\log\hat{\mathbf{m}}_s) + \boldsymbol{\Sigma}(\log\hat{\mathbf{m}}_m) - \boldsymbol{\Sigma}(\log\hat{\mathbf{m}}_p) \\ &= \mathbf{X}(\mathbf{X}'\mathbf{D}_\mathbf{m}\mathbf{X})^{-1}\mathbf{X}' - \boldsymbol{\Lambda}(\mathcal{S}), \qquad (2.30)\end{aligned}$$

where $\hat{\mathbf{m}}_p$ is the MLE of $\mathbf{m}$ given Poisson sampling, and $\boldsymbol{\Lambda}(\mathcal{S})$ is given by (2.28). This generalizes a result by Lang (1996b) to a broader class of sampling schemes.

The distribution of other parameters

In the previous section, the asymptotic covariance matrix of $\log\hat{\mathbf{m}}_{s+m}$ was obtained using results from Appendix A. Here, $\hat{\mathbf{m}}$ is written short for $\hat{\mathbf{m}}_{s+m}$. Using the delta method, the covariance matrices of the estimators $\hat{\mathbf{m}}$, $\mathbf{X}'\hat{\mathbf{m}}$, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\log\hat{\mathbf{m}}$, which are functions of $\log\hat{\mathbf{m}}$, are derived. The derivative matrices that are needed are

$$\frac{\partial\hat{\mathbf{m}}}{\partial\boldsymbol{\theta}'} = \mathbf{D}_\mathbf{m} \qquad \frac{\partial\mathbf{X}'\hat{\mathbf{m}}}{\partial\mathbf{m}'} = \mathbf{X}' \qquad \frac{\partial\hat{\boldsymbol{\beta}}}{\partial\boldsymbol{\theta}'} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Using (2.30), the delta method yields

$$\begin{aligned}\boldsymbol{\Sigma}(\hat{\mathbf{m}}) &= \mathbf{D}_\mathbf{m}\boldsymbol{\Sigma}(\log\hat{\mathbf{m}})\mathbf{D}_\mathbf{m} \\ \boldsymbol{\Sigma}(\mathbf{X}'\hat{\mathbf{m}}) &= \mathbf{X}'\boldsymbol{\Sigma}(\hat{\mathbf{m}})\mathbf{X}' \\ \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}(\log\hat{\mathbf{m}})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

A simplification is possible:

$$\begin{aligned}\boldsymbol{\Sigma}(\mathbf{X}'\hat{\mathbf{m}}) &= \mathbf{X}'\,\boldsymbol{\Sigma}(\hat{\mathbf{m}})\,\mathbf{X} \\ &= \mathbf{X}'\mathbf{D}_\mathbf{m}\mathbf{X} - \mathbf{X}'\mathbf{D}_\mathbf{m}\boldsymbol{\Lambda}(\mathcal{S})\mathbf{D}_\mathbf{m}\mathbf{X} \\ &= \boldsymbol{\Sigma}(\mathbf{X}'\mathbf{n}).\end{aligned}$$

The asymptotic covariance matrix of $\mathbf{X}'\hat{\mathbf{m}}$, the estimated value of the sufficient statistics is equal to the covariance matrix of the observed sufficient statistics. The MLEs of the sufficient statistics are identical to their observed value. Thus, asymptotically, the distribution of the sufficient statistics is independent of the validity of the loglinear model.

The distribution of residuals

In order to calculate the adjusted residuals, the asymptotic variances of the raw residuals $\mathbf{n} - \hat{\mathbf{m}}$ are needed. Equation (A.3), with the loglinear constraints $\mathbf{C}'\boldsymbol{\theta} = \mathbf{0}$, reduces to

$$\mathbf{D}_{\mathbf{m}}^{-1}(\mathbf{n} - \mathbf{m}) + \mathbf{D}_{\mathbf{m}}^{-1}\mathbf{C}\boldsymbol{\lambda} = \mathbf{0}.$$

The MLEs $\hat{\mathbf{m}}$ and $\hat{\boldsymbol{\lambda}}$ are a solution to this equation, so

$$\mathbf{n} - \hat{\mathbf{m}} = -\mathbf{C}\hat{\boldsymbol{\lambda}}.$$

The raw residuals are clearly a function of $\hat{\boldsymbol{\lambda}}$, and, using (A.4), their covariance matrix can be calculated using the delta method. This yields

$$\boldsymbol{\Sigma}(\mathbf{n} - \hat{\mathbf{m}}) = \mathbf{C}(\mathbf{C}'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{C})^{-1}\mathbf{C}'.$$

Using result 2, this can be restated in terms of the design matrix $\mathbf{X}$ as

$$\boldsymbol{\Sigma}(\mathbf{n} - \hat{\mathbf{m}}) = \mathbf{D}_{\mathbf{m}} - \mathbf{D}_{\mathbf{m}}\mathbf{X}(\mathbf{X}'\mathbf{D}_{\mathbf{m}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_{\mathbf{m}}.$$

The square root of the ith diagonal element yields the variance of the residual $n_i - \hat{m}_i$, which can be substituted into (2.27).

For the conditional residuals (2.28), the variances of the elements of $\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2$ are needed, where $\hat{\mathbf{m}}_1$ is the MLE of $\mathbf{m}$ given a model $[\omega_1]$, and $\hat{\mathbf{m}}_2$ is the MLE of $\mathbf{m}$ given a simpler model $[\omega_2]$. Since both estimators are functions of the observed frequencies $\mathbf{n}$, the delta method can be used to obtain the covariance matrix of the residuals. One finds

$$\boldsymbol{\Sigma}(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2) = \mathbf{D}_{\mathbf{m}}\mathbf{X}_1(\mathbf{X}_1'\mathbf{D}_{\mathbf{m}}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{D}_{\mathbf{m}} - \mathbf{D}_{\mathbf{m}}\mathbf{X}_2(\mathbf{X}_2'\mathbf{D}_{\mathbf{m}}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{D}_{\mathbf{m}}.$$

Note that the distribution of the residuals does not depend on the sampling scheme used.

2.6.3 The average precision of MLEs

It is well known that the precision of MLEs is proportional to the sample size, in the sense that, asymptotically, if the sample size is doubled, the variance of MLEs halves. It is also well known that the more parsimonious a model, the smaller the variance of MLEs if the model is true (Altham,

1984). Below, a new result is presented, which shows that the average precision of $\hat{\boldsymbol{\theta}} = \log \hat{\mathbf{m}}$ is, in fact, proportional to the ratio of the sample size and the number of free parameters on which the distribution depends.

Let the probability of an observation falling in cell i be π_i and denote the average variance of the elements of $\hat{\boldsymbol{\theta}}$ as $\bar{\sigma}^2(\hat{\boldsymbol{\theta}})$. The average variance can be defined by using the definition

$$\bar{\sigma}^2(\hat{\boldsymbol{\theta}}) = \sum \pi_i \, \sigma^2(\hat{\theta}_i).$$

Let f equal the number of free parameters of the distribution, defined as the number of identified loglinear parameters minus the number of sampling constraints. As will be demonstrated below, $\bar{\sigma}^2(\hat{\boldsymbol{\theta}})$ reduces to the very simple formula

$$\bar{\sigma}^2(\hat{\boldsymbol{\theta}}) = \frac{f}{n}. \tag{2.31}$$

As a result, if the number of parameters is doubled, the number of observations should also be doubled to keep the same precision.

The result (2.31) can be proven as follows. With the trace of a matrix defined as the sum of its diagonal elements, observe that, for any loglinear model,

$$\begin{aligned} \bar{\sigma}^2(\hat{\boldsymbol{\theta}}) &= \text{trace}\left(\mathbf{D}_{\boldsymbol{\pi}} \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})\right) \\ &= \text{trace}\left(\mathbf{D}_{\hat{\boldsymbol{\pi}}} \mathbf{X} (\mathbf{X}' \mathbf{D}_{\mathbf{m}} \mathbf{X})^{-1} \mathbf{X}'\right) \\ &= n^{-1} \times \text{trace}\left(\mathbf{D}_{\mathbf{m}} \mathbf{X} (\mathbf{X}' \mathbf{D}_{\mathbf{m}} \mathbf{X})^{-1} \mathbf{X}'\right). \end{aligned} \tag{2.32}$$

A result from matrix algebra is that trace($\mathbf{QR}$)=trace($\mathbf{RQ}$), for arbitrary matrices $\mathbf{R}$ and $\mathbf{Q}$ for which $\mathbf{RQ}$ is square (Searle, 1982, p. 45). With $\mathbf{I}_f$ the $f \times f$ identity matrix, moving the leftmost $\mathbf{X}'$ in (2.32) to the left of the expression in brackets yields

$$\begin{aligned} \bar{\sigma}^2(\hat{\boldsymbol{\theta}}) &= n^{-1} \times \text{trace}\left(\mathbf{X}' \mathbf{D}_{\mathbf{m}} \mathbf{X} (\mathbf{X}' \mathbf{D}_{\mathbf{m}} \mathbf{X})^{-1}\right) \\ &= n^{-1} \times \text{trace}(\mathbf{I}_f) \\ &= \frac{f}{n}. \end{aligned}$$

It will be shown in Chapter 4 that the result holds for a much more general class of models.

Result (2.31) is important as it shows the effect of parsimonious modelling. When there are many variables, the number of parameters can be greatly reduced by deleting higher-order interaction parameters, and thereby the precision of estimators is improved. For instance, when there are ten variables with four categories each, the saturated model has more than a million parameters. Unless one has literally millions of observations, the observed log cell frequencies will have such large variances that they are meaningless as estimates of the true expected log cell frequencies. However, the model of complete independence has only 31 parameters, and only, say, several hundred observations are necessary to get reasonable estimates of the log cell frequencies. Thus, if the independence model is true, the average precision of the MLEs $\log \hat{m}_i$ given independence is more than thirty thousand times greater than the average precision of the estimated log frequencies given the saturated model. Of course, because of the inevitable sparseness of such a table, it will be difficult to test whether independence really holds.

Chapter 3

Marginal homogeneity models linear in the expected frequencies

Marginal homogeneity models, which are used to model homogeneity of correlated discrete distributions, and other models which are linear in the expected frequencies are described in this chapter. These models are generally not loglinear in terms of the expected joint frequencies, and, therefore, methods other than the ones used in the previous chapter must be used for estimation and testing. The purpose of this chapter is to provide an overview of the most important literature on the marginal homogeneity model.

In the introductory section 3.1, a short explanation of marginal homogeneity is given, and the notation is presented. Subsequently, maximum likelihood estimation is described in Section 3.2. The existence and uniqueness of maximum likelihood estimates is proven, and several estimation algorithms are presented. Two alternatives to maximum likelihood are considered in Section 3.3, namely, the minimum discrimination estimation method, and minimum modified chi-squared estimation. Testing goodness-of-fit and the asymptotic distribution of MLEs of the expected joint frequencies is described in Section 3.4. Finally, the methods are illustrated with an example.

	Grade of left eye				
	Highest	Second	Third	Lowest	*Total*
Grade of right eye					
Highest	1520	266	124	66	1976
Second	234	1512	432	78	2256
Third	117	362	1772	205	2456
Lowest	36	82	179	492	789
Total	1907	2222	2507	841	7477

Table 3.1: *Unaided distance vision for women (Stuart, 1955)*

3.1 Introduction and notation

Consider the data in Table 3.1. For a group of 7477 women, the quality of the left and right eyes were classified into four categories. This is a classic example, originally analyzed by Stuart (1955). He was interested in testing the hypothesis that, in the population, the quality of the left and right eye are the same. In other words, the question is whether the marginal distributions of Table 3.1 are identical. The corresponding model is referred to as the marginal homogeneity (MH) model. Denoting the expected count of subjects with category i for the right eye and category j for the left eye as m_{ij}, the hypothesis of MH can be written using the equation

$$m_{i+} = m_{+i} \qquad \forall i, \tag{3.1}$$

where the "+" denotes summation over the appropriate subscript.

The MH model for two dimensions, as defined by constraint 3.1, can be generalized to MH for higher dimensional tables. For three variables, e.g., three points in time, MH of the one-dimensional marginals can be represented using the formula

$$m_{i++} = m_{+i+} = m_{++i} \qquad \forall i.$$

Two alternative types of MH for three-way tables can be represented using constraints such as

$$\begin{aligned} m_{ij+} &= m_{+ij} \qquad \forall i, j \\ m_{ij+} = m_{+ij} &= m_{i+j} \qquad \forall i, j. \end{aligned}$$

The first model may be applied, for instance, in panel studies if one is interested in knowing whether change is constant, i.e., if the turnover from time 1 to time 2 is the same as the turnover from time 2 to time 3. Many other variations of the MH model are possible and generalization to higher-dimensional tables is straightforward.

In general, MH models are not loglinear, but linear in the expected joint frequencies. Only for the 2×2 table is the MH model loglinear. In this case, MH is identical to symmetry which can be characterized by the single constraint $m_{12} = m_{21}$. In loglinear form, this is written as $\log m_{12} = \log m_{21}$.

It should be noted that all MH models considered above are defined using linear constraints on the expected frequencies. More generally, for t expected cell frequencies $m_1, \ldots, m_t$, a set of c linear constraints can be written as

$$\sum_{i=1}^{t} c_{ij} m_i = d_j \qquad \forall j, \tag{3.2}$$

where c_{ij} and $d_j \geq 0$ are constants. Let $\mathbf{C}$ be the matrix with elements c_{ij}, $\mathbf{d}$ the vector with elements d_j, and $\mathbf{m}$ the vector with elements m_i. Then the linear constraints (3.2) can be written in matrix notation as

$$\mathbf{C}'\mathbf{m} = \mathbf{d}. \tag{3.3}$$

To illustrate the matrix notation, consider MH for a 3×3 table. The constraints can be written equivalently using three equations or a single matrix equation as follows:

$$\left.\begin{array}{l} m_{1+} - m_{+1} = 0 \\ m_{2+} - m_{+2} = 0 \\ m_{3+} - m_{+3} = 0 \end{array}\right\} \iff$$

$$\begin{pmatrix} 0 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 1 & 1 & 0 \end{pmatrix} \mathbf{m} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

where $\mathbf{m} = (m_{11}, m_{12}, m_{13}, m_{21}, m_{22}, m_{23}, m_{31}, m_{32}, m_{33})'$. Note that any row in matrix $\mathbf{C}'$ is a linear combination of the other two, so if any one of them is deleted the same model is obtained.

In model equation (3.3), any of the sampling constraints described in Section 2.1, which are linear constraints of the form $\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}$, can

be included. For example, for a single multinomial sample, the sampling constraint is the linear equation $m_+ = n_+$.

The constraints $\mathbf{C}'\mathbf{m} = \mathbf{d}$ are said to be consistent if a distribution $\mathbf{m}$, with $m_i > 0$ for all i, satisfying them exists. A redundant constraint is a linear combination of other constraints. A set of constraints is said to be independent if it contains no redundant constraints. The number of degrees of freedom for a model is equal to the number of independent constraints, not counting the multinomial sampling constraints.

3.2 Maximum likelihood estimation

In this section, the likelihood equations and a proof of existence and uniqueness of maximum likelihood estimates are given. Furthermore, three algorithms for solving the likelihood equations are presented: a maximization method similar to the Newton-Raphson method for log-linear models (see Section 2.4.3), a minimization method, which finds MLEs by reparameterizing a Lagrangian log likelihood in terms of Lagrange multipliers, and a "saddle-point method", which finds a saddle point of the Lagrangian log likelihood by solving the likelihood equations simultaneously for Lagrange multipliers and the expected frequencies.

3.2.1 The likelihood equations

MLEs of the expected frequencies are found by maximizing the kernel of the log likelihood function $\mathcal{L} = \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m}$ in terms of $\mathbf{m}$ subject to the linear constraints (3.3). Sampling constraints, as described in Section 2.1, can be included in the linear model constraints. Using a vector of Lagrange multipliers $\boldsymbol{\lambda}$, the MLE $\hat{\mathbf{m}}$ is a saddle point of the Lagrangian log likelihood function

$$L = \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m} + \boldsymbol{\lambda}'(\mathbf{C}'\mathbf{m} - \mathbf{d}). \tag{3.4}$$

Differentiating L with respect to $\log \mathbf{m}$ and equating the result to zero yields

$$\frac{\partial L}{\partial \log \mathbf{m}} = \mathbf{n} - \mathbf{m} + \mathbf{D}_{\mathbf{m}}\mathbf{C}\boldsymbol{\lambda} = \mathbf{0}, \tag{3.5}$$

where $\mathbf{D}_{\mathbf{m}}$ is the diagonal matrix with the elements of $\mathbf{m}$ on the main diagonal. (Note that differentiating with respect to $\log \mathbf{m}$ and equating

to zero yields the same result as differentiating with respect to $\mathbf{m}$ and equating to zero.) Equation (3.5) and the constraint $\mathbf{C}'\mathbf{m} = \mathbf{d}$ determine the MLE $\hat{\mathbf{m}}$.

Interestingly, from (3.5), the vector $\mathbf{m}$ can be expressed as an explicit function of the Lagrange multiplier vector $\boldsymbol{\lambda}$. One obtains

$$\mathbf{m} = \frac{\mathbf{n}}{\mathbf{1} - \mathbf{C}\boldsymbol{\lambda}}, \tag{3.6}$$

where the division of vectors is done elementwise. Substituting this expression into the constraints (3.3) yields a set of equations in $\boldsymbol{\lambda}$ which determine the MLE $\hat{\boldsymbol{\lambda}}$ independently of $\hat{\mathbf{m}}$.

Following Bennett (1967) and Bishop, Fienberg, and Holland (1975), expression (3.6) will be written out explicitly for the models described in Section 3.1. For square tables with constraints $m_{i+} = m_{+i}$, (3.6) becomes

$$m_{ij} = \frac{n_{ij}}{1 - (\lambda_i - \lambda_j)}. \tag{3.7}$$

For three-way tables, with constraints $m_{i++} = m_{+i+} = m_{++i}$, an expression for the MLEs, with Lagrange multipliers λ_i, μ_i, and ν_i, is

$$m_{ijk} = \frac{n_{ijk}}{1 - (\lambda_i - \lambda_j) - (\mu_j - \mu_k) - (\nu_k - \nu_i)}. \tag{3.8}$$

With constraints $m_{ij+} = m_{+ij}$ and $m_{ij+} = m_{+ij} = m_{i+j}$, the MLEs have forms

$$m_{ijk} = \frac{n_{ijk}}{1 - (\lambda_{ij} - \lambda_{jk})}$$

and

$$m_{ijk} = \frac{n_{ijk}}{1 - (\lambda_{ij} - \lambda_{jk}) - (\mu_{ij} - \mu_{ik})}$$

respectively.

If the model specification is $\mathbf{C}'\mathbf{m} = \mathbf{0}$, where $\mathbf{C}'$ is a contrast matrix, a solution $\hat{\mathbf{m}}$ to equation (3.5) and $\mathbf{C}'\mathbf{m} = \mathbf{0}$ automatically satisfies the multinomial sampling constraint $\mathbf{1}'\mathbf{m} = \mathbf{1}'\mathbf{n}$ (which is identical to $m_+ = n_+$ in scalar notation). This can be seen as follows. Suppose $\mathbf{m}$ satisfies $\mathbf{C}'\mathbf{m} = \mathbf{0}$. Premultiplying (3.5) by $\mathbf{1}'$ yields

$$\mathbf{1}'(\mathbf{n} - \mathbf{m} + \mathbf{D}_{\mathbf{m}}\mathbf{C}\boldsymbol{\lambda}) = \mathbf{1}'(\mathbf{n} - \mathbf{m}) + \mathbf{m}'\mathbf{C}\boldsymbol{\lambda} = \mathbf{1}'(\mathbf{n} - \mathbf{m}) = \mathbf{0}.$$

Thus, the same fitting method can be used for both Poisson and multinomial sampling.

Existence and uniqueness of MLEs

To prove the existence and uniqueness of solutions $\hat{\mathbf{m}}$ and $\hat{\boldsymbol{\lambda}}$ to the linear constraint $\mathbf{C}'\mathbf{m} = \mathbf{d}$ and equation (3.5), it will be assumed that $n_i > 0$ for all i, and that there exists an $\mathbf{m}$ with strictly positive elements satisfying the model.

From (3.6), we define $\mathbf{m}$ as a function of $\boldsymbol{\lambda}$, i.e.,

$$\mathbf{m}(\boldsymbol{\lambda}) = \frac{\mathbf{n}}{\mathbf{1} - \mathbf{C}\boldsymbol{\lambda}}. \tag{3.9}$$

Now $\hat{\boldsymbol{\lambda}}$ is a solution to $\mathbf{C}'\mathbf{m}(\boldsymbol{\lambda}) = \mathbf{d}$, such that $m_i(\hat{\boldsymbol{\lambda}}) > 0$ for all i. Using the Lagrange multiplier theorem (e.g., Bertsekas, 1982, p. 67), such a solution must exist because of the concavity of the log likelihood and the assumed consistency of the constraint $\mathbf{C}'\mathbf{m} = \mathbf{d}$. Below, it is shown that such a solution is unique.

Consider the set

$$\mathcal{C} \;=\; \{\boldsymbol{\lambda} \,|\, \mathbf{1} - \mathbf{C}\boldsymbol{\lambda} > \mathbf{0}\}.$$

It follows that $\hat{\boldsymbol{\lambda}} \in \mathcal{C}$, because otherwise $m_i(\hat{\boldsymbol{\lambda}})$ would be negative or undefined for some i. It is easily verified that $\mathcal{C}$ is a convex set. Now consider L as a function of $\boldsymbol{\lambda}$, i.e., $L = L(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$. Define $\mathbf{q}$ and $\mathbf{Q}$ as the vector of first and the matrix of second derivatives of L with respect to $\boldsymbol{\lambda}$ respectively, i.e.,

$$\begin{aligned} \mathbf{q} = \frac{\partial L}{\partial \boldsymbol{\lambda}} &= \mathbf{C}'\mathbf{m} - \mathbf{d} \\ \mathbf{Q} = \frac{\partial L}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'} &= \mathbf{C}'\mathbf{D}_{\mathbf{n}}^{-1}\mathbf{D}_{\mathbf{m}}^{2}\mathbf{C}, \end{aligned} \tag{3.10}$$

where $\mathbf{m}$ is written short for $\mathbf{m}(\boldsymbol{\lambda})$. Since $\mathbf{Q}$ is positive definite for all $\boldsymbol{\lambda} \in \mathcal{C}$, L is a convex function of $\boldsymbol{\lambda}$ on $\mathcal{C}$. Thus, a stationary point of L is a minimum of L. Because such a stationary point exists and $\mathcal{C}$ is convex, L has a unique minimum $\hat{\boldsymbol{\lambda}}$. Furthermore, $\hat{\mathbf{m}} = \mathbf{m}(\hat{\boldsymbol{\lambda}})$ is uniquely defined and has strictly positive elements.

In practice, there will often be zero observed cells in a contingency table. In general, models linear in the expected frequencies do not have unique MLEs in such cases. For instance, consider the model defined by the constraint $m_1 = m_2 + m_3$, with observations $(n_1, n_2, n_3) = (2, 0, 0)$.

The MLEs are $(\hat{m}_1, \hat{m}_2, \hat{m}_3) = (1, q, 1-q)$ for all $q \in [0, 1]$, which is a solution that cannot be identified. The type of models for which the existence and uniqueness of MLEs is still guaranteed when there are zero observed cells still has to be investigated. In practice, no problems with identifiability seem to occur for the marginal homogeneity models described in Section (3.1), even when there are many observed zeroes.

3.2.2 A minimization method

In the proof of uniqueness of MLEs presented in the previous section, it was shown that $\hat{\boldsymbol{\lambda}}$ is the minimum of $L(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$. This suggests the use of a minimization algorithm to find $\hat{\boldsymbol{\lambda}}$. The Newton-Raphson algorithm presented below was previously described by Madansky (1963), Bennett (1967), Bishop (1975), and Haber (1985), though these authors did not put the algorithm in the context of minimization.

Define the "updating" function

$$\mathbf{v}(\boldsymbol{\lambda}, \mathit{step}) = \boldsymbol{\lambda} - \mathit{step}\, \mathbf{Q}^{-1}(\mathbf{C}'\mathbf{m} - \mathbf{d}),$$

with $\mathbf{Q}$ defined by (3.10) and with

$$\mathbf{m} = \mathbf{D}[\mathbf{1} - \mathbf{C}\boldsymbol{\lambda}]^{-1}\mathbf{n}, \tag{3.11}$$

where $\mathbf{D}[.]$ is the diagonal matrix with the elements of the vector in square brackets on the main diagonal. Then, with appropriate initial estimates $\boldsymbol{\lambda}^{(0)}$, for instance, $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$, the following iterative scheme can be used:

$$\boldsymbol{\lambda}^{(k+1)} \quad = \quad \mathbf{v}(\boldsymbol{\lambda}^{(k)}, \mathit{step}^{(k)}),$$

for $k = 0, 1, \ldots$, and with appropriate values of $\mathit{step}^{(k)}$. The step size $\mathit{step}^{(k)}$ has to be taken small enough so that the log likelihood function $L(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ decreases and successive estimated frequencies remain positive, i.e., so that $\boldsymbol{\lambda}^{(k+1)} \in \mathcal{C}$. Since this is a minimization method on a convex domain, the N-R method is guaranteed to converge if all $n_i > 0$.

Of course, in many practical situations there will be observed zeroes. One problem with the method is that zero observed cells are estimated as zero because of (3.11), while the MLEs are not necessarily zero. In some cases this problem can be overcome by adding a small constant to a zero observed cell (say 10^{-8}), though this may have serious undesirable effects on the estimates $\hat{\mathbf{m}}$. Additionally, adding a constant yields a potentially

serious numerical problem for the method described above. This can be illustrated as follows. Consider the MH model for square tables, with $n_{ij} = 0$ for a certain i and j, and

$$\hat{m}_{ij} \approx \frac{\epsilon}{1 + \hat{\lambda}_i - \hat{\lambda}_j}$$

with ϵ as the added constant. Now suppose that $\epsilon = 10^{-17}$, $\hat{\lambda}_i = 0$, and $\hat{\lambda}_j = 1 - 10^{-17}$, so that $\hat{m}_{ij} \approx 1$. On most computers, a number can only be stored with precision up to a maximum of about 16 digits. Assuming a precision of 16 decimal digits, $\hat{\lambda}_j$ can be stored only as $1.00\ldots00$ or as a $0.99\ldots99$, with just 15 0s or 16 9s after the dot respectively. Thus, using the value $\epsilon = 10^{-17}$ gives a value of $\hat{m}_{ij}$ of either 0 or 10 instead of the desired value 1. In order to avoid numerical problems, therefore, the constant that is added should not be too small relative to the numerical precision of the computer system used.

3.2.3 A maximization method

In the previous section, the fact that the likelihood function can be written as a function of the Lagrange multipliers $\boldsymbol{\lambda}$, and that the MLE $\hat{\boldsymbol{\lambda}}$ is the minimum value of the likelihood as a function of $\boldsymbol{\lambda}$ was presented. An alternative method of finding MLEs is to reparameterize the likelihood in terms of model parameters $\boldsymbol{\beta}$, and then maximize the likelihood in terms of $\boldsymbol{\beta}$. In the literature, two methods have been proposed which use this approach to find MLEs: a method by Gokhale (1973) and the Fisher scoring method (see Agresti, 1990, p. 449–451).

The constraint equation $\mathbf{C}'\mathbf{m} = \mathbf{d}$ can be rewritten using model parameters $\boldsymbol{\beta}$ in the following way:

$$\mathbf{m} = \mathbf{C}^{-}\mathbf{d} + \mathbf{X}\boldsymbol{\beta}, \tag{3.12}$$

where $\mathbf{C}^{-}$ is a generalized inverse of $\mathbf{C}'$ (i.e., $\mathbf{C}'\mathbf{C}^{-} = \mathbf{I}$), $\mathbf{X}$ the orthogonal complement of $\mathbf{C}$ (i.e., $\mathbf{C}'\mathbf{X} = \mathbf{0}$), and $\boldsymbol{\beta}$ a vector of unknown parameters. For example, one can use

$$\begin{aligned} \mathbf{C}^{-} &= \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1} \\ \mathbf{X} &= \mathbf{I} - \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'. \end{aligned} \tag{3.13}$$

The choice of $\mathbf{C}^-$ and $\mathbf{X}$ does not affect the results. Matrix $\mathbf{X}$ is square and not of full column rank. If a matrix of full column rank is needed,

$$\mathbf{X} = (\mathbf{I} - \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}')\mathbf{U} \tag{3.14}$$

can be used, where $\mathbf{U}$ is a random $I \times (I - df)$ matrix. Then, with probability 1, $\mathbf{X}$ is of full column rank. For completeness, it is noted that, for the MH model, Firth (1989) and Lipsitz, Laird, and Harrington (1990) gave specific methods for finding $\mathbf{X}$, without resorting to formulas (3.13) or (3.14). However, it seems simpler to use one of the formulae. For some models it may be awkward to find values for β_k yielding strictly positive values for all m_i.

Both Gokhale's method and Fisher scoring are gradient methods. The search direction used by Gokhale is simply the first derivative vector of the log likelihood function. Fisher scoring is a modification of N-R, and uses minus the inverse of the expected value of the matrix of second derivatives of the log likelihood function as a search direction. Let $\mathcal{L}$ be the kernel of the log likelihood function, i.e.,

$$\mathcal{L} = \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m}.$$

The first and second derivatives of $\mathcal{L}$ with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} &= \mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-1}(\mathbf{n} - \mathbf{m}) \\ \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= -\mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-2}\mathbf{D}_{\mathbf{n}}\mathbf{X}. \end{aligned} \tag{3.15}$$

Every element of the matrix of second derivatives is a linear function of $\mathbf{n}$, which has expected value $\mathbf{m}$. Thus, the expected value of the matrix of second derivatives is

$$E\left(\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = -\mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{X}. \tag{3.16}$$

Both Gokhale's method and Fisher scoring use (3.15) as a search direction for $\boldsymbol{\beta}$. Define the "updating" functions $\mathbf{u}$ and $\mathbf{v}$ as

$$\begin{aligned} \mathbf{u}(\boldsymbol{\beta}, step) &= \boldsymbol{\beta} + step\, \mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-1}(\mathbf{n} - \mathbf{m}) \\ \mathbf{v}(\boldsymbol{\beta}, step) &= \boldsymbol{\beta} + step\, (\mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_{\mathbf{m}}^{-1}(\mathbf{n} - \mathbf{m}) \end{aligned}$$

with $\mathbf{m} = \mathbf{m}(\boldsymbol{\beta}) = \mathbf{C}^{-}\mathbf{d} + \mathbf{X}\boldsymbol{\beta}$ and *step* a positive number chosen such that the likelihood increases, while the estimated frequencies must stay strictly positive. One can start with $step = 1$ and, if the above conditions are not satisfied, keep halving *step* until they are. Gokhale's method and Fisher scoring are, for $k = 0, 1, \ldots$,

$$\begin{aligned} \boldsymbol{\beta}_1^{(k+1)} &= \mathbf{u}(\boldsymbol{\beta}_1^{(k)}, step_1^{(k)}) \\ \boldsymbol{\beta}_2^{(k+1)} &= \mathbf{v}(\boldsymbol{\beta}_2^{(k)}, step_2^{(k)}), \end{aligned}$$

respectively. A starting estimate $\boldsymbol{\beta}_h^{(0)}$ has to be found such that the corresponding frequencies (3.12) are positive.

Some disadvantages of the methods are that care has to be taken that no out-of-range estimates are obtained, and for some models, it may be difficult to find starting values for parameters such that the corresponding cell frequencies are positive. In general, Fisher scoring requires fewer steps to reach satisfactory convergence than Gokhale's method.

For MH for square tables, it can be shown that Gokhale's algorithm can be written more simply as follows. Let $f_{ij}^{(k)} = n_{ij}/m_{ij}^{(k)}$. Then, for $k = 0, 1, \ldots$,

$$m_{ij}^{(k+1)} = m_{ij}^{(k)} + step\left(2I(f_{ij}^{(k)} - 1) + (f_{i+}^{(k)} - f_{+i}^{(k)} - f_{j+}^{(k)} + f_{+j}^{(k)})\right). \quad (3.17)$$

Taking $m_{ij}^{(0)} = 1$ for all i and j as starting values suffices.

3.2.4 A saddle point method

The MLE $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ is a saddle point of the Lagrangian likelihood function (3.4). Below, a method for searching the saddle point is presented. The method has the advantages that no out-of-range iterate estimates are obtained and starting values are not difficult to find. "Saddle-point" methods like the one proposed below were previously presented by Aitchison and Silvey (1958; 1960) and Lang (1996a).

It is proposed that the updating function (A.14) in the appendix be used. With

$$\mathbf{k} = \frac{\partial \mathcal{L}}{\partial \log \mathbf{m}} = \mathbf{n} - \mathbf{m} \qquad \mathbf{B} = E\left(-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = \mathbf{D}_{\mathbf{m}},$$

function (A.14) reduces to

$$\mathbf{v}(\log \mathbf{m}, \mathit{step}) = \log \mathbf{m} + \mathit{step}\, \mathbf{D}_{\mathbf{m}}^{-1}(\mathbf{n} - \mathbf{m} + \mathbf{C}\boldsymbol{\lambda}(\mathbf{m})),$$

where

$$\begin{aligned} \boldsymbol{\lambda}(\mathbf{m}) &= -(\mathbf{C}'\mathbf{D}_{\mathbf{m}}\mathbf{C})^{-1}(\mathbf{C}'(\mathbf{n} - \mathbf{m}) + \mathbf{C}'\mathbf{m} - \mathbf{d}) \\ &= -(\mathbf{C}'\mathbf{D}_{\mathbf{m}}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{n} - \mathbf{d}). \end{aligned}$$

With appropriate starting values, for instance $\mathbf{m}^{(0)} = \mathbf{n}$, the algorithm is

$$\log \mathbf{m}^{(k+1)} = \mathbf{v}(\log \mathbf{m}^{(k)}, \mathit{step}^{(k)}) \qquad k = 0, 1, \ldots.$$

The step size $\mathit{step}^{(k)}$ at iteration k must be chosen such that an appropriate distance function $e(\mathbf{m}^{(k)})$ for measuring distance from convergence decreases, if possible. The function

$$e(\mathbf{m}) = \mathbf{k}'\mathbf{D}_{\mathbf{m}}\mathbf{k}$$

can be used though this gives only a very rough indication of whether one is on the right "track" to convergence. It will not be possible to obtain a decrease of $e(\mathbf{m})$ with every iteration, but over several iterations its value must decrease.

3.2.5 Comparison of the methods

The number of iterations required by the minimization, Fisher scoring, and saddle-point method seems to be approximately the same. For each method, one matrix inversion is necessary, for which the number of rows and columns is equal to the number of degrees of freedom for the model. For many useful models, this number grows slowly with the number of variables, and the inversion can be done very fast. Gokhale's method requires consistently more iterations than the other methods. However, the latter method requires only one matrix inversion (or none if an explicit expression for the matrix to be inverted can be found, as in (3.17) for the MH model). Gokhale's method is recommended only if the matrix inversion is a bottleneck for computations.

A serious disadvantage of the minimization method is that a constant must be added to zero observed cells. This has a potentially large effect on estimated frequencies, and can give rise to serious numerical problems.

Fisher scoring and the saddle-point method behave rather similarly. A disadvantage of Fisher scoring is that it may be difficult to find appropriate starting estimates. A disadvantage of the saddle-point method is that it is difficult to find a theoretically justifiable method for choosing a step size. An advantage is that iterate estimates cannot be out-of-range. The saddle point method will be generalized for use with a much broader class of models in Section 5.1.

3.3 Alternatives to maximum likelihood estimation

In the literature, two alternatives to maximum likelihood for estimating expected cell frequencies given MH have been proposed: minimization of the discrimination information and minimization of Neyman's statistic. An advantage of these estimates is that, in general, they can be calculated using simpler algorithms than those needed for maximum likelihood estimation.

3.3.1 Minimizing the discrimination information

The discrimination information statistic (Kullback, 1959, p. 117–119) is an alternative to the likelihood ratio test or Pearson's chi-squared test for testing goodness-of-fit of a model. Given estimated frequencies m_{ij} for a certain model and observed frequencies n_{ij}, the discrimination information is defined as

$$I = \sum_{i,j} m_{ij} \log \frac{m_{ij}}{n_{ij}}.$$

If the model under consideration is true, I has an asymptotic chi-squared distribution, with degrees of freedom equal to the degrees of freedom for the model. Minimum discrimination information (MDI) estimates for a certain model are those frequencies m_i satisfying the model that minimize I.

For MH for square tables, Ireland, Ku & Kullback (1968) developed an algorithm for calculating minimum discrimination information (MDI) estimates, which is described below. An attractive property of the algorithm is that it is simple and intuitive. In fact, it is strongly related to IPF for loglinear models (Darroch & Ratcliff, 1972).

Consider the marginal homogeneity model defined by the constraints $m_{i+} - m_{+i} = 0$, under multinomial sampling. With Lagrange multipliers λ_i and τ, the MDI estimates can be found as a saddle point of the Lagrange function

$$I_{\boldsymbol{\lambda}} = \sum_{i,j} m_{ij} \log \frac{m_{ij}}{n_{ij}} - \sum_i \lambda_i (m_{i+} - m_{+i}) - \tau(m_{++} - n_{++}).$$

It should be noted that, contrary to ML estimation, the sampling constraint $m_{++} = n_{++}$ is not automatically satisfied, so the term $\tau(m_{++} - n_{++})$ must be included in $I_{\boldsymbol{\lambda}}$. Differentiating $I_{\boldsymbol{\lambda}}$ with respect to m_{ij} and equating the result to zero yields

$$\frac{\partial I_{\boldsymbol{\lambda}}}{\partial m_{ij}} = 1 + \log m_{ij} - \log n_{ij} - (\tau + \lambda_i - \lambda_j) = 0.$$

It follows that

$$m_{ij} = n_{ij} \exp(\tau + \lambda_i - \lambda_j).$$

The Lagrange multipliers can be eliminated by noting that these equations are equivalent to

$$\frac{m_{11} m_{ij}}{m_{i1} m_{1j}} = \frac{n_{11} n_{ij}}{n_{i1} n_{1j}} \qquad \prod_k \frac{m_{ik}}{m_{ki}} = \prod_k \frac{m_{ik}}{m_{ki}},$$

i.e., the observed odds ratios are reproduced. These equations together with the constraints $m_{i+} = m_{+i}$ yield the MDI estimates. The estimates thus have to satisfy a set of linear and multiplicative constraints simultaneously, and the so called generalized iterative scaling algorithm can be used to find them. This yields the iterative scheme:

$$m_{ij}^{(k+1)} := \gamma_k \, m_{ij}^{(k)} \sqrt{\frac{m_{+i}^{(k)} m_{j+}^{(k)}}{m_{i+}^{(k)} m_{+j}^{(k)}}}, \tag{3.18}$$

where γ_k is a normalizing constant which should be chosen such that the estimated frequencies add up to N. As starting values the observed frequencies n_{ij} can be taken. When estimated frequencies are found, they can be substituted in I to obtain an asymptotic chi-squared test. A disadvantage of MDI estimation is that zero observed frequencies are estimated

as zero. However, it is not known whether this affects the chi-squared approximation of I very negatively. An interesting property of the MDI estimates is that $\tilde{m}_{ii} \geq n_{ii}$. Ireland, Ku, and Kullback (1968) described the algorithm and proved convergence basically using the convexity property of the discrimination information and the Cauchy-Schwarz inequality. Darroch and Ratcliff (1972) showed that the algorithm is included in a more general class of algorithms and gave a simplified proof of convergence.

3.3.2 Minimizing Neyman's statistic

In addition to maximum likelihood and minimum discrimination information estimation, a third method of estimating expected frequencies is by minimizing Neyman's statistic (Neyman, 1949). An advantage of the latter method is that closed form solutions for the estimates are obtained. Bhapkar (1966) showed that estimates obtained by minimizing Neyman's statistic are the same as weighted least squares estimates.

Neyman's statistic, also referred to as the modified Pearson's chi-squared (X^2_{mod}) statistic, is

$$X^2_{\text{mod}} = \sum_i \frac{(n_i - m_i)^2}{n_i} = (\mathbf{n} - \mathbf{m})' \mathbf{D}_{\mathbf{n}}^{-1} (\mathbf{n} - \mathbf{m}).$$

Minimizing Neyman's statistic subject to the linear constraints $\mathbf{C}'\mathbf{m} - \mathbf{d}$ can be done using the method of Lagrange multipliers. It should be noted that, if (product) multinomial sampling is used, the sampling constraints must be included in the model constraints. With a vector $\boldsymbol{\lambda}$ of Lagrange multipliers, a saddle point is sought of

$$S = \tfrac{1}{2}(\mathbf{n} - \mathbf{m})' \mathbf{D}_{\mathbf{n}}^{-1} (\mathbf{n} - \mathbf{m}) - \boldsymbol{\lambda}'(\mathbf{C}'\mathbf{m} - \mathbf{d}).$$

(The $\frac{1}{2}$ is introduced because it simplifies calculations later on.) Differentiating S with respect to $\mathbf{m}$ and equating the result to zero gives

$$\mathbf{D}_{\mathbf{n}}^{-1}\mathbf{m} - \mathbf{1} - \mathbf{C}\boldsymbol{\lambda} \;=\; \mathbf{0}. \tag{3.19}$$

From this equation, $\mathbf{m}$ can be written in terms of the unknown $\boldsymbol{\lambda}$:

$$\mathbf{m} = \mathbf{n} + \mathbf{D}_{\mathbf{n}}\mathbf{C}\boldsymbol{\lambda}. \tag{3.20}$$

Substituting this expression into the linear constraints yields the following equation in terms of $\boldsymbol{\lambda}$.

$$\mathbf{C}'\mathbf{n} + \mathbf{C}'\mathbf{D_n}\mathbf{C}\boldsymbol{\lambda} - \mathbf{d} = \mathbf{0}. \tag{3.21}$$

From (3.21), $\boldsymbol{\lambda}$ can be derived:

$$\boldsymbol{\lambda} = -(\mathbf{C}'\mathbf{D_n}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{n} - \mathbf{d}). \tag{3.22}$$

The estimated frequencies will be denoted by $\tilde{\mathbf{m}}$. Substituting (3.22) into (3.20) yields $\tilde{\mathbf{m}}$:

$$\tilde{\mathbf{m}} = \mathbf{n} - \mathbf{D_n}\mathbf{C}(\mathbf{C}'\mathbf{D_n}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{n} - \mathbf{d}). \tag{3.23}$$

There are some drawbacks to these estimates. First, from (3.23) it follows that, if $n_i = 0$, then $\tilde{m}_i = 0$ as well. Second, estimated frequencies may be negative.

Substituting the estimates $\tilde{\mathbf{m}}$ into X^2_{mod} yields the asymptotic chi-squared statistic

$$\begin{aligned} X^2_{\text{mod}} &= (\mathbf{n} - \mathbf{m})'\mathbf{D}_\mathbf{n}^{-1}(\mathbf{n} - \mathbf{m}) \\ &= \boldsymbol{\lambda}'\mathbf{C}'\mathbf{D_n}\mathbf{C}\boldsymbol{\lambda} \\ &= (\mathbf{C}'\mathbf{n} - \mathbf{d})'(\mathbf{C}'\mathbf{D_n}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{n} - \mathbf{d}). \end{aligned} \tag{3.24}$$

The number of degrees of freedom is equal to the column rank of $\mathbf{C}$ (not counting multinomial sampling constraints). It should be noted that the sample size is not automatically reproduced, and, with (product) multinomial sampling, the sampling constraints should be included in the model constraints.

3.4 Assessing model goodness-of-fit

3.4.1 Chi-squared test statistics

When estimated expected frequencies satisfying the model have been obtained, the chi-squared statistics described in Section 2.5.1, the discrimination information, or Neyman's statistic can be used to test goodness-of-fit. The number of degrees of freedom is equal to the number of independent constraints of the model.

The Pearson chi-squared statistic has a special form, which will be derived next. First note that, from (3.5),

$$\mathbf{n} - \hat{\mathbf{m}} = -\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C}\hat{\boldsymbol{\lambda}}.$$

Premultiplying both sides of this equation by $-(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C})^{-1}\mathbf{C}'$ yields

$$\hat{\boldsymbol{\lambda}} = -(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{n} - \hat{\mathbf{m}}),$$

so that

$$\mathbf{n} - \hat{\mathbf{m}} = \mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C}(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{n} - \hat{\mathbf{m}}).$$

Thus, since $\mathbf{C}'\hat{\mathbf{m}} = \mathbf{d}$,

$$\begin{aligned} X^2 &= (\mathbf{n} - \hat{\mathbf{m}})'\mathbf{D}_{\hat{\mathbf{m}}}^{-1}(\mathbf{n} - \hat{\mathbf{m}}) = (\mathbf{n} - \hat{\mathbf{m}})'\mathbf{C}(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C})^{-1}\mathbf{C}'(\mathbf{n} - \hat{\mathbf{m}}) \\ &= (\mathbf{C}'\mathbf{n} - \mathbf{d})'(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{n} - \mathbf{d}). \end{aligned} \tag{3.25}$$

To use the Wald statistic, no estimated frequencies are needed. For testing a hypothesis $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, the Wald statistic has the form

$$W^2 = \mathbf{h}(\mathbf{n})'\boldsymbol{\Sigma}(\mathbf{h}(\mathbf{n}))^{-1}\mathbf{h}(\mathbf{n}),$$

with $\boldsymbol{\Sigma}(\mathbf{h}(\mathbf{n}))$ an estimate of the covariance matrix of $\mathbf{h}(\mathbf{n})$. This statistic was used by Stuart (1955) and Bhapkar (1966) to test for MH. Their tests differ because they estimated $\boldsymbol{\Sigma}(\mathbf{h}(\mathbf{n}))$ differently. Stuart only described a test for MH for square tables, which was difficult enough to evaluate, given the state of the art in computing in the mid-fifties. For an $I \times I$ table, with $h_i = n_{i+} - n_{+i}$ and with $i, j < I$, he estimated the covariance between h_i and h_j as

$$\text{cov}(h_i, h_j) = \begin{cases} n_{i+} + n_{+i} - 2n_{ii} & i = j \\ -(n_{ij} + n_{ji}) & i \neq j \end{cases}.$$

This is the sample value of the MLE of the covariance between h_i and h_j under multinomial sampling, assuming MH holds (Stuart, 1955). Alternatively, $\text{cov}(h_i, h_j)$ can be interpreted as the sample value of the MLE of the covariance between h_i and h_j under Poisson sampling, without assuming MH. Note that the test does not use the diagonal counts. Denoting Stuart's test for an $I \times I$ table by S_I^2, yields

$$S_2^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}.$$

for a 2×2 table. This test is best known as *McNemar's test* (McNemar, 1947). Explicit formulae for Stuart's test for 3×3 and 4×4 tables were given by Fleiss and Everitt (1971). Fryer (1971) showed how Stuart's test could be generalized to the multi-dimensional case. In matrix notation, a generalization of Stuart's test for models involving linear constraints is

$$S^2 = (\mathbf{C}'\mathbf{n} - \mathbf{d})'(\mathbf{C}'\mathbf{D_n}\mathbf{C})^{-1}(\mathbf{C}'\mathbf{n} - \mathbf{d}),$$

where the multinomial sampling constraints are *not* included in the model constraints. Note that S^2 is identical to the sample value of (3.25). It is of the same form as the minimum modified chi-squared statistic (3.24), with the difference that there the multinomial sampling constraints are included.

Bhapkar estimated the covariance matrix of $\mathbf{h}(\mathbf{n})$ differently, namely as

$$\mathrm{cov}(h_i, h_j) = \begin{cases} n_{i+} + n_{+i} - 2n_{ii} - (n_{i+} - n_{+i})^2 & i = j \\ -(n_{ij} + n_{ji}) - (n_{i+} - n_{+i})(n_{j+} - n_{+j}) & i \neq j \end{cases}.$$

This is the sample value of the MLE of the covariance between h_i and h_j under multinomial sampling without assuming MH. As it is hard to see why the diagonal counts should have a large influence on goodness-of-fit of MH, Stuart's test seems preferable. Denoting Bhapkar's test as B^2, Bennett (1967) showed that the following relationship holds between the two statistics.

$$B^2 = \frac{S^2}{1 - S^2/n_{++}}. \tag{3.26}$$

It follows that $B^2 > S^2$. Krauth (1985) did some simulations for the 3×3 table, and found that B^2 tends to be liberal, while S^2 was sometimes conservative, sometimes liberal. Thus, it seems better to compare the tests, rather than use one of them blindly.

3.4.2 A conditional test for MH using standard loglinear models

Sometimes a computer program for testing MH directly may not be available, while a program for testing various loglinear models is. In such cases, the following conditional test for MH can sometimes be used.

Caussinus (1965) described a conditional test for MH using the log-linear model of *quasi-symmetry* (QS) and *symmetry* (S), which will be defined below. He noted that

$$\text{QS} \cap \text{MH} = \text{S} . \tag{3.27}$$

A conditional test for MH against the alternative of QS can be obtained by using an asymptotic partitioning of chi-squared.

The symmetry (S) model for a two-way table is defined by the constraint equations

$$m_{ij} = m_{ji}.$$

The sufficient statistics are $n_{ij}+n_{ji}$, and the MLEs are $\hat{m}_{ij} = (n_{ij}+n_{ji})/2$. Quasi-symmetry (QS) is defined by restricting the saturated model

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

by the constraints $\lambda_{ij}^{AB} = \lambda_{ji}^{AB}$. The sufficient statistics are $n_{ij} + n_{ji}$, n_{i+}, and n_{+i} (Haberman, 1979, p. 490).

The sufficient statistics of QS are a linear combination of the marginal frequencies, so, as demonstrated in Section 5.4.2, it follows that QS and MH are asymptotically separable, meaning that a chi-squared statistic can be asymptotically partitioned. That is, if symmetry holds, G^2 can asymptotically be partitioned as

$$G^2(\text{S}) = G^2(\text{MH}) + G^2(\text{QS}).$$

The following conditional test of MH is obtained.

$$G^2(\text{MH}) = G^2(\text{S}) - G^2(\text{QS}), \tag{3.28}$$

based on *df*$= I - 1$. Asymptotically, if QS holds, the conditional test has the same power as the unconditional one. A drawback of the conditional test is that it can only be used if the hypothesis QS holds.

3.4.3 Asymptotic behaviour of MLEs

The asymptotic distribution of the MLE $\hat{\mathbf{m}}$ can be derived using the technique of Aitchison and Silvey, as described in Appendix A. If the model $\mathbf{C}'\mathbf{m} = \mathbf{d}$ is true, $\hat{\mathbf{m}}$ has an asymptotic multivariate normal distribution,

with mean converging to the population value of $\mathbf{m}$, and an asymptotic covariance matrix which can be estimated as

$$\boldsymbol{\Sigma}(\hat{\mathbf{m}}) = \mathbf{D}_{\hat{\mathbf{m}}} - \mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C}(\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}\mathbf{C})^{-1}\mathbf{C}'\mathbf{D}_{\hat{\mathbf{m}}}.$$

It should be noted that multinomial sampling constraints should be included in the model constraints to obtain the correct estimated covariance matrix.

3.5 Example: Unaided distance vision

The methods discussed in this chapter are applied to the data in Table 3.1. The observed frequencies, the MLEs, and the standardized residuals given the marginal homogeneity model

$$m_{i+} = m_{+i} \qquad i = 1, 2, 3, 4,$$

are presented in Table 3.2. Note that the observations on the diagonal elements of the table are reproduced. The frequencies which minimize the discrimination information statistic I and those which minimize Neyman's statistic X^2_{mod} given MH are presented in Table 3.3. It should be noted that the three different types of estimates are remarkably similar. The most conspicuous difference is that, in Table 3.3, the observations on the main diagonal are not reproduced, but are equal to the observed diagonal elements multiplied by a factor greater than one (1.0008 and 1.0016, respectively) for each of the two estimation methods. This is generally the case: for the MH model under multinomial sampling, the two estimation methods yield estimated diagonal frequencies greater or equal than the observed diagonal frequencies.

In Table 3.4, the values of the four different test statistics discussed in this chapter are presented using estimates obtained with the three estimation methods discussed above. Note that the values of all statistics are very similar. For all statistics, the p-value is less than 1%. Though MH does not yield a good fit, the difference between the marginal distributions is not large: on average, the column totals differ from the corresponding row totals by only 3.3%. This is the reason that the various estimators and test statistics have similar values, which would not be expected if the observed marginals were very different. It can be noted from Table 3.2 that the absolute values of the adjusted marginal residuals are greatest in

1520	266	124	66	1976
(1520)	(252.5)	(111.8)	(57.0)	(1941.3)
0	1.47	2.73	3.18	2.39
234	1512	432	78	2256
(247.2)	(1512)	(409.4)	(70.6)	(2239.2)
-1.47	0	1.81	2.37	0.88
117	362	1772	205	2456
(131.3)	(383.1)	(1772)	(195.3)	(2481.7)
-2.73	-1.81	0	1.21	-1.36
36	82	179	492	789
(42.8)	(91.6)	(188.4)	(492)	(814.8)
-3.18	-2.37	-1.21	0	-2.03
1907	2222	2507	841	7477
(1941.3)	(2239.2)	(2481.7)	(814.8)	(7477)
-2.36	-0.90	1.34	2.06	0

Table 3.2: Observed data, maximum likelihood estimates (in brackets), and adjusted residuals for the data of Table 3.1 given marginal homogeneity

1521.22	252.3	111.3	56.3	1941.1
1522.43	252.2	110.7	55.6	1941.0
247.1	1513.2	409.1	70.3	2239.6
246.9	1514.4	408.7	69.9	2240.0
130.6	382.9	1773.4	195.2	2482.1
129.9	382.6	1774.8	195.1	2482.5
42.2	91.2	188.3	492.4	814.1
41.8	90.7	188.2	492.8	813.5
1941.1	2239.6	2482.1	814.1	7477
1941.0	2240.0	2482.5	813.5	7477

Table 3.3: Minimum discrimination information (top) and minimum modified chi-squared estimates (bottom) for the data of Table 3.1 given marginal homogeneity

	Test statistics			
Estimation method	G^2	I	X^2_{mod}	X^2
Maximum likelihood	11.987	12.027	12.089	11.970
min. discr. inform. estim.	12.015	11.998	12.003	12.053
min. X^2_{mod} estimation	12.100	12.027	11.976	12.197

Table 3.4: Chi-squared statistics with different estimation methods for the data of Table 3.1 given marginal homogeneity (df=3)

the first and last categories. Additionally, the adjusted residuals become greater in absolute value towards the upper left and lower right corners. For marginal homogeneity to hold, there are too many observations in the category "Highest" for the right eye and in the category "Lowest" for the left eye. This indicates that lack of fit may be caused by the right eye being on average slightly better than the left eye.

Chapter 4

Marginal models

Two distinct types of models were discussed in the previous two chapters. Chapter 2 demonstrated how the *joint* distribution of a set of variables can be modelled using loglinear models. In Chapter 3, models which induce linear constraints on expected frequencies were presented. The latter models are most useful for testing whether there are differences in various *marginal* distributions. The aim of this chapter is to discuss extensions which are relevant for categorical data analysis of both types of models. To this end, a class of models referred to as *marginal models* is introduced. Though marginal models can also be used to model the joint distribution of a set of variables, the emphasis is on modelling marginal distributions. The three main applications of marginal models are distinguished below.

Firstly, marginal models provide alternatives to the loglinear models of Chapter 2 for modelling the joint distribution of several variables. For example, global odds ratio models can, in certain cases, be used in place of loglinear models (Dale, 1986; Molenberghs & Lesaffre, 1994). Furthermore, the association between two variables may be modelled by means of the measure gamma (Agresti, 1984) instead of odds ratios, as in the loglinear modelling approach. Association is but one aspect of the joint distribution of variables; many other aspects may be of interest. One example is agreement, which can be modelled by means of the measure of agreement kappa (Cohen, 1960).

The second reason for using marginal models is to provide extensions of the marginal homogeneity models of Chapter 3. The models of Chapter 3 are mostly used for testing equality of various marginal distributions.

However, it may be useful to test whether specific aspects of marginal distributions are equal, rather than the complete marginal distributions. As an example, consider a four-way table $ABCD$. Using the models presented in Chapter 3, it is possible to test whether the distributions of AB and CD are equal. A weaker hypothesis is that the association in tables AB and CD is equal. The latter can be tested by means of the local odds ratios in AB and CD. Alternatively, other measures of association can be used, such as global odds ratios, gamma, or Kendall's tau. The homogeneity of aspects other than association may also be of interest, e.g., agreement, which can be tested for homogeneity using the agreement measure kappa. Of course, attention need not be restricted to the two-way marginal tables AB and CD. If A, B, C, and D are interval level variables, the means of the respective variables may be tested for equality. Furthermore, using a regression model, it is possible to test whether some other relation between the means holds. In general, models for testing some form of homogeneity (not necessarily equality) among marginal distributions are referred to as marginal homogeneity models.

Finally, the third application of marginal models is to provide a methodology for simultaneously modelling different models of the types described above. In a four-way table $ABCD$, a marginal homogeneity model for tables AB and CD may be tested simultaneously with the restriction that the linear by linear association model for the joint distribution of table $ABCD$ holds (Lang & Agresti, 1994). Even though sometimes the different models may be fitted separately, there can be several statistical advantages to simultaneous fitting.

Summarizing, marginal models can be used to model the joint distribution of a set of variables, various types of marginal homogeneity, and simultaneous models. Some readers may find the term "marginal model" not entirely appropriate. However, a large body of literature is accumulating in which this term is used in a similar sense. For instance, Liang, Zeger, and Qakish (1992), Molenberghs and Lesaffre (1994), and Becker (1994) used the term "marginal models" to refer to various log-linear models for marginal (or sums of) frequencies. Besides being useful for modelling different forms of marginal homogeneity, an important feature of this latter type is that it can also be used to model the joint distribution of a set of variables. To be consistent with the literature, the models introduced in this chapter are also referred to as marginal models. However, those considered here are much more general than the marginal

models discussed by the aforementioned authors. This is mainly because measures which are not of a loglinear type, such as gamma or kappa, can also be modelled.

A general equation for representing marginal models is given in Section 4.1. In Section 4.2, it is shown how marginal models can be applied. Several measures which are commonly used in categorical data analysis for summarizing aspects of the joint distribution of variables are described in Section 4.3, and a convenient matrix notation for implementing these measures is presented. Finally, the marginal modelling approach is illustrated by an example in Section 4.4. A description of methods for fitting and testing marginal models is dealt with in the next chapter.

4.1 Definition of marginal models

Consider an $I_1 \times \ldots \times I_Q$ contingency table $A_1 \ldots A_Q$ with expected joint frequencies $m_{i_1 \ldots i_Q}$. A marginal model is a model for *marginal frequencies*. A marginal frequency is defined as a frequency formed by (partial) summation over certain indices of the expected frequencies $m_{i_1 \ldots i_Q}$. Formally, let $\mathcal{I}_q = \{i_{q1}, \ldots, i_{qh_q}\} \subseteq \{1, \ldots, I_q\}$ $(1 \leq h_q \leq I_q)$ be a subset of the indices of variable A_q $(q = 1, \ldots, Q)$. A marginal frequency is formed by summing the expected joint frequencies over the subsets of indices $\mathcal{I}_q$:

$$\mu = \sum_{i_1 \in \mathcal{I}_1} \ldots \sum_{i_Q \in \mathcal{I}_Q} m_{i_1 \ldots i_Q}. \tag{4.1}$$

By way of illustration, consider a $3 \times 3 \times 3$ table with expected frequencies m_{ijk}. Let $\mathcal{I}_1 = \{1, 2\}$, $\mathcal{I}_2 = \{2\}$, and $\mathcal{I}_3 = \{1, 2, 3\}$. The marginal frequency formed by summing over these index sets is

$$\sum_{i \in \{1,2\}} \sum_{j \in \{2\}} \sum_{k \in \{1,2,3\}} m_{ijk} = m_{12+} + m_{22+}.$$

It is important to note that using this definition, joint frequencies are also marginal frequencies. With $r = \prod_q I_q$, the expected frequencies of an $I_1 \times \ldots \times I_Q$ contingency table can be arranged in an $r \times 1$ vector $\mathbf{m}$ of expected joint frequencies. A marginal frequency μ, as defined above, is a sum of elements of $\mathbf{m}$. Thus, an $s \times 1$ vector $\boldsymbol{\mu}$ of marginal frequencies μ_i can be represented by the formula

$$\boldsymbol{\mu} = \mathbf{M}'\mathbf{m},$$

where $\mathbf{M}$ is an $r \times s$ matrix of zeroes and ones. However, the results derived in the sequel also apply when $\mathbf{M}$ consists of arbitrary nonnegative elements.

For a vector of marginal frequencies $\boldsymbol{\mu}$, consider a collection of measures $\zeta_i(\boldsymbol{\mu})$ $(i = 1, \ldots, z)$. A marginal model is defined by the equation

$$g_i(\zeta_i(\boldsymbol{\mu})) = \sum_j x_{ij}\beta_j \quad i = 1, \ldots, z. \tag{4.2}$$

Using the generalized linear models terminology, g_i is a *link* function. The β parameters are unknown, and the x_{ij} are known covariates. There may be no covariates, in which case the right hand side of equation (4.2) is zero. Let $\boldsymbol{\zeta}(\boldsymbol{\mu})$ be the $z \times 1$ vector of measures with ith element $\zeta_i(\boldsymbol{\mu})$. In matrix notation, the marginal model equation is

$$\mathbf{g}(\boldsymbol{\zeta}(\boldsymbol{\mu})) = \mathbf{X}\boldsymbol{\beta}, \tag{4.3}$$

where $\mathbf{X}$ may be a zero matrix. Since the μ_i may be correlated, and since $\zeta_i(\boldsymbol{\mu})$ may be a complicated function of $\boldsymbol{\mu}$, marginal models are, in general, not special cases of generalized linear models (McCullagh & Nelder, 1989), and therefore the fitting and testing methods developed for the latter models cannot be used. The testing and fitting methods described in the next chapter are only applicable when $\mathbf{g}$ and $\boldsymbol{\zeta}$ are "smooth" functions, in the sense that they are differentiable over the relevant part of the parameter space.

Equation (4.3) can be used to model a specific property (or aspect) of a (marginal) distribution. It is necessary that this property be summarizable by a set of one or more numbers (or measures). For instance, association can be summarized using a set of odds ratios or using gamma. Various measures commonly used in categorical data analysis for summarizing different aspects of a distribution are described in Section 4.3.

Since a linear predictor $\mathbf{X}\boldsymbol{\beta}$ is used in equation (4.3), it is appropriate that a link function maps a measure ζ monotonically onto the whole real line. For instance, a Poisson expected frequency m_i has range $0 < m_i < \infty$, so the log link can be used. The link function

$$g(\zeta) = \log \frac{1+\zeta}{1-\zeta} \tag{4.4}$$

maps the interval $\langle -1, 1 \rangle$ monotonically onto the whole real line. This link will be referred to as the *log-hyperbolic* link. Many useful measures ζ

have a range $-1 \leq \zeta \leq 1$. If the values $\zeta = \pm 1$ can safely be disregarded, the log-hyperbolic link can be used. Several measures (such as gamma, see page 77) which range from -1 to 1 are written naturally as

$$\zeta = \frac{A - B}{A + B} \tag{4.5}$$

for some functions A and B of frequencies. The maximum value $\zeta = 1$ is attained when $B = 0$, the minimum value $\zeta = -1$ is attained when $A = 0$, and $\zeta = 0$ when $A = B$. The log-hyperbolic link for measures defined by (4.5) yields

$$g(\zeta) = \log \frac{A}{B}.$$

We can see that g transforms ζ into a logit.

The model specification (4.3) is very general. Unfortunately, this is not without danger. Care should be taken that conflicting constraints are not specified simultaneously. For instance, for a square table, the loglinear model of symmetry, which implies marginal homogeneity, is inconsistent with a constraint such as $m_{1+} - m_{+1} = d$, if $d \neq 0$. A given model is consistent if there is at least one vector of frequencies $\mathbf{m}$ satisfying the model. Many interesting models have the equiprobability model, for which all cells have the same expected frequencies, as a special case. This provides an easy test for consistency of a given model. However, if this test fails, inconsistency is not implied. In general, testing consistency is difficult.

4.2 Applying marginal models

In the previous section, marginal models were presented in a somewhat formal manner. This section is intended to show in a less formal way how marginal models can be used. Model equation (4.3) allows two distinct types of models to be specified: those for describing the joint distribution of a set of variables and marginal homogeneity (MH) models. Additionally, several of these types can be specified simultaneously. In the three subsections below, the various types of models are explicated. Occasionally a reference will be made to Section 4.3, where definitions of different measures are presented. In contrast, the emphasis here is on the kind of models that can be constructed using these measures.

4.2.1 Modelling the joint distribution

As explained in the introduction to this chapter, marginal models provide extensions of the loglinear approach to modelling the joint distribution of a set of variables. To see how loglinear models can be extended, it is useful to remember that most interesting loglinear models can be phrased as models for local odds ratios. This can be illustrated as follows. For an $I \times J$ table, let $\zeta_{ij}^{(1)}$ be the (i,j)th local odds ratio, i.e.,

$$\zeta_{ij}^{(1)} = \frac{m_{ij}\, m_{i+1,j+1}}{m_{i+1,j}\, m_{i,j+1}}.$$

$(i = 1, \ldots, I-1,\ j = 1, \ldots, J-1)$. The independence model can be phrased as a model for local odds ratios by the requirement that their logarithms are zero, i.e., by the requirement that

$$\log \zeta_{ij}^{(1)} = 0 \qquad \forall i, j. \tag{4.6}$$

The independence model (4.6) has the same form as (4.3) with $\boldsymbol{\zeta}(\boldsymbol{\mu})$ a vector consisting of the local odds ratios, $\mathbf{g}$ the log link function, and $\mathbf{X}$ the zero matrix. The linear by linear association model (see Section 2.3.2) can be phrased as a model for local odds ratios by the equation

$$\log \zeta_{ij}^{(1)} = x_{ij}\beta \qquad \forall i, j, \tag{4.7}$$

where $x_{ij} = (a_i - a_{i+1})(b_j - b_{j+1})$ for known a_i and b_j and β is an unknown parameter.

The loglinear modelling approach can be extended by substituting the $\zeta_{ij}^{(1)}$ in equations (4.6) and (4.7) by measures other than the local odds ratios. For instance, for ordinal variables, it may be more appropriate to use *global odds ratios* which are defined as

$$\zeta_{ij}^{(2)} = \frac{\left(\sum_{k \le i} \sum_{l \le j} m_{kl}\right)\left(\sum_{k > i} \sum_{l > j} m_{kl}\right)}{\left(\sum_{k \le i} \sum_{l > j} m_{kl}\right)\left(\sum_{k > i} \sum_{l \le j} m_{kl}\right)},$$

with $1 \le i \le I-1$ and $1 \le j \le J-1$ (see also Section 4.3.5). Equating all log global odds ratios to zero yields the independence model, and is therefore equivalent to equating the log local odds ratios to zero. However, defining the linear by linear association model (4.7) using global odds

ratios yields a different model than the linear by linear model using local odds ratios (Clayton, 1974; Wahrendorf, 1980; Anscombe, 1981). More general models for global odds ratios were presented by Semenya and Koch (1980, p. 103–118). In general, models for global odds ratios are not loglinear in expected joint frequencies of the complete table, but rather loglinear in sums of joint frequencies. This implies that, in general, the fitting methods described in Chapter 2 cannot be used. Still other types of odds ratios in addition to the two given above are described by Agresti (1984).

Besides various types of odds ratios, many other measures have been devised to measure association between two categorical variables. For instance, gamma (see Section 4.3.5) can be used as a measure of association for ordinal variables. Other aspects may also be modelled. For example, for two variables which are ratings, a hypothesis about the strength of agreement between two raters may be made. Provided one or more appropriate measures describing an aspect of the joint distribution of a set of variables can be found, this aspect can be modelled using model equation (4.3).

Finally, some specific methods are given which have been described in the literature for modelling a joint distribution of dimensionality other than two. For modelling a univariate "joint" distribution, Agresti (1990, Chapter 9) provides an overview of several types of logit models. Molenberghs and Lesaffre (Molenberghs & Lesaffre, 1994) developed analogues of global odds ratios for three or more dimensions.

4.2.2 Marginal homogeneity models

In the previous section, it was discussed how the joint distribution of a set of variables can be modelled. In this section, marginal homogeneity models, which are used to specify various forms of homogeneity of aspects of marginal distributions are presented.

In Chapter 3, it was shown how equality of marginal distributions can be modelled by imposing linear constraints on expected frequencies. The hypothesis of complete equality is rather strong, and one may wish to model some weaker form of homogeneity, in particular, homogeneity of specific aspects of the marginal distributions. For example, in a four-way table $ABCD$, there may be interest in testing homogeneity hypotheses for tables AB and CD. Using the models in Chapter 3, it is possible to test

whether AB and CD are identically distributed. A weaker homogeneity hypothesis is whether the association between A and B is the same as the association between C and D. This can be tested using an appropriate measure of association, such as local or global odds ratios, or gamma.

More generally, for a set of T bivariate marginal tables, marginal homogeneity hypotheses other than the one asserting equality of association may be tested. For instance, if the T tables represent points in time, it is possible to test whether there is a linear decrease or increase in association over time. To illustrate, there may be interest in how the association between income and level of education changes over time. The hypothesis can be tested that shortly after leaving school there is a relatively strong association which decreases over time.

Of course, other aspects of marginal distributions can also be tested for homogeneity. For univariate marginal distributions of interval level variables, a regression model may be fitted for modelling homogeneity of the means. When bivariate marginal distributions are ratings, one may test for homogeneity of agreement between raters by testing whether kappa is equal in different marginal tables.

Next, some simple but useful regression models for modelling marginal homogeneity are presented. Suppose we are interested in a set of T marginal distributions, denoted by $\boldsymbol{\mu}(t)$. For instance, in a six-way table $ABCDEF$, one may take $\mu_{ij}(1) = m_{ij++++}$, $\mu_{kl}(2) = m_{++kl++}$, and $\mu_{gh}(3) = m_{++++gh}$. Suppose a certain aspect of marginal distribution $\boldsymbol{\mu}(t)$ is summarized by measures $\zeta_h(\boldsymbol{\mu}(t))$ $(h = 1, \ldots, z)$. An aspect may be summarized by a single measure or by a set of measures. For instance, association may be summarized by the single measure gamma or by a set of odds ratios. To specify a regression model using model equation (4.3), an appropriate link function should be chosen for $\zeta_h(\boldsymbol{\mu})$. Suppose the link g_h for $\zeta_h(\boldsymbol{\mu}(t))$ $(t = 1, \ldots, T)$ is chosen. Then let

$$\eta_h(t) = g_h(\zeta_h(\boldsymbol{\mu}(t))) \qquad \forall h, t.$$

A model asserting that $\eta_h(t)$ is identical for all t can be represented by the parameterization

$$\eta_h(t) = \alpha_h \qquad \forall h, t$$

where the α_h are unknown parameters. A second model represents the hypothesis that $\eta_h(t)$ is a linear function of covariates x_t. For known

covariates x_t and with model parameters α_h and β, this model can be specified using the equation

$$\eta_h(t) = \alpha_h + \beta x_t \qquad \forall h, t.$$

If the distributions $\boldsymbol{\mu}(t)$ represent equally spaced points in time, a sensible choice of the covariates may be $x_t = t$. Many other nonlinear regression lines may be of interest in a specific situation. For instance, if the marginal distributions represent points in time, one might entertain the hypothesis that the value of a measure converges monotonically over time to an unknown value α, with unknown initial deviation β, and known "rate" $k > 0$. Then the model specified by the hyperbolic function

$$\eta_h(t) = \alpha_h + \beta x_t^{-k} \qquad \forall h, t$$

can be used.

4.2.3 Simultaneous modelling

Sometimes it can prove useful to test several of the models described above simultaneously. For instance, in a table $ABCD$, homogeneity of association between AB and CD may be tested given the additional restriction that the linear by linear interaction model holds for both tables AB and CD. Sometimes the different models may be fitted separately, and goodness-of-fit may be assessed separately for each model. An advantage of simultaneous fitting rather than separate fitting is that greater efficiency of expected joint frequencies (i.e., the frequencies in the cells of the complete table) is obtained.

However, the efficiency of fitted model parameters or marginal frequencies is not necessarily improved by simultaneous modelling. In particular, if two models imply restrictions for two sets of orthogonal parameters respectively, the asymptotic standard errors of estimators of these parameters are the same whether the models are fitted separately or simultaneously (see Appendix A.3). For instance, in a two-way table with expected frequencies m_{ij}, the marginal frequencies m_{i+} and m_{+j} are orthogonal to the local odds ratios (see Section 5.4.2). Efficiency of the marginal frequencies is not improved by fitting a model for the odds ratios, such as independence. Still, even in such cases, simultaneous fitting may be preferable to separate fitting because of the greater efficiency of estimated expected joint frequencies.

4.3 Definition and representation of measures for categorical data

In model equation (4.3), $\boldsymbol{\zeta}(\boldsymbol{\mu})$ may be any differentiable measure of frequencies or probabilities. In this section, definitions of some widely used measures for categorical data are presented, and it is shown how these measures can be written in matrix notation. It should be noted that in contrast to the previous section, modelling is not discussed here.

The reason for using matrix notation is the following. In order to be able to use the fitting and testing methods of the next chapter, the derivative matrix of $\boldsymbol{\zeta}(\boldsymbol{\mu})$ must be calculated. These derivatives can be calculated directly by hand, but in many computer languages, automatic differentiation is awkward to implement. It is then more convenient to have a single general (matrix) formula with which many different measures can be represented. Only the derivative matrix of this formula needs to be implemented without having to program rules for differentiation. Kritzer (1977) gave matrix formulas for several measures of association. His approach is generalized in Section 4.3.1. All measures are written in terms of expected probabilities π_i, instead of expected frequencies μ_i.

4.3.1 A recursive "exp-log" notation for representing measures

A convenient method for representing all measures described in this chapter is formulated. The method of notation will be referred to as the "exp-log" notation. To illustrate the basic idea of the "exp-log" notation, consider the fraction $(\pi_1 + \pi_2)/(\pi_3 + \pi_4)$. Using matrices this expression can be written as

$$\begin{aligned}\frac{\pi_1 + \pi_2}{\pi_3 + \pi_4} &= \exp(\log(\pi_1 + \pi_2) - \log(\pi_3 + \pi_4)) \\ &= \exp\left[\begin{pmatrix} 1 & -1 \end{pmatrix} \log \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}\right].\end{aligned}$$

In general, any product of strictly positive terms involves exponentiating the sum of the logarithms of the terms. Care must be taken to ensure

that terms of which the logarithm is taken are positive. By repeated application of an "exp-log" function as given above, all measures described below can be obtained. A definition of "exp-log" functions is given below. In the remaining part of this section, "sum" will be used for a possibly weighted sum of terms, and "product" for a product of powers of terms.

First, the simplest nontrivial "exp-log" function is given. Let $\mathbf{A}$ be an $c_0 \times a_0$ matrix with elements $a_{ij} \geq 0$, where it is assumed that every column of $\mathbf{A}$ contains at least one nonzero element, and let $\mathbf{C}$ be an $a_0 \times c_1$ matrix with elements c_{jk}. With $\boldsymbol{\pi}$ a $c_0 \times 1$ vector of strictly positive parameters, a vector $\boldsymbol{\zeta}(\boldsymbol{\mu})$ which can be written in matrix notation as

$$\boldsymbol{\zeta}(\boldsymbol{\pi}) = \exp(\mathbf{C}' \log \mathbf{A}' \boldsymbol{\pi}) \tag{4.8}$$

is an "exp-log" measure of order 1. The elements of $\boldsymbol{\zeta}(\boldsymbol{\pi})$ have the form

$$\zeta_k(\boldsymbol{\pi}) = \prod_j \left(\sum_i a_{ij} \pi_i \right)^{c_{jk}}. \tag{4.9}$$

Using the log link, the marginal model equation (4.3) for measures of the form (4.8) is specified by the equation

$$\mathbf{C}' \log \mathbf{A}' \boldsymbol{\pi} = \mathbf{X} \boldsymbol{\beta}. \tag{4.10}$$

If $\mathbf{C}'$ is a contrast matrix, it can be verified that $\boldsymbol{\zeta}(\boldsymbol{\pi})$ is a homogeneous function of $\boldsymbol{\pi}$ (i.e., $\boldsymbol{\zeta}(\mathbf{m}) = \boldsymbol{\zeta}(\boldsymbol{\pi})$, see Appendix B.2). The model equation (4.10) was studied by Haber (1985), Lang and Agresti (1994), and Lang (1996a). Special cases of (4.10) have been considered by McCullagh and Nelder (1989, p. 219), Liang, Zeger, and Qakish (1992), Molenberghs and Lesaffre (1994), Glonek and McCullagh (1995), and Becker (1994), Many well-known measures can be written in the same form as (4.9). These include various types of odds, such as adjacent-category, cumulative, and continuation-ratio odds, various types of odds ratios, such as local and global odds ratios, and conditional probabilities. In Section 4.3.5, it is described how the matrices $\mathbf{A}$ and $\mathbf{C}$ for local and global odds ratios can be obtained. The matrices for representing various types of odds can be obtained from the matrices for the odds ratios.

Next, a generalization of (4.8) is considered. Suppose that, for $i = 0, \ldots, k-1$, $\mathbf{A}_i$ and $\mathbf{C}_i$ are given $c_i \times a_i$ and $a_i \times c_{i+1}$ matrices respectively, where $\mathbf{A}_i$ contains only nonnegative elements, and each row of $\mathbf{A}_i$ contains

at least one nonzero element. A function $\mathbf{t}_k(\boldsymbol{\pi})$, with $\boldsymbol{\pi}$ a $c_0 \times 1$ vector of strictly positive parameters, can be defined recursively as

$$\begin{aligned} \mathbf{t}_0(\boldsymbol{\pi}) &= \boldsymbol{\pi} && (4.11) \\ \mathbf{t}_{i+1}(\boldsymbol{\pi}) &= \exp[\mathbf{C}_i' \log \mathbf{A}_i' \mathbf{t}_i(\boldsymbol{\pi})] \qquad i = 0, \ldots, k-1. && (4.12) \end{aligned}$$

Thus, $\mathbf{t}_{i+1}(\boldsymbol{\pi})$ is a product of sums of elements of $\mathbf{t}_i(\boldsymbol{\pi})$. Note that measures defined by (4.8) are $\mathbf{t}_1$-functions. Forthofer and Koch (1973) described some useful $\mathbf{t}_2$-functions.

All the measures defined below can be written as

$$\zeta(\boldsymbol{\pi}) = \mathbf{e}' \mathbf{t}_k(\boldsymbol{\pi}), \tag{4.13}$$

for some k and some $c_k \times 1$ vector $\mathbf{e}$. Vector $\mathbf{e}$ can be useful for defining contrasts of $\mathbf{t}_k$ functions, but $\mathbf{e}$ can also equal one or any other vector. For a given vector of measures $\boldsymbol{\zeta}(\boldsymbol{\pi})$, finding appropriate matrices $\mathbf{A}_i$ and $\mathbf{C}_i$ such that $\boldsymbol{\zeta}(\boldsymbol{\pi}) = \mathbf{t}_k(\boldsymbol{\pi})$ as defined in (4.11) and (4.12) can be rather tedious. It is shown below how this can be done for several widely used measures.

The derivative matrix of $\mathbf{t}_i(\boldsymbol{\pi})$ is straightforward. Define

$$\mathbf{T}_i(\boldsymbol{\pi}) = \frac{\partial \mathbf{t}_i(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}'}.$$

Writing $\mathbf{t}_i$ and $\mathbf{T}_i$ for $\mathbf{t}_i(\boldsymbol{\pi})$ and $\mathbf{T}_i(\boldsymbol{\pi})$, respectively, we can verify that

$$\begin{aligned} \mathbf{T}_0 &= \mathbf{I} \\ \mathbf{T}_{i+1} &= \mathbf{Diag}[\mathbf{t}_{i+1}]\, \mathbf{C}_i' \, \mathbf{Diag}[\mathbf{A}_i' \mathbf{t}_i]^{-1} \mathbf{A}_i' \, \mathbf{T}_i \qquad i = 0, \ldots, k-1 \end{aligned}$$

where $\mathbf{Diag}[\cdot]$ represents a diagonal matrix with the elements of the vector in square brackets on the main diagonal. The derivative matrix of $\boldsymbol{\zeta}(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$ is

$$\frac{\partial \boldsymbol{\zeta}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}'} = \mathbf{E}' \mathbf{T}_{i+1}.$$

4.3.2 Homogeneity of measures

Consider a $t \times 1$ vector $\boldsymbol{\pi}$ and a measure $\zeta(\boldsymbol{\pi})$. The measure (or function) $\zeta(\boldsymbol{\pi})$ is said to be *homogeneous* if $\zeta(\boldsymbol{\pi}) = \zeta(c\boldsymbol{\pi})$ for all $c > 0$ (see also Appendix D; note that homogeneity of functions is not related to marginal

homogeneity). Homogeneity ensures that the same value of the measure is obtained whether probabilities or frequencies are used in the argument. In the next chapter, it will be demonstrated that homogeneity of measures is also convenient for ML estimation, ensuring that the same algorithm can be used for both Poisson and multinomial fitting.

The "exp-log" notation can be used to rewrite any measure which is a function of probabilities as a homogeneous function by using the fact that $\pi_+ = 1$. Let $\zeta^*(\boldsymbol{\pi}) = \zeta(\boldsymbol{\pi}/\pi_+)$. Then $\zeta^*(\boldsymbol{\pi}) = \zeta^*(c\boldsymbol{\pi})$, so ζ^* is homogeneous and has the same value as $\zeta(\boldsymbol{\pi})$ because $\pi_+ = 1$. The term $\boldsymbol{\pi}/\pi_+$ can be written using the "exp-log" notation as

$$\frac{1}{\pi_+}\boldsymbol{\pi} = \exp(\log \boldsymbol{\pi} - \log \mathbf{1}_t\mathbf{1}_t'\boldsymbol{\pi}) = \exp\left[\begin{pmatrix} \mathbf{I}_t & -\mathbf{I}_t \end{pmatrix} \log \begin{pmatrix} \mathbf{I}_t \\ \mathbf{1}_t\mathbf{1}_t' \end{pmatrix} \boldsymbol{\pi}\right],$$

where $\mathbf{1}_t$ is the $t \times 1$ vector with elements 1, and $\mathbf{I}_t$ is the $t \times t$ identity matrix. Now, using the definitions (4.11) and (4.12), $\boldsymbol{\pi}/\pi_+ = \mathbf{t}_1(\boldsymbol{\pi})$ with $\mathbf{A}_0 = (\mathbf{I}_t \quad \mathbf{1}_t\mathbf{1}_t')$ and $\mathbf{C}_0 = (\mathbf{I}_t \quad -\mathbf{I}_t)'$. It follows that, if $\zeta(\boldsymbol{\pi})$ can be written using the "exp-log" notation, so can $\zeta^*(\boldsymbol{\pi})$.

A measure specified in terms of probabilities can always be rewritten so that the measure is a homogeneous function. Many measures are already naturally written as a homogeneous function. For instance, the logit function $\zeta(\boldsymbol{\pi}) = \pi_1/\pi_2$ is clearly homogeneous.

4.3.3 Marginal probabilities

Consider an $I \times J$ table with expected probabilities π_{ij}. Let $\boldsymbol{\pi}$ be the vector with elements π_{ij}, ordered with the last index changing fastest. That is,

$$\boldsymbol{\pi}' = (\pi_{11}, \ldots, \pi_{1J}, \pi_{21}, \ldots, \pi_{IJ}).$$

In the sequel, when this method of ordering indexed symbols is used, it will be said that the symbols are ordered with the last index changing fastest.

Below we will see how matrices $\mathbf{M}_r$ and $\mathbf{M}_c$ can be constructed so that $\mathbf{M}_r'\boldsymbol{\pi}$ consists of the row marginals π_{i+} and $\mathbf{M}_c'\boldsymbol{\pi}$ consists of the column marginals π_{+j}. For $i, k = 1, \ldots, I$ and $j, l = 1, \ldots, J$, let

$$q_{kl}(i) = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases} \qquad r_{kl}(j) = \begin{cases} 1 & l = j \\ 0 & l \neq j \end{cases}.$$

Then

$$\pi_{i+} = \sum_{k,l} q_{kl}(i)\pi_{kl} \qquad \pi_{+j} = \sum_{k,l} r_{kl}(j)\pi_{kl}.$$

Let $\mathbf{Q}(i)$ and $\mathbf{R}(j)$ be the matrices with (k,l)th elements $q_{kl}(i)$ and $r_{kl}(j)$ respectively. As an example, consider a 2×3 table, with expected probabilities $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23})'$. Then

$$\mathbf{Q}(1) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad \mathbf{Q}(2) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{R}(1) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \mathbf{R}(2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \mathbf{R}(3) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Let $\mathbf{q}_i$ be the vector with elements $q_{kl}(i)$, and let $\mathbf{r}_j$ be the vector with elements $r_{kl}(j)$, both ordered with the last index changing fastest. Now if $\mathbf{M}_r$ has ith column $\mathbf{q}_i$ and $\mathbf{M}_c$ has ith column $\mathbf{r}_j$, $\mathbf{M}_r'\boldsymbol{\pi}$ consists of the row totals π_{i+} and $\mathbf{M}_c'\boldsymbol{\pi}$ consists of the column totals π_{+j}. With $\mathbf{M} = (\mathbf{M}_r, \mathbf{M}_c)$, the row and column marginals are the elements of $\mathbf{M}'\boldsymbol{\pi}$. For a 2×3 table,

$$\mathbf{M}'\boldsymbol{\pi} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \end{pmatrix} = \begin{pmatrix} \pi_{1+} \\ \pi_{2+} \\ \pi_{+1} \\ \pi_{+2} \\ \pi_{+3} \end{pmatrix}.$$

4.3.4 The difference of proportions

Consider an $I \times J$ contingency table AB. The conditional probability of an observation being in column k, given it is in row i, is denoted as $\pi_{k|i}$. The conditional probabilities of response k for rows i and j can be compared using

$$\epsilon_{ijk} = \pi_{k|i} - \pi_{k|j} = \frac{\pi_{ik}}{\pi_{i+}} - \frac{\pi_{jk}}{\pi_{j+}}.$$

The variables A and B are independent if $\epsilon_{ijk} = 0$ for all i, j, and k, or equivalently, if $\epsilon_{ij_0k} = 0$ for all i and k with j_0 fixed. Using the "exp-log"

notation, the vector of conditional probabilities $\pi_{k|i}$ can be expressed as

$$\exp(\log \boldsymbol{\pi} - \mathbf{M}_r \log \mathbf{M}_r' \boldsymbol{\pi}) = \exp\left[\begin{pmatrix} \mathbf{I} & -\mathbf{M}_r \end{pmatrix} \log \begin{pmatrix} \mathbf{I} \\ \mathbf{M}_r' \end{pmatrix} \boldsymbol{\pi} \right],$$

where $\mathbf{M}_r$ is such that $\mathbf{M}_r' \boldsymbol{\pi}$ produces the row probabilities, as defined in 4.3.3.

4.3.5 Measures of association

For an $I \times J$ contingency table, association may be measured in many different ways. Odds ratios provide a set of numbers describing association in the table. Other measures summarize association by a single number. Below, two types of odds ratios, some summary measures of association for ordinal variables, and Pearson's correlation coefficient, which describes association for interval level variables, are presented. Agresti (1984, Chapter 9) gave an overview of some important measures of association for ordinal variables. Goodman and Kruskal (1979) republished some of their most important articles on measures of association in a book.

Local and global odds ratios

The (i, j)th local odds ratios $\zeta_{ij}^{(1)}$ and global odds ratios $\zeta_{ij}^{(2)}$ are defined as

$$\zeta_{ij}^{(1)} = \frac{\pi_{i,j}\, \pi_{i+1,j+1}}{\pi_{i,j+1}\, \pi_{i+1,j}} \qquad \zeta_{ij}^{(2)} = \frac{\left(\sum_{k \le i} \sum_{l \le j} \pi_{kl}\right) \left(\sum_{k > i} \sum_{l > j} \pi_{kl}\right)}{\left(\sum_{k \le i} \sum_{l > j} \pi_{kl}\right) \left(\sum_{k > i} \sum_{l \le j} \pi_{kl}\right)}, \quad (4.14)$$

for $1 \le i \le I - 1$ and $1 \le j \le J - 1$. Local odds ratios describe the relative magnitudes of "local" association in the table. For a 2×2 table, there is only one local odds ratio which is a measure for the overall association. Global odds ratios are designed for use with ordinal variables. They describe associations that are "global" in both variables. For more information on local, global, and other types of odds ratios, see Agresti (1984, p. 15–23).

Representing local and global odds ratios in the "exp-log" notation can be done as follows. Let $\boldsymbol{\zeta}^{(1)}(\boldsymbol{\pi})$ be the vector of local odds ratios and

$\boldsymbol{\zeta}^{(2)}(\boldsymbol{\pi})$ be the vector of global odds ratios. To write these vectors in the "exp-log" notation, define the following weights,

$$
\begin{array}{ll}
q_{kl}^{(1)}(i,j) = \begin{cases} 1 & k=i, l=j \\ 0 & \text{other cases} \end{cases} & q_{kl}^{(2)}(i,j) = \begin{cases} 1 & k \le i, l \le j \\ 0 & \text{other cases} \end{cases} \\
r_{kl}^{(1)}(i,j) = \begin{cases} 1 & k=i, l=j+1 \\ 0 & \text{other cases} \end{cases} & r_{kl}^{(2)}(i,j) = \begin{cases} 1 & k \le i, l > j \\ 0 & \text{other cases} \end{cases} \\
s_{kl}^{(1)}(i,j) = \begin{cases} 1 & k=i+1, l=j \\ 0 & \text{other cases} \end{cases} & s_{kl}^{(2)}(i,j) = \begin{cases} 1 & k > i, l \le j \\ 0 & \text{other cases} \end{cases} \\
t_{kl}^{(1)}(i,j) = \begin{cases} 1 & k=i+1, l=j+1 \\ 0 & \text{other cases} \end{cases} & t_{kl}^{(2)}(i,j) = \begin{cases} 1 & k > i, l > j \\ 0 & \text{other cases} \end{cases}
\end{array},
$$

for all i, j, k, and l such that $1 \le i \le I-1$, $1 \le j \le J-1$, $1 \le k \le I$, and $1 \le l \le J$. Let $\mathbf{Q}^{(h)}(i,j)$ ($h = 1, 2$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$) be the $I \times J$ matrix with (k,l)th element $q_{kl}^{(h)}(i,j)$, and define $\mathbf{R}^{(h)}(i,j)$, $\mathbf{S}^{(h)}(i,j)$, and $\mathbf{T}^{(h)}(i,j)$ analogously. To illustrate for a 2×3 table, as in Section 4.3.3, we get

$$
\begin{array}{ll}
\mathbf{Q}^{(1)}(1,1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \mathbf{Q}^{(2)}(1,1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\mathbf{Q}^{(1)}(1,2) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \mathbf{Q}^{(2)}(1,2) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\mathbf{R}^{(1)}(1,1) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \mathbf{R}^{(2)}(1,1) = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \\
\mathbf{R}^{(1)}(1,2) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} & \mathbf{R}^{(2)}(1,2) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \\
\mathbf{S}^{(1)}(1,1) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} & \mathbf{S}^{(2)}(1,1) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \\
\mathbf{S}^{(1)}(1,2) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} & \mathbf{S}^{(2)}(1,2) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \\
\mathbf{T}^{(1)}(1,1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} & \mathbf{T}^{(2)}(1,1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \\
\mathbf{T}^{(1)}(1,2) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \mathbf{T}^{(2)}(1,2) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.
\end{array}
$$

Let $\mathbf{q}_{ij}^{(h)}$, $h \in \{1,2\}$, be the $IJ \times 1$ vector with elements $q_{kl}^{(h)}(i,j)$, ordered with l changing faster than k, and let $\mathbf{Q}^{(h)}$ be the $IJ \times (I-1)(J-1)$ matrix with rows $\mathbf{q}_{ij}^{(h)}$, ordered with j changing faster than i. Define $\mathbf{R}^{(h)}$, $\mathbf{S}^{(h)}$, and $\mathbf{T}^{(h)}$ analogously. For a 2×3 table,

$$\mathbf{Q}^{(1)\prime} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{Q}^{(2)\prime} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$
$$\mathbf{R}^{(1)\prime} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \mathbf{R}^{(2)\prime} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$
$$\mathbf{S}^{(1)\prime} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad \mathbf{S}^{(2)\prime} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$
$$\mathbf{T}^{(1)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \mathbf{T}^{(2)\prime} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

For an $I \times J$ table, $\boldsymbol{\zeta}^{(h)}(\boldsymbol{\pi})$ $(h = 1,2)$ can be represented in the "exp-log" notation as

$$\boldsymbol{\zeta}^{(h)}(\boldsymbol{\pi}) = \exp\left[\begin{pmatrix} \mathbf{I}_v & -\mathbf{I}_v & -\mathbf{I}_v & \mathbf{I}_v \end{pmatrix} \log \begin{pmatrix} \mathbf{Q}^{(h)\prime} \\ \mathbf{R}^{(h)\prime} \\ \mathbf{S}^{(h)\prime} \\ \mathbf{T}^{(h)\prime} \end{pmatrix} \boldsymbol{\pi} \right], \qquad (4.15)$$

where $v = (I-1)(J-1)$ and $\mathbf{I}_v$ is the $v \times v$ identity matrix.

For a 2×3 table, we find

$$\mathbf{Q}^{(1)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{11} \\ \pi_{12} \end{pmatrix} \qquad \mathbf{Q}^{(2)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{11} \\ \pi_{11} + \pi_{12} \end{pmatrix}$$
$$\mathbf{R}^{(1)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{12} \\ \pi_{13} \end{pmatrix} \qquad \mathbf{R}^{(2)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{12} + \pi_{13} \\ \pi_{13} \end{pmatrix}$$
$$\mathbf{S}^{(1)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{21} \\ \pi_{22} \end{pmatrix} \qquad \mathbf{S}^{(2)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{21} \\ \pi_{21} + \pi_{22} \end{pmatrix}$$
$$\mathbf{T}^{(1)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{22} \\ \pi_{23} \end{pmatrix} \qquad \mathbf{T}^{(2)\prime}\boldsymbol{\pi} = \begin{pmatrix} \pi_{22} + \pi_{23} \\ \pi_{23} \end{pmatrix}.$$

With $v = (2-1)(3-1) = 2$, $\mathbf{I}_v = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and $h = 1$, (4.15) becomes

$$\begin{aligned}
\boldsymbol{\zeta}^{(1)}(\boldsymbol{\pi}) &= \exp\left[\begin{pmatrix} 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \end{pmatrix} \log \begin{pmatrix} \mathbf{Q}^{(1)\prime} \\ \mathbf{R}^{(1)\prime} \\ \mathbf{S}^{(1)\prime} \\ \mathbf{T}^{(1)\prime} \end{pmatrix} \boldsymbol{\pi}\right] \\
&= \exp\begin{bmatrix} \log \pi_{11} - \log \pi_{12} - \log \pi_{21} + \log(\pi_{22}) \\ \log \pi_{12} - \log \pi_{13} - \log \pi_{22} + \log(\pi_{23}) \end{bmatrix} \\
&= \begin{pmatrix} \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \\ \frac{\pi_{12}\pi_{23}}{\pi_{13}\pi_{22}} \end{pmatrix}.
\end{aligned}$$

With $h = 2$, (4.15) becomes

$$\begin{aligned}
\boldsymbol{\zeta}^{(2)}(\boldsymbol{\pi}) &= \exp\left[\begin{pmatrix} 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \end{pmatrix} \log \begin{pmatrix} \mathbf{Q}^{(2)\prime} \\ \mathbf{R}^{(2)\prime} \\ \mathbf{S}^{(2)\prime} \\ \mathbf{T}^{(2)\prime} \end{pmatrix} \boldsymbol{\pi}\right] \\
&= \exp\begin{bmatrix} \log \pi_{11} - \log(\pi_{12} + \pi_{13}) - \log \pi_{21} + \log(\pi_{22} + \pi_{23}) \\ \log(\pi_{11} + \pi_{12}) - \log \pi_{13} - \log(\pi_{21} + \pi_{22}) + \log \pi_{23} \end{bmatrix} \\
&= \begin{pmatrix} \frac{\pi_{11}(\pi_{22}+\pi_{23})}{\pi_{21}(\pi_{12}+\pi_{13})} \\ \frac{\pi_{23}(\pi_{11}+\pi_{12})}{\pi_{13}(\pi_{21}+\pi_{22})} \end{pmatrix},
\end{aligned}$$

which are the odds ratios in the collapsed tables

$$\left(\begin{array}{c|cc} \pi_{11} & \pi_{12} & \pi_{13} \\ \hline \pi_{21} & \pi_{22} & \pi_{23} \end{array}\right) \qquad \left(\begin{array}{cc|c} \pi_{11} & \pi_{12} & \pi_{13} \\ \hline \pi_{21} & \pi_{22} & \pi_{23} \end{array}\right).$$

Probabilities of concordance and discordance

For ordinal variables, many summary measures of association are based on the probabilities of a pair of observations being concordant or discordant. Suppose observations have been collected on a group of subjects on variables A and B. A pair of observations is *concordant* if the member that ranks higher on A also ranks higher on B, and *discordant* if the

member that ranks higher on A ranks lower on B. Not all pairs of observations are concordant or discordant. A *tie* occurs when two observations rank equally on at least one of the variables.

If the total number of observations is n_{++}, then the total number of pairs of observations is $n_{++}(n_{++}-1)/2$. Let C and D be the number of concordant and discordant pairs respectively, and T_A, T_B, and T_{AB} be the numbers of pairs that are tied on A, B, and both A and B, respectively. Then

$$\tfrac{1}{2}n_{++}(n_{++}-1) = C + D + T_A + T_B - T_{AB}.$$

In this formula, T_{AB} is subtracted because pairs tied on both A and B have been counted twice, once in T_A and once in T_B.

The probabilities of concordance and discordance Π_c and Π_d, that is, the probability that a pair of observations is concordant or discordant respectively, can be shown to be

$$\begin{aligned}
\Pi_c &= 2\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\pi_{ij}\left(\sum_{k>i}\sum_{l>j}\pi_{kl}\right) \\
\Pi_d &= 2\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\pi_{i,j+1}\left(\sum_{k>i}\sum_{l\leq j}\pi_{kl}\right).
\end{aligned}$$

The factor 2 occurs in these formulas because the first observation could be in cell (i,j) and the second in cell (k,l), or vice versa. We can see that Π_c and Π_d are sums of products of sums of probabilities, and can be written using the "exp-log" notation. The probabilities of a tie on A, B, and both A and B are

$$\Pi_{t,A} = \sum_i(\pi_{i+})^2 \qquad \Pi_{t,B} = \sum_j(\pi_{+j})^2 \qquad \Pi_{t,AB} = \sum_{i,j}\pi_{ij}^2.$$

How can the probabilities be written using the "exp-log" notation? Using the definitions of the $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{S}$, and $\mathbf{T}$ matrices as in Section 4.3.5, the probabilities of concordance and discordance can be expressed as

$$\begin{pmatrix}\Pi_c\\ \Pi_d\end{pmatrix} = \begin{pmatrix} 2\ \mathbf{1}_v'\exp(\log\mathbf{Q}^{(1)\prime}\boldsymbol{\pi} + \log\mathbf{T}^{(2)\prime}\boldsymbol{\pi}) \\ 2\ \mathbf{1}_v'\exp(\log\mathbf{R}^{(1)\prime}\boldsymbol{\pi} + \log\mathbf{S}^{(2)\prime}\boldsymbol{\pi})\end{pmatrix}$$

$$= 2(\mathbf{1}'_v \oplus \mathbf{1}'_v) \exp\left[\begin{pmatrix} \mathbf{I}_v & \mathbf{I}_v \end{pmatrix} \oplus \begin{pmatrix} \mathbf{I}_v & \mathbf{I}_v \end{pmatrix} \log \begin{pmatrix} \mathbf{Q}^{(1)\prime} \\ \mathbf{T}^{(2)\prime} \\ \mathbf{R}^{(1)\prime} \\ \mathbf{S}^{(2)\prime} \end{pmatrix} \boldsymbol{\pi} \right], \quad (4.16)$$

where $v = (I-1)(J-1)$ and "$\oplus$" represents the direct sum of matrices:

$$\mathbf{E}_1 \oplus \mathbf{E}_2 = \begin{pmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_2 \end{pmatrix}.$$

For a 2×3 table, (4.16) reduces to

$$\begin{aligned} \begin{pmatrix} \Pi_c \\ \Pi_d \end{pmatrix} &= 2 \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \exp\left[\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \log \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{22}+\pi_{23} \\ \pi_{23} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{21}+\pi_{22} \end{pmatrix} \right] \\ &= 2 \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \exp \begin{bmatrix} \log \pi_{11} + \log(\pi_{22} + \pi_{23}) \\ \log \pi_{12} + \log \pi_{23} \\ \log \pi_{12} + \log \pi_{21} \\ \log \pi_{13} + \log(\pi_{21} + \pi_{22}) \end{bmatrix} \\ &= 2 \begin{pmatrix} \pi_{11}(\pi_{22} + \pi_{23}) + \pi_{12}\pi_{23} \\ \pi_{12}\pi_{21} + \pi_{13}(\pi_{21} + \pi_{22}) \end{pmatrix}. \end{aligned}$$

With $\mathbf{M}_r$ and $\mathbf{M}_c$ as defined in Section 4.3.3, the probabilities of a tie are written as

$$\begin{aligned} \begin{pmatrix} \Pi_{t,A} \\ \Pi_{t,B} \\ \Pi_{t,AB} \end{pmatrix} &= \begin{pmatrix} \mathbf{1}'_I \exp\left(2\mathbf{I}_I \log \mathbf{M}'_r \boldsymbol{\pi}\right) \\ \mathbf{1}'_J \exp\left(2\mathbf{I}_J \log \mathbf{M}'_c \boldsymbol{\pi}\right) \\ \mathbf{1}'_{IJ} \exp\left(2\mathbf{I}_{IJ} \log \boldsymbol{\pi}\right) \end{pmatrix} \\ &= (\mathbf{1}'_I \oplus \mathbf{1}'_J \oplus \mathbf{1}'_{IJ}) \exp\left[2\,\mathbf{I}_{I+J+IJ} \log \begin{pmatrix} \mathbf{M}'_r \\ \mathbf{M}'_c \\ \mathbf{I}_{IJ} \end{pmatrix} \boldsymbol{\pi} \right]. \end{aligned}$$

Gamma

The association measure gamma is given by the formula

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

The range of values of gamma is $-1 \leq \gamma \leq 1$, with $\gamma = -1$ if $\Pi_c = 0$ and $\gamma = 1$ if $\Pi_d = 0$. If there is independence, gamma is zero, but the converse is not true. If gamma is zero, the probabilities of concordance and discordance are equal. An important aspect of gamma is that ties are not counted. For a 2×2 table, gamma simplifies to

$$\gamma = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}},$$

which is also referred to as Yule's Q.

Gamma can be written using the "exp-log" notation as

$$\begin{aligned}
\gamma &= \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d} = \frac{\Pi_c}{\Pi_c + \Pi_d} - \frac{\Pi_d}{\Pi_c + \Pi_d} \\
&= \exp(\log \Pi_c - \log(\Pi_c + \Pi_d)) - \exp(\log \Pi_d - \log(\Pi_c + \Pi_d)) \\
&= \begin{pmatrix} 1 & -1 \end{pmatrix} \exp\left[\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \log\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \Pi_c \\ \Pi_d \end{pmatrix} \right) \right]. \quad (4.17)
\end{aligned}$$

Since Π_c and Π_d can be written in "exp-log" notation, so can gamma, using the recursive definition (4.11) and (4.12). Gamma can be written in the form of (4.13) with $\mathbf{e}' = (1, -1)$.

If a marginal model of the form (4.3) is defined using gamma, it is recommended that the log-hyperbolic link function (4.4) be used because gamma has the same form as (4.5). If this link is used, a logit model is obtained for the ratio Π_c/Π_d, and out-of-range values cannot follow. In this way, the models described by Schollenberger, Agresti, and Wackerly (1979) can be generalized. Interestingly, for a 2×2 table, the log-hyperbolic link transforms gamma (or Yule's Q) to the log odds ratio.

Somers' d

A measure similar to gamma, but for which the pairs untied on A serve as the base rather than pairs untied on both A and B, was proposed by

Somers (1962). The population value of Somers' d is

$$\Delta_{BA} = \frac{\Pi_c - \Pi_d}{1 - \Pi_{t,A}}.$$

This expression is the difference between the proportions of concordant and discordant pairs out of the pairs that are untied on A. This is an asymmetric measure intended for use when B is a response variable.

Analogous to (4.17), Δ_{BA} can be written using the "exp-log" notation in the following way:

$$\Delta_{BA} = \begin{pmatrix} 1 & -1 \end{pmatrix} \exp \left[\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \log \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \Pi_c \\ \Pi_d \\ \mathbf{1}'\boldsymbol{\pi} \\ \Pi_{t,A} \end{pmatrix} \right],$$

where $\mathbf{1}'\boldsymbol{\pi} = \sum_i \pi_i = 1$ (this is done so that a function of $\boldsymbol{\pi}$ is obtained: "1" is not a function of $\boldsymbol{\pi}$). Like gamma, Somers' d can be written in the form of (4.13) with $\mathbf{e}' = (1, -1)$.

Kendall's tau and tau-b

Another variant of gamma was proposed by Kendall (1945). Its population version is

$$\tau_b = \frac{\Pi_c - \Pi_d}{\sqrt{(1 - \Pi_{t,A})(1 - \Pi_{t,B})}}$$

and is referred to as *Kendall's tau-b*.

If there are no ties the common value of gamma, Somers' d, and Kendall's tau-b is

$$\tau = \Pi_c - \Pi_d.$$

This measure is referred to as *Kendall's tau*, and was originally introduced for continuous variables.

Analogous to (4.17), Kendall's tau-b can be written in "exp-log" notation as

$$\tau_b = \begin{pmatrix} 1 & -1 \end{pmatrix} \cdot$$

$$\exp\left[\begin{pmatrix} 1 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \log \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \Pi_c \\ \Pi_d \\ \mathbf{1}'\boldsymbol{\pi} \\ \Pi_{t,A} \\ \Pi_{t,B} \end{pmatrix}\right].$$

Pearson's correlation coefficient

For two random variables A and B, Pearson's correlation coefficient rho is defined as

$$\rho(A,B) = \frac{\text{cov}(A,B)}{\sigma(A)\sigma(B)} = \frac{E(AB) - E(A)E(B)}{\sigma(A)\sigma(B)}.$$

Rho can be a useful measure of association for two interval level variables when these variables are linearly related.

In the "exp-log" notation, rho is written as

$$\rho(A,B) = \begin{pmatrix} 1 & -1 \end{pmatrix} \exp\left[\begin{pmatrix} 0 & 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 1 & 0 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \log \begin{pmatrix} E(A) \\ E(B) \\ E(AB) \\ \sigma^2(A) \\ \sigma^2(B) \end{pmatrix}\right].$$

The variances of A and B can be written as

$$\begin{pmatrix} \sigma^2(A) \\ \sigma^2(B) \end{pmatrix} = \begin{pmatrix} E(A^2) - E(A)^2 \\ E(B^2) - E(B)^2 \end{pmatrix}$$
$$= \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \exp\left[\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \log \begin{pmatrix} E(A) \\ E(A^2) \\ E(B) \\ E(B^2) \end{pmatrix}\right].$$

Suppose A has I categories with scores a_i and B has J categories with scores b_j, and let π_{ij} be the probability of an observation falling in the (i,j)th cell of table AB. Then, for example, $E(A) = \sum_i a_i \pi_{i+}$. Let $\mathbf{M}_r$ and $\mathbf{M}_c$ be such that $\mathbf{M}_r'\boldsymbol{\pi}$ and $\mathbf{M}_c'\boldsymbol{\pi}$ produce the row and column totals respectively. Let $\mathbf{a}$ and $\mathbf{a}^2$ be the vectors with elements a_i and a_i^2 respectively, define $\mathbf{b}$ and $\mathbf{b}^2$ analogously, and let $\mathbf{D}_{\mathbf{ab}'}$ be the diagonal

matrix with elements $a_i b_j$ on the main diagonal (with j changing fastest). Then the expected values that are used are:

$$\begin{pmatrix} E(A) \\ E(A^2) \\ E(B) \\ E(B^2) \\ E(AB) \end{pmatrix} = \begin{pmatrix} \sum a_i \pi_{i+} \\ \sum a_i^2 \pi_{i+} \\ \sum b_j \pi_{+j} \\ \sum b_j^2 \pi_{+j} \\ \sum a_i b_j \pi_{ij} \end{pmatrix} = \begin{pmatrix} \mathbf{a}' \mathbf{M}_r' \\ \mathbf{a}^{2\prime} \mathbf{M}_r' \\ \mathbf{b}' \mathbf{M}_c' \\ \mathbf{b}^{2\prime} \mathbf{M}_c' \\ \mathbf{1}' \mathbf{D}_{\mathbf{ab}'} \end{pmatrix} \boldsymbol{\pi}.$$

Thus, rho is a "sum of products of sums of products of sums" of probabilities, and can be written using the "exp-log" notation.

4.3.6 Measures of agreement

Suppose that both classifications in an $I \times I$ table have the same categories, listed in the same order. To assess the degree to which observations cluster on the main diagonal of the table, one can compare the probability $\Pi_o = \sum_i \pi_{ii}$ that an observation falls on the diagonal to the corresponding probability $\Pi_e = \sum_i \pi_{i+}\pi_{+i}$ that would be expected if the variables were independent. Cohen (1960) introduced the measure of agreement *kappa*, defined as

$$\kappa = \frac{\Pi_o - \Pi_e}{1 - \Pi_e} = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}}.$$

Kappa equals one when there is complete agreement and zero when there is independence.

A weighted version of kappa, introduced by Spitzer et al. (1967), can utilize the distance of cells from the main diagonal and is more useful for ordinal variables. With weights satisfying $0 \leq w_{ij} \leq 1$, the weighted agreement is $\Pi_o(\mathbf{w}) = \sum_{i,j} w_{ij}\pi_{ij}$ and weighted kappa is defined as

$$\kappa_w = \frac{\Pi_o(\mathbf{w}) - \Pi_e(\mathbf{w})}{1 - \Pi_e(\mathbf{w})} = \frac{\sum_{i,j} w_{ij}\pi_{ij} - \sum_{i,j} w_{ij}\pi_{i+}\pi_{+j}}{1 - \sum_{i,j} w_{ij}\pi_{i+}\pi_{+j}}.$$

Kappa is equal to weighted kappa with weights $w_{ij} = 0$ if $i \neq j$ and $w_{ii} = 1$ for all i. For weights $w_{ij} = 1 - (i-j)^2/(I-1)^2$, suggested by Fleiss and Cohen (1973), agreement is greater if more mass is in cells near the main diagonal.

How is kappa written for an $I \times I$ table in matrix notation? Let $\mathbf{M}_r$ and $\mathbf{M}_c$ be defined as in Section 4.3.3. Additionally, let $\mathbf{w}$ be the $I^2 \times 1$ vector with elements w_{ij}. Then

$$\begin{aligned}
\Pi_o(\mathbf{w}) &= \sum_{i,j} w_{ij}\pi_{ij} = \mathbf{w}'\boldsymbol{\pi} \\
\Pi_e(\mathbf{w}) &= \sum_{i,j} w_{ij}\pi_{i+}\pi_{+j} = \sum_{i,j} w_{ij} \exp(\log \pi_{i+} + \log \pi_{+j}) \\
&= \mathbf{w}' \exp\begin{pmatrix} \mathbf{M}_r & \mathbf{M}_c \end{pmatrix} \log \begin{pmatrix} \mathbf{M}_r' \\ \mathbf{M}_c' \end{pmatrix} \boldsymbol{\pi}.
\end{aligned}$$

To produce kappa from these probabilities, we can use

$$\kappa = \begin{pmatrix} 1 & -1 \end{pmatrix} \exp \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \log \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ \Pi_o(\mathbf{w}) \\ \Pi_e(\mathbf{w}) \end{pmatrix}.$$

4.4 Example: modelling political survey data

Consider the data in Table 4.1. They are obtained from a national sample of the Dutch electorate and pertain to the turnover in vote intention and preference for prime minister between February 1977 and March 1977. The observations in the two tables are dependent, as they involve the same people. The full cross-classification of observations, turnover table $ABCD$, is given in Table 4.2. Variables A and B represent vote intention in February and March, respectively, and variables C and D represent Preference for Prime Minister in February and March, respectively. Note that the cell frequencies in Table 4.1 are formed by the marginal frequencies of Table 4.2. Both variables Party Preference and Preference for Prime Minister are classified into three categories. The Christian Democratic Party is considered to be the political center, while observations on the category "other" consist mostly of voters for right-wing parties. Thus, the variables may be considered as ordinal.

Below, we examine in what respects turnover in party preference is similar to turnover in preference for prime minister. That is, the distributions of tables AB and CD are compared. When testing homogeneity hypotheses about AB and CD, it should be noted that the tables are dependent. Hagenaars (1990) considered which characteristic is most

(a) Vote Intention: February 1977 - March 1977

	B. March			
	Left	Chr. Dem.	Other	*Total*
A. February				
1. Left Wing	350	9	26	385
2. Christ. Dem.	17	242	25	284
3. Other	51	43	337	431
Total	418	294	388	1100

(b) Preference for Prime Minister: February 1977 - March 1977

	D. March			
	Left	Chr. Dem.	Other	*Total*
C. February				
1. Left Wing (Den Uyl)	410	16	49	475
2. Christ. Dem. (Van Agt)	19	111	42	172
3. Other	66	53	334	453
Total	495	180	425	1100

Table 4.1: Turnover in Political Preference in The Netherlands: February 1977 - March 1977 (source: Hagenaars, 1990)

		C	1	1	1	2	2	2	3	3	3	
		D	1	2	3	1	2	3	1	2	3	*Total*
A	B											
1	1		293	1	6	4	2	0	22	1	21	350
1	2		2	1	2	1	1	0	0	1	1	9
1	3		8	1	7	0	1	0	0	0	9	26
2	1		8	0	1	1	0	0	2	2	3	17
2	2		13	6	7	9	84	23	8	24	68	242
2	3		2	0	3	1	3	2	3	2	9	25
3	1		31	0	0	1	0	1	9	2	7	51
3	2		5	4	0	1	6	1	1	9	16	43
3	3		48	3	23	1	14	15	21	12	200	337
Total			410	16	49	19	111	42	66	53	334	1100

Note: The symbols A, B, C, and D have the same meaning as in Table 4.1

Table 4.2: Vote Intention and Preference Prime Minister in The Netherlands (source: Hagenaars, 1990)

Measure	*df*	G^2	X^2	p-value using G^2
frequencies	8	132.29	111.98	0.00
local odds ratios	4	30.86	30.75	0.00
global odds ratios	4	71.01	63.77	0.00
gamma	1	1.73	1.72	0.19
Π_c and Π_d	2	82.79	73.24	0.00

Table 4.3: Results of testing whether various measures are identical in tables AB and CD.

changeable, Party Preference or Preference for Prime Minister. Duncan (1979, 1981) suggested conditional testing procedures to test various hypotheses asserting a similarity between tables AB and CD. Hagenaars (1990, p. 169–176) applied these tests to various hypotheses about tables AB and CD. The conditional tests proposed by Duncan generalize the conditional test for marginal homogeneity described in Section 3.4.2. As mentioned there, one disadvantage of a conditional test is that the hypothesis conditioned on must be approximately true. Additionally, a conditional test does not yield expected cell frequencies. For these reasons, the maximum likelihood fitting and testing methods described in the next chapter, which are much more flexible than conditional testing procedures, are used below. It should be noted that a few models discussed below can be represented by constraints which are linear in the expected frequencies and that, therefore, the methods presented in Chapter 3 can also be used to test such models.

To start comparing tables AB and CD, the most restrictive hypothesis, that they are identically distributed, is tested. It can be seen in Table 4.3, that this model yields a bad fit. A conspicuous difference between tables AB and CD is that the marginal distributions are different: the marginal distribution of A differs from C, and B differs from D. Both in February and March, the number of people who preferred the socialist candidate Den Uyl is larger than the number of people who preferred a left-wing party, whereas the opposite is true for the Christian Democratic candidate Van Agt and the Christian Democratic Party. This reflects the "empirical regularity" that, in general, a Prime Minister in office, Den Uyl at that time, is more popular than his or her party, often even among those who do not support the Prime Minister's politics (Hagenaars, 1990,

p. 172). However, the dissimilarity between the marginal distributions A and C and B and D, respectively is not the only reason that the model asserting equality of AB and CD yields a bad fit. The adjusted residuals (see Section 5.4.3) presented in Table 4.4 indicate that a large portion of the dissimilarity of tables AB and CD is caused by the large difference of the observed values in the cells of AB and CD with index $(2, 2)$. (The adjusted residuals, which have a standard normal distribution if the model is correct, are 8.93 and -9.83, respectively.)

To investigate the similarity in turnover between AB and CD, the differences in popularity of political parties and their respective leaders can be disregarded. A weaker hypothesis than complete equality of the distributions of AB and CD is that the association in tables AB and CD is similar. Parsimonious models that can be used for testing this hypothesis are those which assert equality of odds ratios. The following two types of odds ratios can be used: local odds ratios and global odds ratios (see Section 4.3.5). The test results presented in Table 4.3 indicate strong lack of fit. A weaker model asserting homogeneity of association can be tested using the model specifying that gamma is identical in the two tables. This model fits well, though it is not very parsimonious with df=1. The stronger hypothesis that both the probabilities of concordance and discordance are equal in AB and CD was also tested, but does not fit well at all. The adjusted residuals indicate that this lack of fit should be ascribed to the large difference in the observed values of Π_c. Interestingly, homogeneity of gamma or homogeneity of Π_d both fit well, but the simultaneous model, which asserts that both Π_c and Π_d are equal in tables AB and CD does not fit well at all. This demonstrates that very different results can be obtained when fitting models simultaneously or separately.

A different type of question relates to differences in the one-dimensional marginal distributions A, B, C, and D. It is possible to test whether there has been no net change in vote intention, preference for prime minister, or both. It can be seen in Table 4.5, that the difference in the marginal distribution of A and B is significant, while the difference in the marginal distribution of C and D is not. There is no evidence for a net change in preference for prime minister. Another question is whether net changes in Party Preference are identical to net changes in vote intention. This question may be answered by testing whether the differences $A - B$ and $C - D$ are identical. It is reasonable to measure the differences using

	Table AB			Table CD		
index	observed	estimated	adj. res.	observed	estimated	adj. res.
Marginal frequencies						
(1,1)	350	387.43	-5.41	410	387.43	3.26
(1,2)	9	12.48	-1.45	16	12.48	1.46
(1,3)	26	39.25	-3.29	49	39.25	2.42
(2,1)	17	16.17	0.30	19	16.17	1.02
(2,2)	242	179.66	8.93	111	179.66	-9.83
(2,3)	25	31.35	-1.65	42	31.35	2.77
(3,1)	51	55.95	-1.02	66	55.95	2.07
(3,2)	43	47.44	-1.01	53	47.44	1.26
(3,3)	337	330.27	0.83	334	330.27	0.46
Local log odds ratios						
(1,1)	6.32	5.67	2.76	5.01	5.67	-2.32
(1,2)	-3.33	-2.77	-2.16	-2.09	-2.77	2.45
(2,1)	-2.83	-2.45	-1.98	-1.98	-2.45	1.90
(2,2)	4.33	3.60	4.91	2.81	3.60	-4.03
Global log odds ratios						
(1,1)	4.56	4.12	3.27	3.69	4.12	-3.35
(1,2)	2.65	2.71	-0.37	2.57	2.71	-1.15
(2,1)	2.20	2.42	-1.77	2.45	2.42	-0.23
(2,2)	3.77	3.33	3.81	2.84	3.33	-4.13
Gamma						
	0.86	0.84	1.30	0.83	0.84	-1.34
Probabilities of concordance and discordance						
Π_c	0.263	0.244	5.24	0.225	0.244	-5.76
Π_d	0.020	0.021	-0.82	0.021	0.021	0.05

Table 4.4: MLEs and adjusted residuals of various measures given homogeneity

Model	*df*	G^2	X^2	p-val.(G^2)
1. $A = B$	2	14.60	14.27	0.00
2. $C = D$	2	4.00	3.98	0.14
3. (1) $\cap$ (2)	4	16.36	15.92	0.00
4. $A - B = C - D$	3	1.82	1.81	0.61
5. (2) $\cap$ (a)*	3	5.83	6.01	0.12
6. (4) $\cap$ (a)*	4	4.29	4.22	0.37
7. (5) against (a)*	2	4.04	4.03	0.13
8. (6) against (a)*	3	2.59	2.57	0.30

Table 4.5: Test result for various models for tables AB and CD.
(*): (a) is the model asserting that gamma is identical in AB and CD (see Table 4.3).

odds (i.e., to test whether $P(A = i)/P(B = i) = P(C = i)/P(D = i)$, $i = 1, 2, 3$). According to Table 4.5, there is no reason to reject this hypothesis.

Apparently, the data do not allow a choice between the models asserting that there is no net change in preference for Prime Minister ($C = D$) and that the net changes in both turnover tables are identical ($A-B = C-D$), since both fit well but cannot simultaneously be true in the population because there is a significant change in Party Preference ($A \neq B$). The model specifying that gamma in table AB equals gamma in table CD simultaneously fitted with $C = D$ and $A - B = C - D$, respectively, yields a good fit. Thus, still no choice between the latter two models can be made. Finally, the conditional tests described in Section 2.5.3 can be used to test the models against the alternative that gamma is identical in the two tables. The test results presented in Table 4.5 shows that neither conditional test yields a significant result.

To conclude, the stability of vote intention and preference for prime minister do not differ significantly when measured in terms of gamma. Both the local and global odds ratios do differ significantly, however. Since the absolute values of the observed local log odds ratios in table AB are all greater than the corresponding local log odds ratios in table CD (see Table 4.4), there is evidence that vote intention is more stable than preference for prime minister.

Chapter 5

Fitting and testing marginal models

In this chapter, methods for fitting and testing the goodness-of-fit of the marginal models described in the previous chapter are presented. In Section 5.1, a computationally simple and efficient maximum likelihood fitting method is described. In Section 5.2, a regularity condition is given for loglinear models for sums of frequencies. Regularity of a model is sufficient to ensure that there is a unique vector of frequencies maximizing the likelihood subject to the model constraints. The weighted least squares method and generalized estimating equations approach are briefly presented in Section 5.3. Testing goodness-of-fit is considered in Section 5.4. Finally, asymptotic behaviour of MLEs is described in Section 5.5.

5.1 Maximum likelihood estimation

A standard approach to maximum likelihood fitting of models for categorical data involves solving the score equations using iterative methods such as Newton-Raphson or Fisher scoring. Such methods are very suitable for the loglinear models described in Chapter 2 (see Section 2.4), or for the models linear in the expected joint frequencies described in Chapter 3 (see Section 3.2.3). Several authors have proposed generalizations of this approach to classes of marginal models (Dale, 1986; McCullagh & Nelder, 1989, p. 216; Lipsitz et al., 1990; Becker & Balagtas, 1991; Molenberghs & Lesaffre, 1994; Glonek & McCullagh, 1995). A Fisher

scoring method can be used by reparameterization of the expected cell frequencies in terms of "marginal" model parameters. However, a serious drawback of this approach is that this reparameterization is typically computationally awkward and expensive. Additionally, it seems difficult to generalize the approach to the complete class of marginal models as defined in Section 4.1.

An alternative method, which has been discussed by Aitchison and Silvey (1958, 1960), utilizes the Lagrange multiplier method for constrained optimization. The model is viewed as inducing constraints on the expected cell frequencies, and the likelihood is maximized subject to these constraints. Haber (1985), Lang and Agresti (1994), and Lang (1996a) have shown how this approach can be applied to models for categorical data. In particular, these authors considered loglinear models for sums of frequencies, as defined by equation (4.10). A modified version of Aitchison and Silvey's method is given in Appendix A.2 . Below, this method is applied to marginal models.

With $\mathbf{n}$ an $r \times 1$ vector of observations and $\mathbf{m}$ an unknown $r \times 1$ vector of expected cell frequencies, which are all assumed to be strictly positive, the kernel of the log likelihood function is

$$\mathcal{L}(\mathbf{m}) = \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m}. \tag{5.1}$$

For the sampling distributions considered in Section 2.1, the vector of expected frequencies $\mathbf{m}$ is either unconstrained, in the case of Poisson sampling, or subject to the sampling constraint

$$\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}. \tag{5.2}$$

The marginal model specification, presented in Section 4.1, is

$$\mathbf{g}(\boldsymbol{\zeta}(\boldsymbol{\mu})) = \mathbf{X}\boldsymbol{\beta}, \tag{5.3}$$

where $\boldsymbol{\beta}$ is an unknown parameter vector, and $\boldsymbol{\mu} = \mathbf{M}'\mathbf{m}$ is a vector of marginal frequencies. It should be noted that the methods described below also apply when $\boldsymbol{\mu}$ is an arbitrary linear combination of frequencies, provided $\boldsymbol{\zeta}(\boldsymbol{\mu})$ is properly defined. A precise description of the model equation is given in the previous chapter. A maximum likelihood estimate (MLE) $\hat{\mathbf{m}}$ is sought which maximizes $\mathcal{L}$ as a function of $\mathbf{m}$ subject to the marginal model constraint (5.3) and, additionally, if the sampling scheme

is not Poisson, subject to the sampling constraint (5.2). ML estimation for Poisson sampling is simplest since there are no sampling constraints, and is therefore discussed first.

5.1.1 Estimation given Poisson sampling

Following Lang (1996a), optimization is done by reparameterizing the likelihood in terms of $\boldsymbol{\theta} = \log \mathbf{m}$ (all m_i being assumed to be positive). The advantage of this choice for $\boldsymbol{\theta}$ is that out-of-range iterate estimates are avoided.

Before the Lagrange multiplier technique is used, the model equation (5.3) is rewritten as a constraint equation for $\boldsymbol{\theta} = \log \mathbf{m}$, i.e., the vector of freedom parameters $\boldsymbol{\beta}$ is eliminated. Let $\mathbf{U}$ be the orthogonal complement of $\mathbf{X}$. Then (5.3) is equivalent to

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{U}'\mathbf{g}(\boldsymbol{\zeta}(\boldsymbol{\mu})) = \mathbf{0}. \tag{5.4}$$

It is assumed that $\mathbf{h}$ is differentiable. In the sequel, the derivative matrix of $\mathbf{h}$ is needed. Define

$$\mathbf{H} = \frac{\partial \mathbf{h}'}{\partial \boldsymbol{\theta}} \qquad \mathbf{G} = \frac{\partial \mathbf{g}(\boldsymbol{\zeta})'}{\partial \boldsymbol{\zeta}} \qquad \mathbf{Z} = \frac{\partial \boldsymbol{\zeta}(\boldsymbol{\mu})'}{\partial \boldsymbol{\mu}}. \tag{5.5}$$

Note that $\mathbf{G}$ is a diagonal matrix with elements $\partial g_i(\zeta_i)/\partial \zeta_i$ on the main diagonal. Using the chain rule for matrix derivatives, matrix $\mathbf{H}$ is given by the equation

$$\mathbf{H} = \mathbf{D_m M Z G U}.$$

Consider the Lagrangian log likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) &= \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m} + \boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\theta}) \\ &= \mathbf{n}'\boldsymbol{\theta} - \mathbf{1}' \exp(\boldsymbol{\theta}) + \boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\theta}). \end{aligned}$$

Differentiating L with respect to $\boldsymbol{\theta}$ and equating the result to zero yields

$$\mathbf{l}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{n} - \mathbf{m} + \mathbf{H}\boldsymbol{\lambda} = \mathbf{0}. \tag{5.6}$$

Let $\hat{\boldsymbol{\theta}}$ be a local maximum of $\mathcal{L}$ (see (5.1)) subject to the constraint $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$. A classical result (Bertsekas, 1982) is that if $\mathbf{H}(\hat{\boldsymbol{\theta}})$ is of full

column rank, there is a unique $\hat{\boldsymbol{\lambda}}$ such that $\mathbf{l}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}) = \mathbf{0}$. In the sequel, it is assumed that the MLE $\hat{\boldsymbol{\theta}}$ is a solution to the equations (5.4) and (5.6).

We propose using the updating function (A.14) for finding the MLE $\hat{\boldsymbol{\theta}}$. With

$$\mathbf{k} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \mathbf{n} - \mathbf{m} \qquad \mathbf{B} = E\left(-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = \mathbf{D_m},$$

equation (A.14) reduces to

$$\mathbf{v}(\boldsymbol{\theta}, step) = \boldsymbol{\theta} + step\, \mathbf{D_m^{-1}}(\mathbf{n} - \mathbf{m} + \mathbf{H}\boldsymbol{\lambda}(\boldsymbol{\theta})),$$

where

$$\boldsymbol{\lambda}(\boldsymbol{\theta}) = -(\mathbf{H}'\mathbf{D_m^{-1}}\mathbf{H})^{-1}(\mathbf{H}'\mathbf{D_m^{-1}}(\mathbf{n} - \mathbf{m}) + \mathbf{h}).$$

With appropriate starting values, for instance, $\boldsymbol{\theta}^{(0)} = \log(\mathbf{n})$, the algorithm is

$$\boldsymbol{\theta}^{(k+1)} = \mathbf{v}(\boldsymbol{\theta}^{(k)}, step^{(k)}) \qquad k = 0, 1, \ldots, \tag{5.7}$$

with appropriate values of $step^{(k)}$. Choosing the right step size is a difficult problem. For the algorithm described in Section 2.4.3 for finding MLEs for loglinear models, the step size could be chosen such that new iterate estimates would yield a higher value of the log likelihood. However, this approach cannot be used here since a saddle point of the Lagrangian log likelihood is sought.

Before proposing a method for choosing a step size, a termination criterion for the algorithm is presented. In order to be able to monitor convergence, an appropriate measure $e(\boldsymbol{\theta})$ for the distance of $\boldsymbol{\theta}$ from a solution $\hat{\boldsymbol{\theta}}$ to the likelihood equations may be chosen. The iterations can be stopped at iteration k when $e(\boldsymbol{\theta}^{(k)}) < \epsilon$, a sufficiently small constant chosen a priori (say $\epsilon = 10^{-10}$). As a measure for the distance $e(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ from $\hat{\boldsymbol{\theta}}$, the quadratic form

$$e(\boldsymbol{\theta}) = (\mathbf{v}(\boldsymbol{\theta}, 1) - \boldsymbol{\theta})'\mathbf{D_m}(\mathbf{v}(\boldsymbol{\theta}, 1) - \boldsymbol{\theta}) \tag{5.8}$$

can be used. Any solution $\hat{\boldsymbol{\theta}}$ to the equations (5.4) and (5.6) yields a global minimum of $e(\boldsymbol{\theta})$ such that $e(\hat{\boldsymbol{\theta}}) = 0$.

Since $\hat{\boldsymbol{\theta}}$ is a minimum of the "error function" $e(\boldsymbol{\theta})$, one can use $e(\boldsymbol{\theta})$ to choose a step size. If possible, the step size $step^{(k)}$ at iteration k is chosen

such that there is a decrease in the error function. One can start with $step^{(k)} = 1$, and keep halving its value while $e(\mathbf{v}(\boldsymbol{\theta}^{(k)}, step^{(k)}) > e(\boldsymbol{\theta}^{(k)})$. However, it is not always possible to obtain a decrease in the error because the iterative scheme (5.7) does not utilize the gradient of $e(\boldsymbol{\theta})$, while, additionally, $e(\boldsymbol{\theta})$ may be nonconvex. Thus, the error function can only give a rough indication over several iterations of whether we are on the right "track" to convergence. Choosing the right step size is often a question of trial and error.

If the MLE $\hat{\boldsymbol{\theta}}$ has been found, the MLE of the vector of model parameters can be calculated using the formula $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{g}(\boldsymbol{\zeta}(\hat{\boldsymbol{\mu}}))$, with $\hat{\boldsymbol{\mu}} = \mathbf{M}'\hat{\mathbf{m}} = \mathbf{M}'\exp(\hat{\boldsymbol{\theta}})$.

5.1.2 Estimation given sampling constraints

The above estimation method is designed to find Poisson estimates. Next, ML estimation when the expected cell frequencies are subject to the sampling constraint (5.2) is discussed. This sampling constraint can be included in the model constraint $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, though this is not necessary. Assume for the moment that full multinomial sampling is used, i.e., the sampling constraint is $\mathbf{1}'\mathbf{m} = \mathbf{1}'\mathbf{n}$. It was shown in Section 4.3.2 how $\mathbf{h}(\mathbf{m})$ can be written as a homogeneous function of $\mathbf{m}$, i.e., such that $\mathbf{h}(\mathbf{m}) = \mathbf{h}(\boldsymbol{\pi})$, with $\boldsymbol{\pi}$ the vector of cell probabilities. Using Result 5 in Appendix D, it follows that

$$\mathbf{H}'\mathbf{1} = \frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}'}\mathbf{1} = \frac{\partial \mathbf{h}}{\partial \mathbf{m}'}\mathbf{D}_{\mathbf{m}}\mathbf{1} = \frac{\partial \mathbf{h}}{\partial \mathbf{m}'}\mathbf{m} = \mathbf{0}.$$

Thus, premultiplying (5.6) by $\mathbf{1}'$ yields

$$\mathbf{1}'(\mathbf{n} - \mathbf{m}) + \mathbf{1}'\mathbf{H}\boldsymbol{\lambda} = \mathbf{1}'(\mathbf{n} - \mathbf{m}) = \mathbf{0}.$$

The sample size $\mathbf{1}'\mathbf{n}$ is automatically reproduced. Therefore, no separate fitting procedure is needed for multinomial sampling provided that $\mathbf{h}(\mathbf{m})$ is written as a homogeneous function. Similarly, it can be shown that the same result holds for any sampling restriction of the form (5.2), provided $\mathbf{h}(\mathbf{m})$ is written as a homogeneous function of the appropriate parameters.

5.1.3 Remarks on maximum likelihood estimation

The algorithm defined by (5.7) is a modified version of the algorithm provided by Aitchison and Silvey (1958, 1960, see Appendix A.1) because

the Lagrange multiplier vector $\boldsymbol{\lambda}$ has been "eliminated" from the iterative scheme that there is no need to keep track of its value during the iterative process. Lang (1996a) demonstrated how to apply Aitchison and Silvey's method to categorical data problems. The algorithm defined by (5.7) can also be viewed as a modification and generalization of Lang's algorithm. It should be noted that if the step size is equal to one for all iterations, then, with the same starting estimates, both Aitchison and Silvey's algorithm and the iterative scheme defined by (5.7) yield the same iterate estimates.

The question when to terminate the algorithm and how to choose a step size were not addressed by either Aitchison and Silvey or Lang. Above, a reasonable termination criterion was given, plus an indication of how to choose a step size. Choosing a step size remains troublesome for certain problems. For many problems, the iterative scheme (5.7) leads to convergence in relatively few iterations, even for large and highly sparse tables, and the number of iterations does not increase rapidly with the number of cells in the table. All estimates presented in the example in Section 4.4 were found in less than 100 iterations. In Appendix E, a listing is given of a Mathematica program for calculating the MLE $\hat{\mathbf{m}}$.

However, for certain problems, it may appear impossible to obtain convergence using iterative scheme (5.7). Bertsekas (1982) described various "Lagrangian methods" that can be used instead. A method with guaranteed convergence for a large class of loglinear models for marginal frequencies, described in Section 5.2, is suggested in Appendix B.3.

It may be noted that Haber (1985) described a different method than the one used here. He used a Newton-Raphson scheme for fitting loglinear models for sums of frequencies. A problem with this method is that the Hessian matrix to be inverted may become ill-conditioned during the iterative process. Aitchison and Silvey proposed instead that the expected value of the Hessian matrix be used, thereby avoiding the problems with ill-conditioning. This approach was also used by Lang; we use it as well.

Interestingly, if a marginal model is loglinear, the algorithm defined by (5.7) yields identical iterate estimates as the Newton-Raphson algorithm for loglinear models presented in Section 2.4. This is not the case for Aitchison and Silvey's method, unless a step size of one is used for all iterations. It should be further noted that if the marginal model to be fitted implies a certain loglinear model, then the sufficient statistics for the loglinear model are also sufficient statistics for the marginal model. In Chapter 2 it was shown that the sufficient statistics for loglinear mod-

els are reproduced. This is generally not the case for marginal models, unless the marginal model is equivalent to a loglinear model.

It seems that few results on the existence of MLEs for marginal models have been provided in the literature. Aitchison and Silvey gave regularity conditions for the log likelihood function and the constraint function which, if satisfied, guarantee that the MLEs are a solution to the Lagrangian likelihood equations with probability going to one as the sample size goes to infinity. The regularity conditions are satisfied in most practical cases. Lang (1996a) showed that the regularity conditions for the log likelihood and for the constraints induced by a class of loglinear models for sums of frequencies (see equation 4.10) are satisfied. The general small sample case remains unresolved however.

As a final remark, it is noted that the matrix which has to be inverted during the iterative process has a number of rows and columns equal to the number of degrees of freedom for that model. For large tables, if a very parsimonious model is specified, the matrix inversion may be a computational bottleneck. This can happen when a parsimonious loglinear model is specified. For instance, for ten dichotomous dependent variables, the loglinear model of no three or higher-order interaction has 946 degrees of freedom, so that a 946×946 matrix must be inverted. In such cases, computations are eased by optimizing in terms of the loglinear parameters, which are relatively few, rather than the log expected frequencies. (That is, the iterative method implied by (A.14) should be used with $\boldsymbol{\theta}$ equal to the parameters for the loglinear model.)

5.2 Uniqueness of MLEs for marginal models

In the previous section, an algorithm was presented for finding MLEs for marginal models. The problem with that algorithm (or any other algorithm) is that, upon convergence, it is difficult to determine whether a global maximum of the likelihood or a non-global local maximum has been reached. Below, a class of regular models is defined for which it is shown in Appendix B that models in this class have a unique solution $\hat{\mathbf{m}}$ maximizing the likelihood subject to the model constraints, provided all observed frequencies are constrained to be greater than zero. Regularity is only defined for loglinear models for sums of frequencies, and is a sufficient but not necessary condition for uniqueness of MLEs to be guaranteed for

all data. Below, examples are given of regular models as well as non-regular models with multiple solutions to the likelihood equations.

Various authors have shown that there is a unique solution $\hat{\mathbf{m}}$ maximizing the likelihood subject to the model constraints for different types of models for categorical data. Examples are loglinear models and models linear in expected frequencies, as demonstrated in Chapters 2 and 3. For cumulative logit models, uniqueness follows from results by Burridge (1981) and Pratt (1981). Haber and Brown (1986, Theorem 1) claim that models involving both linear and loglinear constraints have unique MLEs, but their proof is incorrect. In Section 5.2.2, a counterexample is given, and the mistake in their proof is pointed out.

5.2.1 Definition of regular models

With $\mathbf{m}$ an $r \times 1$ vector of variables which are constrained to be positive, consider a $z \times 1$ vector of measures $\boldsymbol{\zeta}(\mathbf{m})$ with kth element

$$\zeta_k = \prod_j \left(\sum_i a_{ij} m_i \right)^{c_{jk}},$$

with $a_{ij} \geq 0$ and arbitrary c_{jk}. It is assumed that for all j, there is an i such that $a_{ij} > 0$. Let $\mathbf{A}$ and $\mathbf{C}$ be the matrices with elements a_{ij} and c_{jk}, respectively. Assuming $m_i > 0$ for all i, the vector $\boldsymbol{\zeta}(\mathbf{m})$ with kth element $\zeta_k(\mathbf{m})$ can be written in matrix notation as

$$\boldsymbol{\zeta}(\mathbf{m}) = \exp(\mathbf{C}' \log \mathbf{A}'\mathbf{m}). \tag{5.9}$$

First, regularity of a vector is defined, then this definition is used to define regularity of models. The vector $\boldsymbol{\zeta}(\mathbf{m})$ is said to be *regular* if, for all λ_k and $\mu_j > 0$, with $t_j = \sum_k c_{jk}\lambda_k$,

$$\sum_j \frac{a_{ij}}{\mu_j} t_j < 1 \ \forall i \quad \Rightarrow \quad \sum_j \frac{a_{ij}}{\mu_j} \max(0, t_j) < 1 \ \forall i. \tag{5.10}$$

It can be verified that a subvector of a regular vector of measures is regular, but when two regular vectors of measures are concatenated, regularity may be destroyed.

A loglinear model of the form

$$\log \boldsymbol{\zeta}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta} \tag{5.11}$$

is said to be regular if $\boldsymbol{\zeta}(\mathbf{m})$ is defined by (5.9) and regular. In Appendix B, it is proven that regular models of the form (5.11) have a unique solution maximizing the likelihood subject to the model constraints.

It seems difficult to give an intuitive interpretation of the regularity condition. In some of the examples of non-regular models given below, the model constraints can be factored in the positive domain (that is, the domain where all expected frequencies $m_i > 0$), in the sense that the model can be written as the union of two or more distinct models. For instance, a constraint of the form $m_1^2 - 3m_1m_2 + 2m_2^2 = 0$ can be factored in the positive domain as $(m_1 - m_2)(m_1 - 2m_2) = 0$, so that either $m_1 = m_2$ or $m_1 = 2m_2$. Clearly, if the constraints defining a model can be factored, one would expect a solution maximizing the likelihood for each of the component models. Therefore, such models cannot be regular.

5.2.2 Examples of regular models

The following lemma will be used in the examples.

Lemma 1 *Suppose certain constants t_i ($i = 1, \ldots, I$) are such that $t_+ = 0$. Then, for all i and $\mu_i > 0$, and a certain i^*,*

$$\frac{\max(0, t_i)}{\mu_i} \leq \frac{t_{i^*}}{\mu_{i^*}}.$$

Proof Since $t_+ = 0$, there must be some k such that $t_k \geq 0$. Therefore, for a certain i^*,

$$\frac{\max(0, t_i)}{\mu_i} \leq \max_i \left(\frac{t_i}{\mu_i}\right) = \frac{t_{i^*}}{\mu_{i^*}}.$$

□

Loglinear models

Loglinear models are regular models; this can be shown as follows. Suppose a certain loglinear model is defined by the equation $\log \boldsymbol{\zeta}(\mathbf{m}) = \mathbf{C}' \log \mathbf{m} = \mathbf{X}\boldsymbol{\beta}$, with $\mathbf{C}'$ a contrast matrix. Thus, the model is of the form (5.11), with $\mathbf{A}$ the identity matrix (i.e., $a_{ii} = 1$ for all i, and $a_{ij} = 0$

if $i \neq j$). It is sufficient to show that (5.10) holds for $\boldsymbol{\zeta}(\mathbf{m})$, i.e., that $\boldsymbol{\zeta}(\mathbf{m})$ is regular. For a certain $\boldsymbol{\lambda}$, let $\mathbf{t} = \mathbf{C}\boldsymbol{\lambda}$. Since $\mathbf{C}'$ is a contrast matrix, $t_+ = \mathbf{1}'\mathbf{t} = \mathbf{1}'\mathbf{C}\boldsymbol{\lambda} = 0$, so there is a k such that $t_k \geq 0$. Assume that for arbitrary $\mu_i > 0$

$$\sum_j \frac{a_{ij}}{\mu_j} t_j = \frac{t_i}{\mu_i} < 1 \qquad \forall i$$

(note that $a_{ii} = 1$ and $a_{ij} = 0$ for $i \neq j$). It follows that, using Lemma 1, and for certain i^*,

$$\sum_j \frac{a_{ij}}{\mu_j} \max(0, t_j) = \frac{\max(0, t_i)}{\mu_i} \leq \frac{t_{i^*}}{\mu_{i^*}} < 1 \qquad \forall i.$$

Thus, (5.10) holds so $\boldsymbol{\zeta}(\mathbf{m}) = \exp(\mathbf{C}' \log \mathbf{m})$ is regular, and, by definition, the loglinear model defined by $\log \boldsymbol{\zeta}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta}$ is also regular.

Models for one-dimensional marginal frequencies

Consider an $I \times J$ table with expected frequencies m_{ij}. A vector $\boldsymbol{\zeta}(\mathbf{m})$ with any of the odds m_{i+}/m_{k+} and m_{+j}/m_{+l} as elements is regular. It follows that loglinear models for these odds are regular. The result is generalizable to higher-way tables.

More generally, for $I \times J$ tables, let $\mathbf{A}_1$ be such that $\mathbf{A}_1'\mathbf{m}$ has elements m_{i+} and let $\mathbf{A}_2$ be such that $\mathbf{A}_2'\mathbf{m}$ has elements m_{+j}. Consider a vector of measures of the form

$$\boldsymbol{\zeta}(\mathbf{m}) = \exp\left[\begin{pmatrix} \mathbf{C}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2' \end{pmatrix} \log \begin{pmatrix} \mathbf{A}_1' \\ \mathbf{A}_2' \end{pmatrix} \mathbf{m}\right], \tag{5.12}$$

where $\mathbf{C}_1'$ and $\mathbf{C}_2'$ are contrast matrices. For instance, in a 2×2 table, the marginal odds m_{1+}/m_{2+} and m_{+1}/m_{+2} can be represented in the form (5.12) as

$$\exp\left[\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \log \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} m_{11} \\ m_{12} \\ m_{21} \\ m_{22} \end{pmatrix}\right].$$

Vector $\boldsymbol{\zeta}(\mathbf{m})$ is regular, which can be demonstrated as follows. For a certain arbitrary $\boldsymbol{\lambda}$, let

$$\mathbf{u} = \mathbf{C}_1\boldsymbol{\lambda} \qquad \mathbf{v} = \mathbf{C}_2\boldsymbol{\lambda}.$$

Since $\mathbf{C}_1'$ and $\mathbf{C}_2'$ are contrast matrices, it follows that $u_+ = v_+ = 0$. Using different indices i and j than in (5.10), the left hand side of the equation reduces to

$$\frac{u_i}{\mu_{1i}} + \frac{v_j}{\mu_{2j}} < 1 \quad \forall i, j. \tag{5.13}$$

Assuming that (5.13) is true and using Lemma 1, it follows that, for certain i^* and j^*,

$$\frac{\max(0, u_i)}{\mu_{1i}} + \frac{\max(0, v_j)}{\mu_{2j}} \quad \leq \quad \left(\frac{u_{i^*}}{\mu_{1i^*}}\right) + \left(\frac{v_{j^*}}{\mu_{2j^*}}\right) \quad < \quad 1 \quad \forall i, j.$$

Thus, (5.10) is satisfied, and it follows that a model defined as $\log \boldsymbol{\zeta}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta}$, with $\boldsymbol{\zeta}(\mathbf{m})$ defined by (5.12), is regular. The argument can be generalized to models for one-dimensional marginals of higher-way tables.

Models for two-dimensional marginal frequencies

Consider an $I \times J \times K$ table with expected frequencies m_{ijk}. A vector $\boldsymbol{\zeta}(\mathbf{m})$ containing any of the odds $m_{i_1 j +}/m_{i_2 j +}$ or m_{+jk_1}/m_{+jk_2} as elements is regular. It follows that loglinear models for these odds are regular.

More generally, for an $I \times J \times K$ table, let $\mathbf{A}_1$ be such that $\mathbf{A}_1'\mathbf{m}$ is a vector with elements m_{ij+} (in any order) and let $\mathbf{A}_2$ be such that $\mathbf{A}_2'\mathbf{m}$ has elements m_{+jk} (also in any order). Consider a vector of measures of the form

$$\boldsymbol{\zeta}(\mathbf{m}) = \exp\left[\begin{pmatrix} \mathbf{C}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2' \end{pmatrix} \log \begin{pmatrix} \mathbf{A}_1' \\ \mathbf{A}_2' \end{pmatrix} \mathbf{m}\right], \tag{5.14}$$

where $\mathbf{C}_1'$ and $\mathbf{C}_2'$ are contrast matrices. Then $\boldsymbol{\zeta}(\mathbf{m})$ is regular, as shown below.

For a certain arbitrary $\boldsymbol{\lambda}$, let $\mathbf{u} = \mathbf{C}_1\boldsymbol{\lambda}$ and $\mathbf{v} = \mathbf{C}_2\boldsymbol{\lambda}$. Note that for all j, $u_{+j} = v_{j+} = 0$. Using different indices i, j, and k than in (5.10), the left hand side of (5.10) reduces to

$$\frac{u_{ij}}{\mu_{1ij}} + \frac{v_{jk}}{\mu_{2jk}} < 1 \quad \forall i, j, k. \tag{5.15}$$

Assuming that (5.15) is true and using Lemma 1, it follows that, for certain i_j^* and k_j^*,

$$\begin{aligned}
\frac{\max(0,u_{ij})}{\mu_{1ij}}+\frac{\max(0,v_{jk})}{\mu_{2jk}} &\leq \max_i \frac{\max(0,u_{ij})}{\mu_{1ij}}+\max_k \frac{\max(0,v_{jk})}{\mu_{2jk}} \\
&\leq \frac{\max(0,u_{i_j^*j})}{\mu_{1i_j^*j}}+\frac{\max(0,v_{jk_j^*})}{\mu_{2jk_j^*}} \\
&= \frac{u_{i_j^*j}}{\mu_{1i_j^*j}}+\frac{v_{jk_j^*}}{\mu_{2jk_j^*}} \\
&< 1 \quad \forall i,j,k.
\end{aligned}$$

Thus, (5.10) is satisfied, and it follows that a model defined as $\log \boldsymbol{\zeta}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta}$, with $\boldsymbol{\zeta}(\mathbf{m})$ defined by (5.14), is regular.

The argument can be generalized to models for two-dimensional marginals of higher-way tables, though with caution. If odds for the marginals m_{i+k} are added to $\boldsymbol{\zeta}(\mathbf{m})$, regularity is destroyed. In a four-way table $ABCD$, one might wish to model odds in the marginal tables AB, BC, and CD, with expected frequencies m_{ij++}, m_{+jk+}, and m_{++kl}, respectively. In such a case, using a similar derivation, $\boldsymbol{\zeta}(\mathbf{m})$ can be shown to be regular when it has elements odds and odds *ratios* of the form

$$\frac{m_{i_1j++}}{m_{i_2j++}} \qquad \frac{m_{+j_1k_1+}}{m_{+j_1k_2+}} \times \frac{m_{+j_2k_2+}}{m_{+j_2k_1+}} \qquad \frac{m_{++kl_1}}{m_{++kl_2}}.$$

Note that, for the marginal table BC odds ratios must be taken, instead of odds as for the tables AB and CD. Again, if odds from any other two-dimensional marginal table (such as AD) are added to $\boldsymbol{\zeta}(\mathbf{m})$, regularity is destroyed.

Models involving conditional probabilities

In many cases, it may be unclear whether a certain model of the form (5.11) is regular or not. Sometimes, however, it is possible to rewrite the model constraints in such a way that a loglinear model for a regular vector of measures is obtained.

As an example, consider a four-way table $ABCD$ with cell probabilities π_{ijkl}; the conditional probability distribution of A given B is given by

$$P(A=i|B=j)=\frac{\pi_{ij++}}{\pi_{+j++}}.$$

Suppose we wish to test whether the probability of A given B is identical to the probability of C given D. The model implied by the constraints

$$\frac{\pi_{ij++}}{\pi_{+j++}} = \frac{\pi_{++ij}}{\pi_{+++j}} \qquad \forall i, j. \tag{5.16}$$

may be used. Both one- and two-dimensional marginals are involved in the constraints, so the results of the previous two sections cannot be used. However, it is possible to rewrite the constraints. One can verify that (5.16) is equivalent to

$$\frac{\pi_{i_1 j++}}{\pi_{i_2 j++}} = \frac{\pi_{++i_1 j}}{\pi_{++i_2 j}} \qquad \forall i_1, i_2, j.$$

This equation is loglinear in odds in the marginal tables AB and CD, so the corresponding model is regular, as was shown above.

5.2.3 Examples of non-regular models with multiple local maxima

In certain cases, the methods described above cannot be used to decide whether a model of the form (5.11) is regular. It may be helpful to see some examples of non-regular models that have multiple local maxima. In the first two examples presented below, there are almost always two local maxima, whatever the data. Only in certain special cases do these maxima "overlap". The first two examples are easy to solve analytically; the last is not.

Example 1. Consider a 2×2 table AB, with expected probabilities $\pi_{ij} > 0$. Suppose one wishes to model independence simultaneously with the probability of agreement $\sum_i \pi_{ii}$ being equal to the probability of disagreement $\sum_{i \neq j} \pi_{ij}$, i.e.,

$$\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1 \qquad \wedge \qquad \pi_{11} + \pi_{22} = \pi_{12} + \pi_{21} \tag{5.17}$$

(assuming all $\pi_{ij} > 0$). From straightforward calculations, it follows that these equations are equivalent to

$$\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1 \qquad \wedge \qquad (\pi_{11} - \pi_{12})(\pi_{11} - \pi_{21}) = 0.$$

The equation on the right is factored, yielding two solutions

$$\pi_{11} = \pi_{12} \qquad \vee \qquad \pi_{11} = \pi_{21}.$$

In a 2×2 contingency table, the two solutions to (5.17) can be represented as

π	π
$\frac{1}{2} - \pi$	$\frac{1}{2} - \pi$

π	$\frac{1}{2} - \pi$
π	$\frac{1}{2} - \pi$

for an unknown $\pi \in \langle 0, \frac{1}{2}\rangle$. Each of the two possible solution sets yields a loglinear model. One has

$$\log \pi_{ij} = \lambda + \lambda_i^A \qquad \vee \qquad \log \pi_{ij} = \lambda + \lambda_j^B,$$

i.e., the simultaneous model reduces to the *union* of two loglinear models. Each loglinear model has a unique MLE, so the union of the models has two solutions to the likelihood equations.

Closed form expressions exist for the MLEs. The local maxima of the likelihood function are

$$\hat{\pi}_{11} \;=\; \hat{\pi}_{12} \;=\; \tfrac{1}{2}(p_{11} + p_{12}) \qquad \hat{\pi}_{21} \;=\; \hat{\pi}_{22} \;=\; \tfrac{1}{2}(p_{21} + p_{22})$$

and

$$\hat{\pi}_{11} \;=\; \hat{\pi}_{21} \;=\; \tfrac{1}{2}(p_{11} + p_{21}) \qquad \hat{\pi}_{12} \;=\; \hat{\pi}_{22} \;=\; \tfrac{1}{2}(p_{12} + p_{22}).$$

This is, in fact, a counterexample to Theorem 1 of Haber and Brown's (1986) article, which stated that the likelihood equations have a unique solution when the expected frequencies are simultaneously subject to both linear and loglinear constraints. (On page 478 of the article, a vector $\mathbf{r}$ was defined and it was implicitly assumed, in the last three lines of the proof of Theorem 1 that the estimate $\hat{\mathbf{r}}$ has only strictly positive elements. That assumption is, in general, incorrect.)

Example 2. Suppose one wishes to model equality of variance of the marginal distributions A and B of a 2×2 table AB with expected probabilities π_{ij}. The hypothesis of equality of variance may be expressed by the constraint equation

$$\pi_{1+}\pi_{2+} = \pi_{+1}\pi_{+2}. \tag{5.18}$$

Straightforward calculations show that the constraint (5.18) reduces to

$$(\pi_{12} - \pi_{21})(\pi_{11} - \pi_{22}) = 0,$$

i.e., either $\pi_{12} = \pi_{21}$, or $\pi_{11} = \pi_{22}$. One local maximum of the likelihood occurs at the MLE $\hat{\pi}_{11} = p_{11}$, $\hat{\pi}_{12} = \hat{\pi}_{21} = \frac{1}{2}(p_{12} + p_{21})$, and $\hat{\pi}_{22} = p_{22}$, the other local maximum at the MLE $\hat{\pi}_{11} = \hat{\pi}_{22} = \frac{1}{2}(p_{11} + p_{22})$, $\hat{\pi}_{12} = p_{12}$, and $\hat{\pi}_{21} = p_{21}$.

Example 3. Consider a three-way table ABC. It was shown above that a loglinear model specified for marginal odds ratios of the two-dimensional marginal tables AB and BC is regular. However, if, a loglinear model is also specified for odds ratios in the marginal table AC, local maxima may occur. For instance, consider a $2 \times 2 \times 2$ table with expected frequencies m_{ijk}. The model defined by the constraints

$$\frac{m_{11+}m_{22+}}{m_{12+}m_{21+}} = \frac{m_{+11}m_{+22}}{m_{+12}m_{+21}} = 4 \qquad \frac{m_{1+1}m_{2+2}}{m_{1+2}m_{2+1}} = \tfrac{1}{4} \tag{5.19}$$

was found to yield two local maxima (Tamas Rudas, personal communication). Local maxima of the likelihood for two sets of observed frequencies are presented in Table 5.1. The (observed) frequencies in the table are arranged as

n_{111}	n_{112}
n_{121}	n_{122}

n_{211}	n_{212}
n_{221}	n_{222}

.

For the frequencies on the left in Table 5.1, both local maxima yield $G^2 = 5.71$, i.e., the MLE cannot be identified. The frequencies on the right yield $G^2 = 39.15$ for the first local maximum, and $G^2 = 26.24$ for the second local maximum, i.e., the second maximum is the global maximum of the log likelihood. In Tables 5.2 and 5.3, the observed and estimated marginal frequencies corresponding to the frequencies of Table 5.1 are presented.

5.3 Alternatives to maximum likelihood

Two alternatives to maximum likelihood which have received considerable attention in the literature are discussed below. The *weighted least*

Observed frequencies:

1	1	1	1
1	1	1	1

Observed frequencies:

1	2	7	11
3	5	13	17

Local maximum 1:

0.52	1.98	0.14	0.52
1.98	0.37	0.52	1.98

Local maximum 1:

0.62	5.43	0.90	6.60
6.50	1.13	8.78	29.04

Local maximum 2:

1.98	0.52	0.37	1.98
0.52	0.14	1.98	0.52

Local maximum 2:

3.75	0.98	2.55	16.42
1.32	0.75	19.56	13.66

Table 5.1: Observed frequencies and local maxima of the log likelihood subject to (5.19)

Observed two-way marginals:

2	2	2	2	2	2
2	2	2	2	2	2

Local maximum 1:

2.49	2.35	0.66	2.49	2.49	2.35
0.66	2.49	2.49	2.35	0.66	2.49

Local maximum 2:

2.49	0.66	2.35	2.49	2.49	0.66
2.35	2.49	2.49	0.66	2.35	2.49

Table 5.2: Observed marginals and local maxima of the log likelihood subject to (5.19) corresponding to the frequencies on the left-hand side of Table 5.1

Observed two-way marginals:

3	8	8	13	4	7
18	30	16	22	20	28

Local maximum 1:

6.05	7.62	1.52	12.03	7.12	6.55
7.50	37.82	15.28	30.16	9.68	35.64

Local maximum 2:

4.73	2.07	6.31	17.40	5.08	1.73
18.98	33.22	20.88	14.40	22.11	30.08

Table 5.3: Observed marginals and local maxima of the log likelihood subject to (5.19) corresponding to the frequencies on the right-hand side of Table 5.1

squares (WLS) method was popularized for categorical data analysis by Grizzle, Starmer, and Koch (1969). WLS is a non-iterative method yielding optimal estimates of the β parameters in the marginal model (5.3), and therefore the necessary computations can be done much faster than the ML computations. However, as the sample size increases, convergence of WLS estimates to the true parameter values is much slower than with ML. A more recent approach is the *generalized estimating equations* (GEE) method (Liang et al., 1992). This approach was developed in part because ML methods were considered infeasible (Liang et al., 1992, p. 9) for many problems. However, it appears that ML is preferred by many statisticians (see, for instance, the discussion of the paper by Liang, Zeger & Qakish).

5.3.1 Weighted least squares

The weighted least squares (WLS) approach can be used to estimate model parameters $\boldsymbol{\beta}$ for any of the marginal models described in Section (4.1). With this approach, the fact is employed that, under appropriate regularity conditions, a differentiable measure has an asymptotic normal distribution. Since the approach was popularized in a paper by

Grizzle, Starmer & Koch (1969) it is also called the *GSK* method in their honour.

Suppose the vector of expected marginal frequencies $\boldsymbol{\mu} = \mathbf{M}'\mathbf{m}$ has observed value $\mathbf{y}$. Let $\boldsymbol{\eta}(\boldsymbol{\mu}) = \mathbf{g}(\boldsymbol{\zeta}(\boldsymbol{\mu}))$, so that the marginal model (5.3) can be written as $\boldsymbol{\eta}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. Let $\boldsymbol{\Sigma}(\boldsymbol{\eta}(\mathbf{y}))$ be the asymptotic covariance matrix of $\boldsymbol{\eta}(\mathbf{y})$, which is given by the formula

$$\boldsymbol{\Sigma}(\boldsymbol{\eta}(\mathbf{y})) = \mathbf{G}'\mathbf{Z}'\boldsymbol{\Sigma}(\mathbf{y})\mathbf{Z}\mathbf{G},$$

where $\mathbf{G}$ and $\mathbf{Z}$ are defined in (5.5) and where the covariance matrix of the observed marginals $\mathbf{y}$ is

$$\boldsymbol{\Sigma}(\mathbf{y}) = \mathbf{M}'\mathbf{D}_\mathbf{m}\mathbf{M}. \tag{5.20}$$

Assuming that $\bar{\boldsymbol{\Sigma}}(\boldsymbol{\eta}(\mathbf{y}))$, the sample value of $\boldsymbol{\Sigma}(\boldsymbol{\eta}(\mathbf{y}))$, is invertible, the WLS estimate $\tilde{\boldsymbol{\beta}}$ of the true parameter value minimizes the quadratic form

$$W^2 = (\boldsymbol{\eta}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta})'\bar{\boldsymbol{\Sigma}}(\boldsymbol{\eta}(\mathbf{y}))^{-1}(\boldsymbol{\eta}(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}), \tag{5.21}$$

The *weight* matrix in the quadratic form is $\bar{\boldsymbol{\Sigma}}(\boldsymbol{\eta}(\mathbf{y}))$. The WLS estimate $\tilde{\boldsymbol{\beta}}$ is

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\bar{\boldsymbol{\Sigma}}(\boldsymbol{\eta}(\mathbf{y}))^{-1}\mathbf{X})^{-1}\mathbf{X}'\bar{\boldsymbol{\Sigma}}(\boldsymbol{\eta}(\mathbf{y}))^{-1}\boldsymbol{\eta}(\mathbf{y}),$$

The statistic (5.21) with $\tilde{\boldsymbol{\beta}}$ substituted for $\boldsymbol{\beta}$ is the Wald statistic and has an asymptotic chi-squared distribution, with degrees of freedom equal to the number of linearly independent elements of $\boldsymbol{\eta}(\mathbf{y})$ minus the number of linearly independent β parameters.

An advantage of using WLS to estimate model parameters is that computational costs are low compared to ML, especially when many variables are used. ML methods require evaluation of all expected cell frequencies in the contingency table. Since this number increases exponentially with the number of variables, the complexity of ML also increases exponentially with the number of variables. WLS does not require evaluation of all cells, and its complexity increases much more slowly than ML with the number of cells.

However, there are some potentially serious disadvantages of WLS compared to ML. First, WLS is very sensitive to sparse data. In fact, for many models, almost all observed marginals should exceed about 5-10

(Agresti, Lipsitz, & Lang, 1992) to get reasonable estimates and a good approximation to the chi-squared distribution of W^2. Second, WLS does not usually yield estimated expected frequencies, but only estimates of the β parameters.

5.3.2 Generalized estimating equations

Consider a model $\boldsymbol{\eta}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ for a vector of expected frequencies $\boldsymbol{\mu}$ ($\boldsymbol{\mu} = \mathbf{M}'\mathbf{m}$), which has observed value $\mathbf{y}$. It is assumed that the elements of $\boldsymbol{\mu}$ are linearly independent. Suppose that $\boldsymbol{\eta}$ is (at least implicitly) invertible, so that $\boldsymbol{\mu} = \boldsymbol{\eta}^{-1}(\mathbf{X}\boldsymbol{\beta})$, and assume differentiability of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\beta}$. The *generalized estimating equation* (GEE) for estimating $\boldsymbol{\beta}$ is a multivariate analogue of the quasi-score function introduced by Wedderburn (1974) and has the form

$$\left(\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}}\right) \boldsymbol{\Sigma}(\mathbf{y})^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \tag{5.22}$$

(Liang & Zeger, 1986; Diggle et al., 1994, p. 149). For a loglinear model for marginal frequencies, specified as $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\partial \boldsymbol{\mu}'/\partial \boldsymbol{\beta} = \mathbf{X}'\mathbf{D}_{\boldsymbol{\mu}}^{-1}$, and equation (5.22) reduces to

$$\mathbf{X}'\mathbf{D}_{\boldsymbol{\mu}}^{-1}\boldsymbol{\Sigma}(\mathbf{y})^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}.$$

If the covariance matrix of the marginal frequencies $\boldsymbol{\Sigma}(\mathbf{y})$ is known, equation (5.22) is solvable for $\boldsymbol{\beta}$. In practice, $\boldsymbol{\Sigma}(\mathbf{y})$ is usually unknown, and several methods have been proposed to estimate it (Liang et al., 1992; Diggle et al., 1994). In most cases, however, this approach does not yield a true statistical model based on a probability distribution (Lindsey, 1993, Section 2.9). When the covariances are misspecified, efficiency is usually lost.

The simplest method of estimating the covariances, apparently not considered by Diggle, Liang, and Zeger, is to use the observed value of $\boldsymbol{\Sigma}(\mathbf{y})$, i.e., to use

$$\bar{\boldsymbol{\Sigma}}(\mathbf{y}) = \mathbf{M}'\mathbf{D}_{\mathbf{n}}\mathbf{M}.$$

This choice gives optimal estimates of $\boldsymbol{\beta}$. When $\boldsymbol{\eta}(\boldsymbol{\mu}) = \boldsymbol{\mu}$, i.e., for models linear in the expected frequencies, WLS estimates are obtained for $\boldsymbol{\beta}$. If the MLE $\hat{\boldsymbol{\Sigma}}(\mathbf{y})$ of $\boldsymbol{\Sigma}(\mathbf{y})$ is used, the estimated parameter value is the MLE $\hat{\boldsymbol{\beta}}$ (Fitzmaurice et al., 1993). However, to find the $\hat{\boldsymbol{\Sigma}}(\mathbf{y})$, the complete maximum likelihood estimation procedure has to be performed.

5.4 Assessing model goodness-of-fit

5.4.1 Chi-squared tests

The chi-squared statistics described in Section 2.5 can be used to test goodness-of-fit. An alternative chi-squared statistic was proposed by Aitchison and Silvey (1958, 1960) and Silvey (1959). This is the Lagrange multiplier statistic L^2, defined as

$$L^2 = \hat{\boldsymbol{\lambda}}' \boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}})^{-1} \hat{\boldsymbol{\lambda}}.$$

It can be shown that the Pearson chi-squared and Lagrange multiplier statistics are numerically equal when they are evaluated at the maximum likelihood estimates. From the likelihood equation (5.6), $\mathbf{n} - \hat{\mathbf{m}} = -\hat{\mathbf{H}}\hat{\boldsymbol{\lambda}}$, and since $\boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}}) = (\hat{\mathbf{H}}' \mathbf{D}_{\hat{\mathbf{m}}}^{-1} \hat{\mathbf{H}})^{-1}$ (see (A.4)), we obtain

$$X^2 = (\mathbf{n} - \hat{\mathbf{m}})' \mathbf{D}_{\hat{\mathbf{m}}}^{-1} (\mathbf{n} - \hat{\mathbf{m}}) = \hat{\boldsymbol{\lambda}} \hat{\mathbf{H}}' \mathbf{D}_{\hat{\mathbf{m}}}^{-1} \hat{\mathbf{H}} \hat{\boldsymbol{\lambda}} = L^2.$$

The number of degrees of freedom for a marginal model is equal to the number of functionally independent constraints, which is equal to the column rank of $\mathbf{H}$, say h. In a set of constraints, a constraint which is implied by others is said to be redundant. Molenberghs and Lesaffre (1994) and Glonek and McCullagh (1995) provided (different) reparameterizations of the expected cell frequencies in terms of "marginal" model parameters, which they showed to be functionally independent, by demonstrating that the derivative matrix of these parameters with respect to the expected cell frequencies is of full column rank. Say these parameters are $\boldsymbol{\theta}$, then a model $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}$ specified for these parameters does not contain redundant parameters $\boldsymbol{\beta}$ when $\mathbf{X}$ is of full column rank, or, equivalently, with $\mathbf{U}$ the orthogonal complement of $\mathbf{X}$, the constraints $\mathbf{U}'\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$ are functionally independent when $\mathbf{U}$ is of full column rank. Lang and Agresti (1994) also gave a characterization of some models having functionally independent constraints.

A complication in determining the number of degrees of freedom, *df*, is that the number of independent constraints may depend on $\mathbf{m}$. By way of illustration, consider the model defined by the simultaneous constraints

$$\begin{aligned} (m_1 + m_4)(m_1 + m_2) &= 2m_1(m_2 + m_4) \\ (m_1 + m_4)(m_1 + m_3) &= 2m_1(m_3 + m_4). \end{aligned}$$

(No application for this model is known.) Straightforward algebra shows that the constraints are equivalent to

$$\begin{aligned}(m_1 - m_4)(m_1 - m_2) &= 0\\ (m_1 - m_4)(m_1 - m_3) &= 0,\end{aligned}$$

i.e.,

$$m_1 = m_4 \qquad \vee \qquad m_1 = m_2 = m_3.$$

The model is defined by a *disjunction* of two sets of constraints rather than a more common *conjunction* of constraints. As a consequence, the number of degrees of freedom becomes a random variable, leading to statistical complications.

5.4.2 Partitioning chi-squared statistics

Aitchison (1962) introduced the concept of *asymptotic separability* for hypotheses for which the asymptotic chi-squared goodness-of-fit statistics can be asymptotically partitioned. If the parameter vectors $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are orthogonal (which implies that the respective MLEs are asymptotically independent), it is shown in Appendix A.3 that model $[\omega_1 \cap \omega_2]$ is asymptotically separable when $[\omega_1]$ constrains $\boldsymbol{\phi}$ but not $\boldsymbol{\psi}$, and $[\omega_2]$ constrains $\boldsymbol{\psi}$ but not $\boldsymbol{\phi}$. In that case, the Wald statistic partitions exactly as

$$W^2(\omega_1 \cap \omega_2) = W^2(\omega_1) + W^2(\omega_2). \tag{5.23}$$

It follows that goodness-of-fit statistics which are asymptotically equivalent to W^2, such as G^2 and X^2, can be asymptotically partitioned. (Note that (5.23) concerns nonnested models, whereas the exact partitioning of G^2, given in (2.24), concerns nested models.)

It is of particular interest for the sampling distributions described in Section 2.1 that certain products of frequencies are orthogonal to certain sums of frequencies. For instance, in a two-way table, the local odds ratios are orthogonal to the marginal frequencies. This fact can be used in the following example. Suppose we wish to test whether the linear by linear association model holds simultaneously with marginal homogeneity. MH can be tested using, for instance, Stuart's test, the linear by linear

model using G^2, and the simultaneous model by adding the values together. However, in practice there is often quite a big difference between G^2 for the simultaneous model and G^2 for the separate models added together. It is therefore recommended that models be fit simultaneously using the ML methods described in Section 5.1 and that goodness-of-fit be tested subsequently, rather than adding separate goodness-of-fit statistics together.

In general, a sum ϕ and product ψ defined as

$$\phi = \sum_i a_i m_i \qquad \psi = \prod_i m_i^{c_i}$$

are orthogonal, according to definition (A.17), if $\sum_i a_i c_i = 0$. It follows that the respective MLEs of ϕ and ψ are independent. Analogously, a vector of sums $\boldsymbol{\phi} = \mathbf{A}'\mathbf{m}$ and a vector of products $\boldsymbol{\psi} = \exp(\mathbf{C}' \log \mathbf{m})$ are orthogonal when $\mathbf{C}'\mathbf{A} = \mathbf{0}$. Thus, when $\mathbf{C}'\mathbf{A} = \mathbf{0}$, a model $[\omega_1]$ constraining $\boldsymbol{\phi}$ and a model $[\omega_2]$ constraining $\boldsymbol{\psi}$ are asymptotically separable. This generalizes a result proven by Lang (1996c) to a broader class of models.

The above result can be applied when simultaneously modelling loglinear and marginal models. If the marginal frequencies $\boldsymbol{\mu}$ constrained by the marginal model are a linear combination of the sufficient statistics of the loglinear model, then the marginal and loglinear model are asymptotically separable. For instance, for a three-way table ABC, a marginal model for marginal tables AB and BC and the loglinear model of no-three-factor interaction are asymptotically separable, while the same marginal model and the loglinear model of independence are not separable. In a two-way table, quasi-symmetry and marginal homogeneity are asymptotically separable, justifying the conditional test of MH given in Section 3.4.2.

5.4.3 Adjusted residuals

Following Lang and Agresti (1994), the definition of the adjusted residuals described in Section 2.5.4 can be generalized to adjusted residuals for marginal frequencies μ_i (which are sums of joint frequencies). Generalizing once more, adjusted residuals for measures $\zeta_i = \zeta_i(\boldsymbol{\mu})$ are also defined below. With y_i the observed value of μ_i, and z_i the observed value of ζ_i,

adjusted residuals are defined as

$$r_i(\mu_i) = \frac{y_i - \hat{\mu}_i}{\sigma(y_i - \hat{\mu}_i)} \qquad r_i(\zeta_i) = \frac{z_i - \hat{\zeta}_i}{\sigma(z_i - \hat{\zeta}_i)}.$$

Consider two nested marginal models $[\omega_1]$ and $[\omega_2]$ such that $[\omega_2]$ implies $[\omega_1]$, with fitted marginals $\hat{\mu}_{1i}$ and $\hat{\mu}_{2i}$, respectively, and with fitted measures $\hat{\zeta}_{1i}$ and $\hat{\zeta}_{2i}$, respectively. Conditional adjusted residuals (which were not given by Lang and Agresti) can be defined as

$$r_i(\mu_i; \omega_2|\omega_1) = \frac{\hat{\mu}_{1i} - \hat{\mu}_{2i}}{\sigma(\hat{\mu}_{1i} - \hat{\mu}_{2i})} \qquad r_i(\zeta_i; \omega_2|\omega_1) = \frac{\hat{\zeta}_{1i} - \hat{\zeta}_{2i}}{\sigma(\hat{\zeta}_{1i} - \hat{\zeta}_{2i})}. \tag{5.24}$$

Formulas for the standard errors are derived in the next section.

5.5 Asymptotic behaviour of MLEs

Using the methods employed by Aitchison and Silvey (1958) described in Appendix A.1 and the delta method, the asymptotic distributions of MLEs of some relevant parameters are derived given one of the sampling schemes described in Section 2.1 and the model constraint $\mathbf{h}(\mathbf{m}) = \mathbf{0}$. For a class of loglinear models for sums of frequencies, a formal derivation of the asymptotic distribution of various estimators was presented by Lang (1996a).

With $\boldsymbol{\theta} = \log \mathbf{m}$, and the log likelihood $\mathcal{L}$ defined by (5.1), let

$$\mathbf{B} = E\left(-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = \mathbf{D}_{\mathbf{m}}$$

be the information matrix. Suppose $\hat{\mathbf{m}}_s$ is the MLE of the true value of $\mathbf{m}$ given one of the sampling distributions described in Section 2.1, i.e., either Poisson sampling or Poisson sampling with sampling constraint $\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}$. In Section 2.6.2, it was shown that the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_s$ is

$$\boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_s) = \mathbf{D}_{\mathbf{m}}^{-1} - \boldsymbol{\Lambda}(\mathcal{S}),$$

with $\boldsymbol{\Lambda}(\mathcal{S})$ defined by (2.28). Let $\mathbf{h}(\mathbf{m}) = \mathbf{U}'\mathbf{g}(\boldsymbol{\zeta}(\boldsymbol{\mu}))$ be the marginal model constraint function (cf. 5.4) and let $\mathbf{H} = \partial \mathbf{h}'/\partial \boldsymbol{\theta}$. Suppose $\hat{\boldsymbol{\theta}}_h =$

$\log \hat{\mathbf{m}}_h$ is the MLE of the true value of $\boldsymbol{\theta}$ given Poisson sampling given that $\mathbf{h}(\mathbf{m}) = \mathbf{0}$. Using (A.5), the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_h$ is

$$\begin{aligned} \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_h) &= \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{B}^{-1} \\ &= \mathbf{D}_{\mathbf{m}}^{-1} - \mathbf{D}_{\mathbf{m}}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{H})^{-1}\mathbf{H}\mathbf{D}_{\mathbf{m}}^{-1}. \end{aligned}$$

Consider full multinomial sampling, i.e., the sampling constraint is $\mathbf{1}'\mathbf{m} = \mathbf{1}'\mathbf{n}$. Assuming that $\mathbf{h}(\mathbf{m})$ is a homogeneous function of the expected frequencies (see Section 4.3.2 and Appendix D), i.e., that $\mathbf{h}(\mathbf{m}) = \mathbf{h}(\boldsymbol{\pi})$, it follows that

$$\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}'} \cdot \mathbf{B}^{-1} \cdot \frac{\partial \mathbf{m}'\mathbf{1}}{\partial \boldsymbol{\theta}} = \frac{\partial \mathbf{h}'}{\partial \mathbf{m}} \mathbf{D}_{\mathbf{m}} \cdot \mathbf{D}_{\mathbf{m}}^{-1} \cdot \mathbf{m} = \frac{\partial \mathbf{h}'}{\partial \mathbf{m}} \mathbf{m} = \mathbf{0}.$$

Thus, by definition (A.17) of orthogonality of parameters, $\mathbf{h}(\mathbf{m})$ is orthogonal to $\mathbf{1}'\mathbf{m}$. More generally, with sampling constraints $\mathbf{W}'\mathbf{m} = \mathbf{W}'\mathbf{n}$, it can be shown that $\mathbf{W}'\mathbf{m}$ is orthogonal to $\mathbf{h}(\mathbf{m})$ when $\mathbf{h}(\mathbf{m})$ is a homogeneous function with respect to the parameters of each sample. Therefore, the partitioning (A.18) can be used to calculate the asymptotic covariance matrix of $\log \hat{\mathbf{m}}_{s+h}$, the MLE of $\mathbf{m}$ given both sampling scheme $\mathcal{S}$ and $\mathbf{h}(\mathbf{m}) = \mathbf{0}$. With $\boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_p) = \mathbf{D}_{\mathbf{m}}^{-1}$ the covariance matrix of $\log \hat{\mathbf{m}}$ given Poisson sampling without further restrictions, the partitioning (A.18) yields

$$\begin{aligned} \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_{s+h}) &= \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_s) + \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_h) - \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}_p) \\ &= \mathbf{D}_{\mathbf{m}}^{-1} - \mathbf{D}_{\mathbf{m}}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{D}_{\mathbf{m}}^{-1} - \boldsymbol{\Lambda}(\mathcal{S}), \end{aligned}$$

where $\boldsymbol{\Lambda}(\mathcal{S})$ is defined by (2.28). This generalizes a result by Lang (1996b) to broader classes of sampling schemes and models.

Using the delta method described in Section (2.6.1), it is possible to derive the asymptotic covariance matrices of some relevant parameters which are functions of $\log \mathbf{m}$. Omitting the subscript $s + h$, we obtain

$$\begin{aligned} \boldsymbol{\Sigma}(\hat{\mathbf{m}}) &= \mathbf{D}_{\mathbf{m}} \boldsymbol{\Sigma}(\log \hat{\mathbf{m}}) \mathbf{D}_{\mathbf{m}} \\ \boldsymbol{\Sigma}(\hat{\boldsymbol{\mu}}) &= \mathbf{M}' \boldsymbol{\Sigma}(\hat{\mathbf{m}}) \mathbf{M} \\ \boldsymbol{\Sigma}(\hat{\boldsymbol{\zeta}}) &= \mathbf{Z}' \boldsymbol{\Sigma}(\hat{\boldsymbol{\mu}}) \mathbf{Z} \\ \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}'\boldsymbol{\Sigma}(\hat{\boldsymbol{\zeta}})\mathbf{G}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

where $\mathbf{G}$ and $\mathbf{Z}$ are defined by (5.5). For the adjusted residuals (5.24), the standard errors of the differences $\hat{m}_{1i} - \hat{m}_{2i}$ are needed. Since both estimators depend on the observed frequencies $\mathbf{n}$, the delta method can be used to obtain the asymptotic covariance matrix, yielding

$$\begin{aligned}
\boldsymbol{\Sigma}(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2) &= \mathbf{H}_2(\mathbf{H}_2'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{H}_2)^{-1}\mathbf{H}_2' - \mathbf{H}_1(\mathbf{H}_1'\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{H}_1)^{-1}\mathbf{H}_1' \quad (5.25)\\
\boldsymbol{\Sigma}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) &= \mathbf{M}'\boldsymbol{\Sigma}(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2)\mathbf{M} \quad (5.26)\\
\boldsymbol{\Sigma}(\hat{\boldsymbol{\zeta}}_1 - \hat{\boldsymbol{\zeta}}_2) &= \mathbf{Z}'\boldsymbol{\Sigma}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)\mathbf{Z}. \quad (5.27)
\end{aligned}$$

The standard error of, for instance, $\sigma(\hat{m}_{1i} - \hat{m}_{2i})$ is the square root of the (i,i)th diagonal element of (5.25). Note that the values of the adjusted residuals (5.24) do not depend on which one of the sampling distributions $\mathcal{S}$ described in Section (2.1) is used.

Finally, it is noted that the result on the average precision of estimated log expected cell frequencies derived in Section 2.6.3 can be generalized to marginal models. The average precision of $\log \hat{\mathbf{m}}$ is defined as

$$\bar{\sigma}^2(\log \hat{\mathbf{m}}) = \sum_i \pi_i \sigma^2(\log \hat{m}_i).$$

As was done in Section 2.6.3, this formula can be reduced to

$$\bar{\sigma}^2(\log \hat{\mathbf{m}}) = \frac{f}{n},$$

with f the number of free parameters of the loglinear model, and n the sample size. That is, f is equal to the number of cells minus the number of degrees of freedom for the model, minus the number of sampling constraints.

5.6 Conclusion

A simple and computationally efficient maximum likelihood estimation algorithm for fitting marginal models was presented in this chapter. This is a modification of Aitchison and Silvey's (1958) method and a generalization and modification of Haber (1985) and Lang's (1996a) techniques. A method for monitoring convergence was also presented, which was not done by the above authors. Our method can be used for large contingency tables with millions of cells. Major advantages of the algorithm

are that it is simple to program (see Appendix E) and can be used for a very large class of models. A drawback is that it requires evaluation of MLEs of all expected frequencies (which is the case with most algorithms for maximum likelihood estimation). Since the number of cell frequencies increases exponentially with the number of variables, only a limited number of variables can be handled.

A class of loglinear models for marginal frequencies was described for which the likelihood function is uniquely maximized subject to the model constraints. This greatly eases the maximum likelihood estimation, since, if a local maximum has been found, it is for certain the global maximum.

A brief description was given of the weighted least squares and the generalized estimating equation approaches. Though maximum likelihood is preferred by many statisticians to these methods, they can be of use when a contingency tables is too large to be fitted using maximum likelihood.

Standard chi-squared statistics can be used for testing goodness-of-fit of marginal models. Lang's (1996c) results on asymptotic partitioning of goodness-of-fit statistics were generalized. Additionally, a generalization of adjusted residuals (Haberman, 1973; Lang, 1996a) was presented. In particular, it was shown how conditional adjusted residuals and adjusted residuals for various measures can be calculated. Finally, the asymptotic behaviour of MLEs of various parameters given various sampling distributions was derived.

Chapter 6

Future research

A general class of models for analyzing categorical data has been presented in this book, together with maximum likelihood fitting and testing methods. Marginal models are especially useful for testing whether specific aspects of various marginal distributions are similar in certain respects. A summary of some of the most important limitations of the models and methods discussed in this book are described below, delineating areas where future work has to be done.

First of all, several types of questions concerning homogeneity of marginal distributions still cannot be answered using marginal models. The models described here are "linear" in the model parameters in the sense that a linear predictor is used in the marginal model equation (4.3). An example of a model that does not fit into this equation is Goodman's (1979) *row and column effects* model because it is *logmultiplicative* in the model parameters. A marginal model involving the row and column effects model arises when one wishes to test whether the association parameters in several bivariate marginal tables are equal. To accommodate for logmultiplicative and various other "nonlinear" models, the marginal model equation (4.3) might be generalized by replacing the linear predictor $\mathbf{X}\boldsymbol{\beta}$ with an arbitrary function $\mathbf{u}(\boldsymbol{\beta})$. The greater generality brings with it possible additional estimation and testing problems, however.

Another type that does not fit easily into the marginal model equation of this book is the latent class model. In a panel study, it may be hypothesized that certain manifest variables are the realization of an underlying latent variable. It may be interesting to test whether there are

certain changes in the latent variable, instead of in the manifest variables (Hagenaars, 1992). It appears that the EM algorithm (Dempster, Laird, & Rubin, 1977) can be used in combination with those described in this book.

Another area where more research is needed is the extension of methods presented in this book so that inequality constraints on parameters can be handled. A simple question leading to inequality constraints on parameters is: Does association between two variables increase over time? If there are two points in time, the question may be answered by testing whether all the odds ratios at time point 2 are greater than the corresponding odds ratios at time point 1. Instead of odds ratios, gamma may also be used to measure the association. Standard chi-squared tests cannot generally be used for such hypotheses (Robertson, Wright, & Dykstra, 1988).

A difficult problem which haunts categorical data analysis in general and therefore also the models of this book is the testing of models when only sparse data are available. When there are many observed zeroes in the marginal distributions of interest, large sample methods, such as asymptotic chi-squared statistics, cannot be used to test marginal models. A promising approach to this problem is the use of posterior predictive checks or various other Bayesian methods (Gelman, Carlin, Stern, & Rubin, 1995; Carlin & Louis, 1996). Problems with such methods are the heavy computations which are often required. Alternatively, bootstrapping methods may be used.

Finally, a fundamental, purely technical problem with current maximum likelihood methods for categorical data is that the computational complexity increases explosively with the number of variables. Only for limited classes of models, such as various Markov chain models and modified path models, can the estimation process be simplified so that extremely large tables can be handled (Vermunt, 1997). For most models, however, current maximum likelihood methods require evaluation of every cell frequency in the full cross-classification of variables. This means that only a limited number of variables can be handled. This is not a limitation of maximum likelihood itself, however, but rather of the methods used to calculate estimates. It is very important that other techniques for maximum likelihood estimation, which do not require the evaluation of every cell frequency, be developed.

Appendix A

Maximum likelihood theory

A.1 Aitchison and Silvey's method

Consider a $y \times 1$ vector $\mathbf{y}$ of observations with log likelihood function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$, where $\boldsymbol{\theta}$ is a $b \times 1$ vector of unknown parameters. The essence of many statistical problems is to test whether or not $\boldsymbol{\theta}$ satisfies certain constraints. In this section, some of Aitchison and Silvey's (1958, 1960) results on maximum likelihood (ML) estimation and deriving the asymptotic distribution of the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ are presented. Not all of the regularity conditions will be addressed explicitly.

It is assumed that the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$, which maximize $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ with respect to $\boldsymbol{\theta}$, are a solution to the *score* equation

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{A.1}$$

In many cases, some iterative method is needed to solve equation (A.1). The Fisher information matrix is defined as

$$\mathbf{B} = E\left(-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right).$$

Assuming that $\mathbf{B}$ is nonsingular and that the population value of $\boldsymbol{\theta}$ lies in the interior of the parameter space, $\hat{\boldsymbol{\theta}}$ has an asymptotic normal distribution, with expected value the population value, and with covariance matrix

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \mathbf{B}^{-1}.$$

For many statistical problems, one would like to test whether or not the parameter $\boldsymbol{\theta}$ satisfies the constraints

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}, \tag{A.2}$$

where $\mathbf{h}(\boldsymbol{\theta})$ is an $h \times 1$ vector with elements $h_i(\boldsymbol{\theta})$. A traditional approach to find the MLE $\hat{\boldsymbol{\theta}}$ is to try to eliminate the constraint equations $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$, by finding a function $\mathbf{g}$ such that

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0} \;\; \Leftrightarrow \;\; \boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\psi})$$

for a $(b-h) \times 1$ vector of parameters $\boldsymbol{\psi}$. Then, $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) = \mathcal{L}(\mathbf{g}(\boldsymbol{\psi}); \mathbf{y})$ is a function of $\boldsymbol{\psi}$, and standard methods can be used to find a maximum of $\mathcal{L}$ in terms of $\boldsymbol{\psi}$. However, in many cases it is awkward or impossible to find such a function $\mathbf{g}$. For this reason, Aitchison and Silvey developed an alternative method for finding MLEs using Lagrange's method of undetermined multipliers. Under appropriate conditions, with $\boldsymbol{\lambda}$ an $h \times 1$ vector of unknown Lagrange multipliers, the MLE $\hat{\boldsymbol{\theta}}$ is found at a saddle point of the Lagrangian log likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) + \boldsymbol{\lambda}' \mathbf{h}(\boldsymbol{\theta}).$$

It is important to note that the problem of finding MLEs is now the problem of finding a *saddle point* of L, rather than a maximum. Let

$$\mathbf{k} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \qquad \mathbf{H} = \frac{\partial \mathbf{h}'}{\partial \boldsymbol{\theta}}.$$

Differentiating L with respect to $\boldsymbol{\theta}$ and equating the result to zero yields

$$\mathbf{l}(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}) = \frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbf{k} + \mathbf{H}\boldsymbol{\lambda} = \mathbf{0}. \tag{A.3}$$

Aitchison and Silvey gave regularity conditions, which, if satisfied, guarantee that the MLEs $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\lambda}}$ are a solution to equations (A.2) and (A.3) with probability going to one as the sample size approaches infinity. If the regularity conditions are satisfied and if the MLEs are a solution to equations (A.2) and (A.3), they have an asymptotic multivariate normal distribution, with mean equal to the population value of $\boldsymbol{\theta}$ and $\mathbf{0}$, respectively. Furthermore, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\lambda}}$ are asymptotically independent. Assuming that $\mathbf{H}$ is of full column rank h, the covariance matrices are

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}}) = (\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1} \tag{A.4}$$

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{B}^{-1}, \tag{A.5}$$

where $\hat{\boldsymbol{\theta}}$ is substituted for $\boldsymbol{\theta}$ in $\mathbf{B}$ and $\mathbf{H}$ on the right-hand sides of these formulas. It must be noted that $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is singular, with rank $b-h$. To test goodness-of-fit, Aitchison and Silvey proposed the Lagrange multiplier statistic, defined as

$$L^2 = \hat{\boldsymbol{\lambda}}' \boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}})^{-1} \hat{\boldsymbol{\lambda}}.$$

Provided that the necessary regularity conditions given by Aitchison and Silvey are satisfied, L^2 has an asymptotic chi-squared distribution with $df = h$.

A.2 Estimation of parameters

A solution $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}})$ is sought to the equations

$$\mathbf{l}(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y}) = \mathbf{0} \qquad \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$$

which are defined by (A.3) and (A.2). In the sequel, $\mathbf{l}$ and $\mathbf{h}$ are short for $\mathbf{l}(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{y})$ and $\mathbf{h}(\boldsymbol{\theta})$, respectively. Let

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\lambda} \end{pmatrix} \qquad \mathbf{f}(\boldsymbol{\xi}) = \begin{pmatrix} \mathbf{l} \\ \mathbf{h} \end{pmatrix} \qquad \mathbf{F}(\boldsymbol{\xi}) = E\left(-\frac{\partial \mathbf{f}'}{\partial \boldsymbol{\xi}}\right). \tag{A.6}$$

Then $\mathbf{F}(\boldsymbol{\xi})$ evaluates to

$$\mathbf{F}(\boldsymbol{\xi}) = \begin{pmatrix} \mathbf{B} & -\mathbf{H} \\ -\mathbf{H}' & \mathbf{0} \end{pmatrix}. \tag{A.7}$$

In order to solve $\mathbf{f}(\boldsymbol{\xi}) = \mathbf{0}$ for $\boldsymbol{\xi}$, Aitchison and Silvey proposed the iterative scheme

$$\boldsymbol{\xi}^{(k+1)} = \mathbf{u}(\boldsymbol{\xi}^{(k)}) \tag{A.8}$$

using the "updating" function

$$\mathbf{u}(\boldsymbol{\xi}) = \boldsymbol{\xi} + \mathbf{F}(\boldsymbol{\xi})^{-1} \mathbf{f}(\boldsymbol{\xi}). \tag{A.9}$$

Since $\mathbf{F}(\boldsymbol{\xi})$ is the expected value of the matrix of second derivatives of the Lagrangian likelihood function, the iterative scheme is a type of Fisher

scoring. Appropriate initial estimates $\boldsymbol{\xi}^{(0)} = \text{vec}(\boldsymbol{\theta}^{(0)}, \boldsymbol{\lambda}^{(0)})$ should be chosen. The initial parameter vector $\boldsymbol{\theta}^{(0)}$ should be in the parameter space and preferably close to the MLE $\hat{\boldsymbol{\theta}}$. For the initial Lagrange parameter vector a suitable choice is $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$. The inversion of $\mathbf{F}$ can be simplified by using result 4 for the inverse of a partitioned matrix. To further simplify calculations, Aitchison and Silvey proposed using the initial matrix $\mathbf{F}(\boldsymbol{\xi}^{(0)})$ for all iterations, which has the advantage that only one matrix inversion is necessary for the iterative process. A drawback is that, if the initial value $\boldsymbol{\theta}^{(0)}$ is far from the MLE $\hat{\boldsymbol{\theta}}$, many iterations may be needed before convergence is reached.

The proposed algorithm (A.8) does not always converge when starting estimates are not close enough to the MLEs, in which case it is necessary to introduce a step size into the updating equation (A.14). This issue was not addressed by the Aitchison and Silvey. The standard approach to choosing a step size in optimization problems is to use a value for which the objective function to be maximized increases. However, since a saddle point of the Lagrangian likelihood L is sought, this standard approach cannot be used. A method for introducing a step size, which involves rewriting the updating function (A.9) is proposed below.

Let $\mathbf{z} = \mathbf{F}^{-1}\mathbf{f}$, i.e., $\mathbf{z}$ is the solution to the equation $\mathbf{F}\mathbf{z} = \mathbf{f}$. Suppose $\mathbf{z}$ is partitioned into the $b \times 1$ and $h \times 1$ vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ such that $\mathbf{z}' = (z_{11}, \ldots, z_{1b}, z_{21}, \ldots, z_{2h})$, $\mathbf{z}_1' = (z_{11}, \ldots, z_{1b})$, and $\mathbf{z}_2' = (z_{21}, \ldots, z_{2h})$. Using (A.6) and (A.7), the component vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ of $\mathbf{z}$ are the solution to the simultaneous linear equations

$$\mathbf{B}\mathbf{z}_1 - \mathbf{H}\mathbf{z}_2 = \mathbf{l} \tag{A.10}$$

$$-\mathbf{H}'\mathbf{z}_1 = \mathbf{h}. \tag{A.11}$$

By substitution into (A.11), one can verify that the solution is

$$\mathbf{z}_1 = \mathbf{B}^{-1}(\mathbf{l} + \mathbf{H}\mathbf{z}_2) \tag{A.12}$$

$$\mathbf{z}_2 = -(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{l} + \mathbf{h}). \tag{A.13}$$

Now suppose that $\mathbf{u}' = (v_1, \ldots, v_b, w_1, \ldots, w_h)$, $\mathbf{v}' = (v_1, \ldots, v_b)$, and $\mathbf{w}' = (w_1, \ldots, w_h)$, i.e., $\mathbf{u} = \text{vec}(\mathbf{v}, \mathbf{w})$. Then (A.9) is equivalent to

$$\mathbf{v}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \mathbf{z}_1 \qquad \mathbf{w}(\boldsymbol{\lambda}) = \boldsymbol{\lambda} + \mathbf{z}_2.$$

Substituting (A.12), (A.13), and $\mathbf{l} = \mathbf{k} + \mathbf{H}\boldsymbol{\lambda}$ into $\mathbf{v}(\boldsymbol{\theta})$ and $\mathbf{w}(\boldsymbol{\lambda})$ yields

$$\mathbf{v}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \mathbf{B}^{-1}(\mathbf{l} + \mathbf{H}\mathbf{z}_2) = \boldsymbol{\theta} + \mathbf{B}^{-1}(\mathbf{k} + \mathbf{H}(\boldsymbol{\lambda} + \mathbf{z}_2))$$

$$\begin{aligned} &= \boldsymbol{\theta} + \mathbf{B}^{-1}(\mathbf{k} + \mathbf{H}\mathbf{w}(\boldsymbol{\lambda})) \\ \mathbf{w}(\boldsymbol{\lambda}) &= \boldsymbol{\lambda} - (\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{l} + \mathbf{h}) \\ &= -(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{k} + \mathbf{h}). \end{aligned}$$

This rewrite is important, because $\boldsymbol{\lambda}$ does not appear on the right-hand sides of the above equations. In particular, $\hat{\boldsymbol{\theta}}$ is a fixed point of $\mathbf{v}(\boldsymbol{\theta})$, i.e., it is a solution to the equation $\mathbf{v}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. A step size *step* for which $0 < step \leq 1$ can now be introduced into $\mathbf{v}(\boldsymbol{\theta})$ as follows:

$$\mathbf{v}(\boldsymbol{\theta}, step) = \boldsymbol{\theta} + step\, \mathbf{B}^{-1}(\mathbf{k} + \mathbf{H}\mathbf{w}(\boldsymbol{\lambda})). \tag{A.14}$$

Note that, because of the rewrite, a different function is obtained than when a step size is introduced into (A.9). The iterative scheme that can be used is

$$\boldsymbol{\theta}^{(k+1)} = \mathbf{v}(\boldsymbol{\theta}^{(k)}, step^{(k)}).$$

Two questions must now be answered: 1) how do we choose the step size and 2) how do we know when the algorithm is sufficiently close to convergence? To answer both questions, an appropriate measure $e(\boldsymbol{\theta})$ for the distance from convergence can be chosen. Then, if possible, the step size at iteration k is chosen such that $e(\mathbf{v}(\boldsymbol{\theta}^{(k)}, step^{(k)})) < e(\boldsymbol{\theta}^{(k)})$, and the iterations can be stopped at iteration k when $e(\boldsymbol{\theta}^{(k)}) < \epsilon$, a sufficiently small constant chosen a priori. As a measure for the distance $e(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ from satisfying $\mathbf{v}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, we propose the statistic:

$$e(\boldsymbol{\theta}) = (\mathbf{v}(\boldsymbol{\theta}) - \boldsymbol{\theta})'\mathbf{B}(\mathbf{v}(\boldsymbol{\theta}) - \boldsymbol{\theta}).$$

At iteration k, one can start with $step^{(k)} = 1$, and keep halving $step^{(k)}$ while $e(\mathbf{v}(\boldsymbol{\theta}^{(k)}, step^{(k)})) \geq e(\boldsymbol{\theta}^{(k)})$. However, it is not always possible to obtain a decrease in the error function $e(\boldsymbol{\theta})$, partly because the iterative scheme is not based on the gradient of the error function. In fact, $e(\boldsymbol{\theta})$ can only give a very rough indication over several iterations of whether one is on the right "track" to obtaining convergence. Choosing the right step size remains an unsolved problem. As a rough practical guideline, it is recommended that at most only a few, say 5 or 6, halvings be performed.

Finally, we note that the covariance matrix of $\boldsymbol{\theta}$ is a by-product of the iterative scheme. Let

$$\mathbf{S}(\boldsymbol{\theta}) = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{B}^{-1}.$$

Then $\mathbf{S}(\boldsymbol{\theta}^{(k)})$ converges to $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ as $k \to \infty$. Writing out $\mathbf{v}(\boldsymbol{\theta})$ yields

$$\begin{aligned}\mathbf{v}(\boldsymbol{\theta}) &= \boldsymbol{\theta} + [\mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{H}'\mathbf{B}^{-1}]\mathbf{k} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{h} \\ &= \boldsymbol{\theta} + \mathbf{S}(\boldsymbol{\theta})\mathbf{k} - \mathbf{B}^{-1}\mathbf{H}(\mathbf{H}'\mathbf{B}^{-1}\mathbf{H})^{-1}\mathbf{h}.\end{aligned}$$

A.3 Parameter orthogonality

Assume the log likelihood function $\mathcal{L}(\boldsymbol{\theta})$ is a function of an unknown parameter $\boldsymbol{\theta}$ and suppose $\boldsymbol{\theta}' = (\phi_1, \ldots, \phi_q, \psi_1, \ldots, \psi_r)$ $\boldsymbol{\phi}' = (\phi_1, \ldots, \phi_q)$, and $\boldsymbol{\psi}' = (\psi_1, \ldots, \psi_r)$, i.e., $\boldsymbol{\theta}$ is partitioned into $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. Cox and Reid (1987) defined $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ to be orthogonal when

$$E\left(-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\phi}' \partial \boldsymbol{\psi}}\right) = \mathbf{0}. \tag{A.15}$$

An immediate result is that MLEs $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\psi}}$ are asymptotically independent. A disadvantage of definition (A.15) is that it cannot be used to define orthogonality of parameter vectors $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ when the log likelihood is not a function of these parameters. This happens when $\boldsymbol{\phi} = \mathbf{g}_1(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \mathbf{g}_2(\boldsymbol{\theta})$ for non-invertible functions $\mathbf{g}_1$ and $\mathbf{g}_2$. For this reason, the following broader definition is proposed. Assuming that $\mathbf{g}_1$ and $\mathbf{g}_2$ are differentiable, let their Jacobians be

$$\mathbf{G}_1 = \frac{\partial \mathbf{g}_1'}{\partial \boldsymbol{\theta}} \qquad \mathbf{G}_2 = \frac{\partial \mathbf{g}_2'}{\partial \boldsymbol{\theta}}. \tag{A.16}$$

With $\mathbf{B}$ the information matrix of $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are defined to be orthogonal if

$$\mathbf{G}_1'\mathbf{B}^{-1}\mathbf{G}_2 = \mathbf{0}. \tag{A.17}$$

It can be verified that parameters which are orthogonal as defined by Cox and Reid, are also orthogonal according to this definition.

Orthogonality of parameters has several interesting consequences. In the remaining part of this section, the vectors $\boldsymbol{\phi} = \mathbf{g}_1(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \mathbf{g}_2(\boldsymbol{\theta})$ are assumed to be orthogonal, with Jacobians given by (A.16). An important result is that the MLEs $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\psi}}$ are asymptotically independent. This can be demonstrated as follows. Using the delta method, with $\mathbf{g} = \text{vec}(\boldsymbol{\phi}, \boldsymbol{\psi}) = \text{vec}(\mathbf{g}_1(\boldsymbol{\theta}), \mathbf{g}_2(\boldsymbol{\theta}))$, and Jacobian $\mathbf{G} = (\mathbf{G}_1 \quad \mathbf{G}_2)$, and

$\mathbf{B}^{-1}$ the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$, the asymptotic covariance matrix of $\hat{\mathbf{g}}$ is

$$\boldsymbol{\Sigma}(\hat{\mathbf{g}}) = \mathbf{G}'\mathbf{B}^{-1}\mathbf{G} = \begin{pmatrix} \mathbf{G}_1'\mathbf{B}^{-1}\mathbf{G}_1 & \mathbf{G}_1'\mathbf{B}^{-1}\mathbf{G}_2 \\ \mathbf{G}_2'\mathbf{B}^{-1}\mathbf{G}_1 & \mathbf{G}_2'\mathbf{B}^{-1}\mathbf{G}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{G}_1'\mathbf{B}^{-1}\mathbf{G}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2'\mathbf{B}^{-1}\mathbf{G}_2 \end{pmatrix}.$$

It is seen that the covariances between $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\psi}}$ are zero.

Next, consider the unrestricted model $[\omega_0]$, the model $[\omega_1]$ defined by the constraint $\mathbf{h}_1(\boldsymbol{\phi}) = \mathbf{0}$, and the model $[\omega_2]$ defined by the constraint $\mathbf{h}_2(\boldsymbol{\psi}) = \mathbf{0}$. Thus, $[\omega_1]$ only constrains $\boldsymbol{\phi}$, and $[\omega_2]$ only constrains $\boldsymbol{\psi}$. Let $\mathbf{H}_1 = \partial\mathbf{h}_1'/\partial\boldsymbol{\theta}$ and $\mathbf{H}_2 = \partial\mathbf{h}_2'/\partial\boldsymbol{\theta}$ be the derivative matrices of the constraints. Let $\hat{\boldsymbol{\theta}}_i$ be the MLE of $\boldsymbol{\theta}$ under model $[\omega_i]$, and let $\hat{\boldsymbol{\theta}}_{12}$ be the MLE of $\boldsymbol{\theta}$ under model $[\omega_1 \cap \omega_2]$. (Note that $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, and $\hat{\boldsymbol{\theta}}_{12}$ are of the same dimensionality.) The following results can be proven.

1. Given that the model $[\omega_1 \cap \omega_2]$ is true, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_{12}$ can be partitioned as

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_{12}) = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_1) + \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_2) - \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_0), \tag{A.18}$$

   where the covariance matrices are estimated using the estimates $\hat{\boldsymbol{\theta}}_{12}$. This result is useful for deriving the standard errors of MLEs, if formulas for the standard errors of MLEs under the separate models are available. One can also see the contribution of the separate models to the standard errors.

2. Given that the model $[\omega_1 \cap \omega_2]$ is true, the asymptotic distribution of $\hat{\boldsymbol{\phi}}_{12}$ is identical to the asymptotic distribution of $\hat{\boldsymbol{\phi}}_1$, i.e.,

$$\begin{aligned} E(\hat{\boldsymbol{\phi}}_{12}) &= E(\hat{\boldsymbol{\phi}}_1) \\ \boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}_{12}) &= \boldsymbol{\Sigma}(\hat{\boldsymbol{\phi}}_1). \end{aligned}$$

   It follows that both estimators have the same efficiency. Of course, the same result holds for the MLEs $\hat{\boldsymbol{\psi}}_2$ and $\hat{\boldsymbol{\psi}}_{12}$. This result is important if, for example, we are interested in $\boldsymbol{\phi}$, while $\boldsymbol{\psi}$ is regarded as a nuisance parameter. More details on this subject are given by Cox and Reid (1987).

3. The Wald statistic for testing the simultaneous model $[\omega_1 \cap \omega_2]$ partitions exactly as

$$W^2(\omega_1 \cap \omega_2) = W^2(\omega_1) + W^2(\omega_2), \tag{A.19}$$

where $W^2(\omega)$ is the Wald statistic for testing whether the model $[\omega]$ is true. It follows that other chi-squared statistics which are asymptotically identically distributed as W^2, e.g., G^2, can be asymptotically partitioned. Aitchison (1962) called such models $[\omega_1]$ and $[\omega_2]$ *asymptotically separable*.

The results can be verified by writing out formula (A.5).

Appendix B

Uniqueness of MLEs for regular models

B.1 A theorem about uniqueness of MLEs

Consider an $r \times 1$ vector $\mathbf{m}$ of parameters with elements m_i and parameter space

$$\Omega = \mathbf{R}_+^r = \{\mathbf{m} = (m_1, \ldots, m_r)' : m_i > 0 \, \forall i\}.$$

(where $\mathbf{R}_+$ is the set of strictly positive reals). With $\mathbf{n}$ an $r \times 1$ vector of constants with elements $n_i > 0$, consider the kernel of the multivariate Poisson log likelihood function

$$\mathcal{L}(\mathbf{m}) = \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m}.$$

A maximum of $\mathcal{L}$ is sought subject to the constraint

$$\mathbf{h}(\mathbf{m}) = \mathbf{C}' \log \mathbf{A}'\mathbf{m} = \mathbf{0}, \tag{B.1}$$

where $\mathbf{A}$ is a known $r \times a$ matrix with elements $a_{ij} \geq 0$, such that every column of $\mathbf{A}$ contains at least one nonzero element, and $\mathbf{C}$ is a known $a \times c$ matrix with elements c_{jk}. Let $\boldsymbol{\mu} = \mathbf{A}'\mathbf{m}$. It follows from the restrictions on the elements of $\mathbf{A}$ that $\mu_j > 0$ for all j and $\mathbf{m} \in \Omega$. Therefore the constraint function $\mathbf{h}(\mathbf{m})$ is differentiable for all $\mathbf{m} \in \Omega$.

In general, there may be multiple values of $\mathbf{m}$ maximizing $\mathcal{L}$ subject to (B.1). However, if the vector $\mathbf{C}' \log \mathbf{A}'\mathbf{m}$ is regular in the following sense, it is shown below that there is only one such $\mathbf{m}$. The

vector $\mathbf{C}' \log \mathbf{A}'\mathbf{m}$ is said to be regular if, for all λ_k and $\mu_j > 0$, with $t_j = \sum_k c_{jk}\lambda_k$,

$$\sum_j \frac{a_{ij}}{\mu_j} t_j < 1 \; \forall i \quad \Rightarrow \quad \sum_j \frac{a_{ij}}{\mu_j} \max(0, t_j) < 1 \; \forall i. \tag{B.2}$$

It seems hard to give an intuitive interpretation of (B.2). It can be verified that a subvector of a regular vector of measures is regular, but when two regular vectors of measures are concatenated, regularity may be (and often is) destroyed. Note that a vector of linear combinations of elements of a vector of regular measures is regular.

Assuming that the solution set

$$\mathcal{S} = \{\mathbf{m} \in \Omega : \mathbf{h}(\mathbf{m}) = \mathbf{0}\}$$

is nonempty, i.e., that the constraints are consistent, we have the following central result.

Theorem 1 *Assume that (B.2) holds and that $n_i > 0$ for all i. Then there is a unique $\mathbf{m} \in \Omega$ maximizing $\mathcal{L}(\mathbf{m})$ subject to (B.1).*

From Theorem 1 it follows that (B.2) ensures connectedness of the solution set $\mathcal{S}$.

B.2 Proof of uniqueness of MLEs

This section is devoted to the proof of Theorem 1. To begin with, it is shown that the constrained maximization problem, namely, maximizing $\mathcal{L}(\mathbf{m})$ subject to (B.1), can be reformulated as the problem of solving two sets of equations by making use of Lagrange multipliers.

With $\boldsymbol{\lambda}$ an $h \times 1$ vector of Lagrange multipliers, consider the Lagrangian log likelihood function

$$\begin{aligned} L(\mathbf{m}, \boldsymbol{\lambda}) &= \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m} + \boldsymbol{\lambda}'\mathbf{C}' \log \mathbf{A}'\mathbf{m} \\ &= \mathbf{n}' \log \mathbf{m} - \mathbf{1}'\mathbf{m} + \mathbf{t}' \log \boldsymbol{\mu}, \end{aligned} \tag{B.3}$$

where $\mathbf{t} = \mathbf{C}\boldsymbol{\lambda}$ and $\boldsymbol{\mu} = \mathbf{A}'\mathbf{m}$. Using scalar notation, we have

$$L(\mathbf{m}, \boldsymbol{\lambda}) = \sum_i n_i \log m_i - \sum_i m_i + \sum_j t_j \log \mu_j.$$

The first derivative of L with respect to $\log \mathbf{m}$ equated to zero is

$$\frac{\partial L(\mathbf{m}, \boldsymbol{\lambda})}{\partial \log \mathbf{m}} = \mathbf{n} - \mathbf{m} + \mathbf{H}\boldsymbol{\lambda} = \mathbf{0}, \tag{B.4}$$

where $\mathbf{H}$ is the derivative matrix of $\mathbf{h}(\mathbf{m})$ with respect to $\log \mathbf{m}$, given by

$$\mathbf{H}(\mathbf{m}) = \mathbf{D}_{\mathbf{m}}\mathbf{A}\mathbf{D}_{\boldsymbol{\mu}}^{-1}\mathbf{C}.$$

Thus, (B.4) reduces to

$$\mathbf{n} - \mathbf{m} + \mathbf{D}_{\mathbf{m}}\mathbf{A}\mathbf{D}_{\boldsymbol{\mu}}^{-1}\mathbf{t} = \mathbf{0}. \tag{B.5}$$

In scalar notation, we have

$$n_i - m_i + m_i \sum_j \frac{a_{ij}}{\mu_j} t_j = 0.$$

The following lemma translates the constrained maximization problem to the problem of finding a stationary point of $L(\mathbf{m}, \boldsymbol{\lambda})$.

Lemma 2 *Assume that* $\hat{\mathbf{m}}$ *maximizes* $\mathcal{L}(\mathbf{m})$ *subject to (B.1) and that* $\mathbf{H}(\hat{\mathbf{m}})$ *is of full column rank. Then there is a unique vector* $\hat{\boldsymbol{\lambda}}$ *such that the pair* $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ *satisfies (B.5).*

Since $\mathcal{L}(\mathbf{m})$ and $\mathbf{h}(\mathbf{m})$ are differentiable on the whole parameter space, the lemma follows directly from a classical result about Lagrange multipliers (see, e.g., Bertsekas, 1982, p. 67).

It follows from Lemma 2 that the original constrained optimization problem can be solved by finding solutions to equations (B.1) and (B.5). To prove Theorem 1, it is sufficient to show that there is a unique pair $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ satisfying equations (B.1) and (B.5), such that $\hat{\mathbf{m}}$ maximizes $\mathcal{L}(\mathbf{m})$ subject to (B.1). The proof consists of two parts: 1) it is proven that there is *at least* one pair $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ which satisfies the appropriate conditions, and 2) it is proven that there is *at most* one such pair. The first part is easiest and is given by the following lemma.

Lemma 3 *Given that all* $n_i > 0$*, a solution* $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ *to equations (B.1) and (B.5) such that* $\hat{\mathbf{m}}$ *maximizes* $\mathcal{L}(\mathbf{m})$ *subject to (B.1) exists.*

Proof. Since $\mathbf{h}(\mathbf{m})$ is continuous and differentiable for all $\mathbf{m}$ with positive elements, the solution set $\mathcal{S}$ is closed in Ω. Since $\mathcal{L}(\mathbf{m})$ is strictly concave, with terms $n_i \log m_i - m_i$ going to minus infinity if $m_i \downarrow 0$ or $m_i \to \infty$, and since $\mathcal{S}$ is assumed to be nonempty, there must be at least one $\hat{\mathbf{m}}$ maximizing $\mathcal{L}$ subject to the constraint $\mathbf{h}(\mathbf{m}) = \mathbf{0}$. Thus, by Lemma 2, there must be a corresponding $\hat{\boldsymbol{\lambda}}$ such that $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ satisfies (B.5).

□

The second part of the proof, namely, that there is at most one solution $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ to equations (B.1) and (B.5) is the most difficult. An outline of the proof is as follows. It is shown that equation (B.5), with $\boldsymbol{\lambda}$ fixed, has a solution $\mathbf{m}(\boldsymbol{\lambda})$ if and only if $\boldsymbol{\lambda}$ is in a certain convex domain $\mathcal{C}$, independent of $\mathbf{m}$. Additionally, this solution is proven to be unique. Then it is shown that equation (B.1) with $\mathbf{m}$ substituted by $\mathbf{m}(\boldsymbol{\lambda})$ (i.e., $\mathbf{h}(\mathbf{m}(\boldsymbol{\lambda})) = \mathbf{0}$) has at most one solution $\hat{\boldsymbol{\lambda}}$. Thus, there is at most one pair $(\mathbf{m}(\hat{\boldsymbol{\lambda}}), \hat{\boldsymbol{\lambda}})$ which satisfies equations (B.1) and (B.5). This result combined with Lemma 3 proves the existence and uniqueness of such a pair.

To investigate solutions to (B.5) for fixed $\boldsymbol{\lambda}$, the behaviour of $L(\mathbf{m}, \boldsymbol{\lambda})$ as a function of $\mathbf{m}$ is investigated. The matrix of second derivatives of $L(\mathbf{m}, \boldsymbol{\lambda})$ with respect to $\log \mathbf{m}$ is

$$\begin{aligned} \mathbf{Q}(\mathbf{m}, \boldsymbol{\lambda}) &= \frac{\partial^2 L(\mathbf{m}, \boldsymbol{\lambda})}{\partial \log \mathbf{m} \partial \log \mathbf{m}'} \\ &= -\mathbf{D_m D}[\mathbf{1} - \mathbf{A}\mathbf{D}_{\boldsymbol{\mu}}^{-1}\mathbf{t}] - \mathbf{D_m A}\mathbf{D}_{\boldsymbol{\mu}}^{-2}\mathbf{D_t A}'\mathbf{D_m}, \end{aligned}$$

where $\mathbf{D}[\cdot]$ is the diagonal matrix with the elements of the vector in square brackets on the main diagonal. We have the following result.

Lemma 4 *Assume that $n_i > 0$ for all i and that (B.2) holds. Then matrix $\mathbf{Q}(\mathbf{m}, \boldsymbol{\lambda})$ is negative definite for all $(\mathbf{m}, \boldsymbol{\lambda})$ satisfying equation (B.5).*

Proof. With $t_j = \sum_k c_{jk}\lambda_k$, (B.5) yields

$$n_i - m_i + m_i \sum_j \frac{a_{ij}}{\mu_j} t_j = 0. \qquad \forall i$$

Dividing by m_i, it follows that

$$\sum_j \frac{a_{ij}}{\mu_j} t_j = 1 - \frac{n_i}{m_i} < 1, \qquad \forall i$$

since $n_i > 0$ and $m_i > 0$. Because of the regularity assumption (B.2), it follows that

$$\sum_j \frac{a_{ij}}{\mu_j} \max(0, t_j) < 1 \qquad \forall i. \tag{B.6}$$

It remains to be shown that, given (B.6), matrix $\mathbf{Q}(\mathbf{m}, \boldsymbol{\lambda})$ is negative definite. To simplify the notation, let

$$\mathbf{K}(\mathbf{m}) = -\mathbf{D}_{\mathbf{m}}^{-1} \mathbf{Q}(\mathbf{m}, \boldsymbol{\lambda}) \mathbf{D}_{\mathbf{m}}^{-1}.$$

Then $\mathbf{Q}(\mathbf{m}, \boldsymbol{\lambda})$ is negative definite if and only if $\mathbf{K}(\mathbf{m})$ is positive definite. Using the definition, $\mathbf{K}$ is positive definite if and only if, for all $\mathbf{x} = (x_1, x_2, \ldots)' \neq \mathbf{0}$,

$$\mathbf{x}'\mathbf{K}\mathbf{x} > 0.$$

The latter inequality may be proven as follows. Let $\mathbf{y} = \mathbf{A}'\mathbf{x}$. Writing out $\mathbf{x}'\mathbf{K}\mathbf{x}$ then yields

$$\begin{aligned} \mathbf{x}'\mathbf{K}\mathbf{x} &= \sum_i \frac{x_i^2}{m_i} \left(1 - \sum_j a_{ij} \frac{t_j}{\mu_j} \right) + \sum_j y_j^2 \frac{t_j}{\mu_j^2} \\ &= \sum_i \frac{x_i^2}{m_i} - \sum_j \frac{t_j}{\mu_j} \left(\sum_i a_{ij} \frac{x_i^2}{m_i} - \frac{y_j^2}{\mu_j} \right). \end{aligned} \tag{B.7}$$

Using Cauchy's inequality (Hardy, Littlewood, and Pólya 1967) it can be shown that the term in brackets is nonnegative:

$$\sum_i a_{ij} \frac{x_i^2}{m_i} - \frac{y_j^2}{\mu_j} = \sum_i \frac{(a_{ij} x_i)^2}{a_{ij} m_i} - \frac{(\sum_i a_{ij} x_i)^2}{\sum_i a_{ij} m_i} \geq 0.$$

Because of this inequality, substituting t_j in (B.7) by $\max(0, t_j)$ yields a lower bound for (B.7). Hence,

$$\begin{aligned} \mathbf{x}'\mathbf{K}\mathbf{x} &\geq \sum_i \frac{x_i^2}{m_i} - \sum_j \frac{\max(0, t_j)}{\mu_j} \left(\sum_i a_{ij} \frac{x_i^2}{m_i} - \frac{y_j^2}{\mu_j} \right) \\ &\geq \sum_i \frac{x_i^2}{m_i} - \sum_j \frac{\max(0, t_j)}{\mu_j} \sum_i a_{ij} \frac{x_i^2}{m_i} \\ &= \sum_i \frac{x_i^2}{m_i} \left(1 - \sum_j \frac{a_{ij}}{\mu_j} \max(0, t_j) \right) \\ &> 0, \end{aligned}$$

with the last inequality because of (B.6). This proves that $\mathbf{K}$ is positive definite, and therefore that $\mathbf{Q}$ is negative definite.

□

Lemma 4 implies that $L(\mathbf{m}, \boldsymbol{\lambda})$ as a function of $\mathbf{m}$ with $\boldsymbol{\lambda}$ fixed is concave at all its stationary points. It follows that every stationary point of $L(\mathbf{m}, \boldsymbol{\lambda})$ as a function of $\mathbf{m}$ with $\boldsymbol{\lambda}$ fixed is a strict maximum, implying that there can be at most one stationary point, which, if it exists, is the global maximum. Thus, (B.5) has, at most, one solution for all $\boldsymbol{\lambda}$. A solution does not exist for all $\boldsymbol{\lambda}$, but, as stated in the next lemma, if and only if $\boldsymbol{\lambda}$ is an element of a certain convex set.

Lemma 5 *Assume that all $n_i > 0$ and that (B.2) holds. Then there is a convex set $\mathcal{C} \subset \mathbf{R}^h$ (independent of $\mathbf{m}$ or $\boldsymbol{\lambda}$) such that, for fixed $\boldsymbol{\lambda}$: 1) equation (B.5) has a unique solution $\mathbf{m}(\boldsymbol{\lambda})$ if $\boldsymbol{\lambda} \in \mathcal{C}$, and 2) if $\boldsymbol{\lambda} \notin \mathcal{C}$, equation (B.5) does not have a solution.*

Proof. Let

$$\mathcal{C} = \left\{ \boldsymbol{\lambda} \,\middle|\, \left(\sum_{i:a_{ij}>0} n_i + t_j > 0 \; \forall j \right) \wedge \left(n_i + \sum_{j \in \mathcal{J}_i} t_j > 0 \; \forall i \right) \right\},$$

where $t_j = \sum_k c_{jk}\lambda_k$ and

$$\mathcal{J}_i = \{j \;:\; a_{ij} > 0,\; a_{hj} = 0 \; (h \neq i)\}.$$

Note that $\mathcal{J}_i$ is the index set consisting of those indices j for which $\mu_j = a_{ij}m_i$. It can easily be verified that $\mathcal{C}$ is convex. It is shown below that the lemma holds for this particular choice of $\mathcal{C}$.

From (B.3) we have

$$L(\mathbf{m}, \boldsymbol{\lambda}) = \sum_i n_i \log m_i - \sum_i m_i + \sum_j t_j \log \mu_j,$$

where $\mu_j = \sum_i a_{ij}m_i$ and $t_j = \sum_k \lambda_k c_{jk}$. From Lemma 4, any stationary point of L as a function of $\mathbf{m}$ is a strict maximum. Therefore, there can be at most one stationary point, which, if it exists, is the global maximum. To prove the lemma, we look at specific components of $L(\mathbf{m}, \boldsymbol{\lambda})$. Let

$$\mathcal{J}_i^c = \{j \notin \mathcal{J}_i : \exists_h a_{hj} > 0\}$$

be the set of indices j for which μ_j is the sum of $a_{ij}m_i$ and at least one other term. Thus, $\mathcal{J}_i \cup \mathcal{J}_i^c$ consists of those indices j for which μ_j contains $a_{ij}m_i$. Then the sum of all terms of $L(\mathbf{m}, \boldsymbol{\lambda})$ containing a certain m_i is

$$b_i(m_i) = n_i \log m_i - m_i + \sum_{j \in \mathcal{J}_i \cup \mathcal{J}_i^c} t_j \log \mu_j.$$

Note that $b_i(m_i)$ certainly contains m_i, but not necessarily any other elements of $\mathbf{m}$. Let $\mathbf{m}_j$ be the vector of those m_i contributing to μ_j (i.e., the m_i for which the index i is such that $a_{ij} > 0$), ordered by increasing indices. The sum of all terms of $L(\mathbf{m}, \boldsymbol{\lambda})$ containing an element of $\mathbf{m}_j$ is

$$d_j(\mathbf{m}_j) = \sum_{i:a_{ij}>0} n_i \log m_i - \sum_{i:a_{ij}>0} m_i + t_j \log \mu_j.$$

Note that $d_j(\mathbf{m}_j)$ contains all m_i contributing to μ_j but not any other elements of $\mathbf{m}$. To prove the lemma, the behaviour of all the terms $b_i(m_i)$ and $d_j(\mathbf{m}_j)$ is investigated. It is sufficient to show that: (i) all $b_i(m_i)$ and $d_j(\mathbf{m}_j)$ attain a (usually different) maximum in the interior of their parameter spaces if $\boldsymbol{\lambda} \in \mathcal{C}$, and (ii) there is some $b_i(m_i)$ or $d_j(\mathbf{m}_j)$ which does not have exactly one stationary point which is a maximum in the interior of the parameter space if $\boldsymbol{\lambda} \notin \mathcal{C}$. If (i) holds, then $L(\mathbf{m}, \boldsymbol{\lambda})$ must also attain a maximum in the interior of the parameter space if $\boldsymbol{\lambda} \in \mathcal{C}$. If (ii) holds, then $L(\mathbf{m}, \boldsymbol{\lambda})$ has either zero stationary points, so there can be no $\mathbf{m}(\boldsymbol{\lambda})$ we are seeking to find, or $L(\mathbf{m}, \boldsymbol{\lambda})$ cannot have only stationary points which are maxima, thereby contradicting Lemma 4. It should be remembered that the interior of the parameter space is the space where all $m_i > 0$. Thus, a maximum cannot be attained where some $m_i = 0$ (i.e., a strictly decreasing function of m_i does not attain a maximum in the interior of the parameter space). Points (i) and (ii) are proven below.

(i) It is shown that if $\boldsymbol{\lambda} \in \mathcal{C}$, all $b_i(m_i)$ and $d_j(\mathbf{m}_j)$ attain a maximum in the interior of the parameter space, for any i and j.

First, $b_i(m_i)$ is investigated. It is easily verified that $b_i(m_i)$ goes to $-\infty$ as $m_i \to \infty$. It remains to be seen what happens if $m_i \downarrow 0$. We have

$$\begin{aligned} b_i(m_i) &= n_i \log m_i - m_i + \sum_{j \in \mathcal{J}_i} t_j \log a_{ij} m_i + \sum_{j \in \mathcal{J}_i^c} t_j \log \mu_j \\ &= n_i \log m_i - m_i + \sum_{j \in \mathcal{J}_i} t_j \log m_i + \sum_{j \in \mathcal{J}_i} t_j \log a_{ij} + \sum_{j \in \mathcal{J}_i^c} t_j \log \mu_j \end{aligned}$$

$$= \left(n_i + \sum_{j \in \mathcal{J}_i} t_j \right) \log m_i - m_i + \sum_{j \in \mathcal{J}_i} t_j \log a_{ij} + \sum_{j \in \mathcal{J}_i^c} t_j \log \mu_j.$$

The term in brackets is positive because $\boldsymbol{\lambda} \in \mathcal{C}$, so $b_i(m_i) \to -\infty$ as $m_i \downarrow 0$. Thus, $b_i(m_i)$ attains a maximum for all $\boldsymbol{\lambda} \in \mathcal{C}$.

Next, the behaviour of $d_j(\mathbf{m}_j)$ is investigated. We have

$$d_j(\mathbf{m}_j) = \left(\sum_{i:a_{ij}>0} n_i \log m_i + t_j \log \mu_j \right) - \sum_{i:a_{ij}>0} m_i.$$

It is easily verified that as the elements of $\mathbf{m}_j$ go to infinity (at possibly different rates), that $d_j(\mathbf{m}_j) \to -\infty$, because the term on the right goes to minus infinity faster than the term in brackets goes to plus infinity. Hence, to show that $d_j(\mathbf{m}_j)$ has a maximum it is sufficient to demonstrate the following two assertions. First, that the gradient of $d_j(\mathbf{m}_j)$ as a function of m_i (m_i an element of $\mathbf{m}_j$) is positive if $m_i \in \langle 0, \delta \rangle$, for some $\delta > 0$. Second, that as all elements of $\mathbf{m}_j$ approach zero at the same rate in a certain way, the gradient of $d_j(\mathbf{m}_j)$ is positive. Differentiating with respect to $\log m_i$, the gradient of $d_j(\mathbf{m}_j)$ is

$$g_{ij} = \frac{\partial d_j(\mathbf{m}_j)}{\partial \log m_i} = n_i + \frac{a_{ij} m_i}{\mu_j} t_j - m_i.$$

It must first be shown that, for any i, $g_{ij} > 0$ for small enough m_i. If t_j is nonnegative, then $g_{ij} \geq n_i - m_i > 0$ if m_i is small enough since $n_i > 0$ is fixed. If t_j is negative and $\mu_j = a_{ij} m_j$, then $g_{ij} = n_i + t_j - m_i = \epsilon - m_i$ for some $\epsilon > 0$ by the first inequality of $\mathcal{C}$. If t_j is negative and $\mathbf{m}_j$ has more than one element, then as $m_i \downarrow 0$ (m_i in $\mathbf{m}_j$), $g_{ij} \to n_i > 0$.

The second assertion to be shown is that as all elements of $\mathbf{m}_j$ approach zero at the same rate in a certain way, the gradient of $d_j(\mathbf{m}_j)$ is positive. For all elements m_i^* of $\mathbf{m}_j^*$ let $m_i^* = n_i z$. Then as $z \downarrow 0$, all elements of $\mathbf{m}_j^*$ approach zero at the same rate in a specified way. Now $d_j(\mathbf{m}_j^*)$ can be differentiated with respect to $\log z$. We obtain

$$\begin{aligned} \frac{\partial d_j(\mathbf{m}_j^*)}{\partial \log z} &= \frac{\partial}{\partial \log z} \left(\sum_{i:a_{ij}>0} n_i \log(n_i z) - \sum_{i:a_{ij}>0} n_i z + t_j \log \sum_{i:a_{ij}>0} a_{ij} n_i z \right) \\ &= \sum_{i:a_{ij}>0} n_i - z \sum_{i:a_{ij}>0} n_i + t_j \end{aligned} \tag{B.8}$$

$$= \epsilon - z \sum_{i:a_{ij}>0} n_i$$

for some $\epsilon > 0$ (using the first inequality of $\mathcal{C}$). Thus, if $z \in \langle 0, \delta \rangle$, for some $\delta > 0$, then the gradient of $d_j(\mathbf{m}_j^*)$ as a function of z is positive.

(ii) Next it is shown that, if $\boldsymbol{\lambda} \notin \mathcal{C}$, some $b_i(m_i)$ or $d_j(\mathbf{m}_j)$ does not have exactly one stationary point which is a maximum. Two cases are distinguished below: a) the first inequality used in the definition of $\mathcal{C}$ is violated, and b) the second inequality is violated:

a) First, suppose that for some j:

$$\sum_{h:a_{hj}>0} n_h + t_j \leq 0. \tag{B.9}$$

The first inequality in the definition of $\mathcal{C}$ is violated and therefore $\boldsymbol{\lambda} \notin \mathcal{C}$. For all elements m_i^* of $\mathbf{m}_j^*$ suppose $m_i^* = n_i z$. Then as $z \downarrow 0$, all elements of $\mathbf{m}_j^*$ approach zero at the same rate in a specified way. The derivative of $d_j(\mathbf{m}_j^*)$ with respect to $\log z$ (see (B.8)) becomes

$$\begin{aligned} \frac{\partial d_j(\mathbf{m}_j^*)}{\partial \log z} &= \sum_{i:a_{ij}>0} n_i - z \sum_{i:a_{ij}>0} n_i + t_j \\ &\leq -z \sum_{i:a_{ij}>0} n_i \end{aligned}$$

Thus, $d_j(\mathbf{m}_j^*)$ as a function of z is a strictly decreasing function, and therefore $d_j(\mathbf{m}_j)$ as a function of $\mathbf{m}_j$ cannot have exactly one stationary point which is a maximum.

b) Second, suppose that

$$n_i + \sum_{j \in \mathcal{J}_i} t_j \leq 0 \tag{B.10}$$

for some i. The second inequality in the definition of $\mathcal{C}$ is violated and therefore $\boldsymbol{\lambda} \notin \mathcal{C}$. It is shown that the gradient of $b_i(m_i)$ with respect to m_i is negative if $m_i \in \langle 0, \delta_i]$ for some sufficiently small $\delta_i > 0$. Note that, for $j \in \mathcal{J}_i^c$,

$$\frac{a_{ij} m_i}{\mu_j} t_j = \frac{a_{ij} m_i}{a_{ij} m_i + \sum_{h \neq i} a_{hj} m_h} t_j.$$

Thus we have, with $k_j = \sum_{h \neq i} a_{hj} m_h$, and using (B.10),

$$\begin{aligned} \frac{\partial b_i(m_i)}{\partial \log m_i} &= n_i + \sum_{j \in \mathcal{J}_i} t_j - m_i + \sum_{j \in \mathcal{J}_i^c} \frac{a_{ij} m_i}{\mu_j} t_j \\ &\leq -m_i + \sum_{j \in \mathcal{J}_i^c} \frac{a_{ij} m_i}{\mu_j} t_j \\ &= -m_i + \sum_{j \in \mathcal{J}_i^c} \frac{a_{ij} m_i}{a_{ij} m_i + k_j} t_j, \end{aligned}$$

which is negative if m_i is sufficiently small and k_j is sufficiently large. Since the gradient of $b_i(m_i)$ is negative when $m_i \downarrow 0$ (given certain values of m_h, $h \neq i$) and $b_i(m_i) \to -\infty$ when $m_i \to \infty$, $b_i(m_i)$ cannot have exactly one stationary point which is a maximum.

□

Since $\mathbf{m}(\boldsymbol{\lambda})$ is uniquely defined for all $\boldsymbol{\lambda} \in \mathcal{C}$, we can investigate the implicit function $\mathbf{h}(\mathbf{m}(\boldsymbol{\lambda}))$ for $\boldsymbol{\lambda} \in \mathcal{C}$.

Lemma 6 *Assume that all $n_i > 0$ and that $\mathbf{H}(\mathbf{m}(\boldsymbol{\lambda}))$ is of full column rank for all $\boldsymbol{\lambda} \in \mathcal{C}$. Then $\mathbf{h}(\mathbf{m}(\boldsymbol{\lambda})) = \mathbf{0}$ has at most one solution $\hat{\boldsymbol{\lambda}} \in \mathcal{C}$.*

Proof. From Lemma 4 it follows that $\mathbf{m}(\boldsymbol{\lambda})$ is differentiable with respect to $\boldsymbol{\lambda}$. Differentiating $\mathbf{h}(\mathbf{m}(\boldsymbol{\lambda}))$ with respect to $\boldsymbol{\lambda}$ yields

$$\mathbf{S} = \frac{\partial \mathbf{h}(\mathbf{m}(\boldsymbol{\lambda}))}{\partial \boldsymbol{\lambda}} = -\mathbf{H}' \mathbf{Q}(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})^{-1} \mathbf{H}.$$

Since $(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ is a solution to (B.5) for all $\boldsymbol{\lambda} \in \mathcal{C}$, $\mathbf{Q}(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ is negative definite for all $\boldsymbol{\lambda} \in \mathcal{C}$ by Lemma 4. Since, additionally, $\mathbf{H}$ is assumed to be of full column rank, $\mathbf{S}$ is positive definite for all $\boldsymbol{\lambda} \in \mathcal{C}$. Thus, because $\mathcal{C}$ is convex, there can be at most one $\hat{\boldsymbol{\lambda}}$ satisfying $\mathbf{h}(\mathbf{m}(\boldsymbol{\lambda})) = \mathbf{0}$.

□

Now Theorem 1 can be proven. From Lemma 6, it follows that there is at most one $\hat{\boldsymbol{\lambda}}$ satisfying $\mathbf{h}(\mathbf{m}(\hat{\boldsymbol{\lambda}})) = \mathbf{0}$. For any such $\hat{\boldsymbol{\lambda}}$, it follows from Lemma 5 that there is a unique $\hat{\mathbf{m}} = \mathbf{m}(\hat{\boldsymbol{\lambda}})$ such that the pair $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$ satisfies (B.5). Since Lemma 3 states that there is at least one such pair, Theorem 1 is proven.

B.3 A suggestion for an algorithm to find MLEs

From the proof given above it follows that the MLE $\hat{\boldsymbol{\lambda}}$ is the minimum of $L(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ as a function of $\boldsymbol{\lambda}$, with $\mathbf{m}(\boldsymbol{\lambda})$ defined as the solution to (B.5). The MLE $\hat{\mathbf{m}}$ of the expected frequencies equals $\mathbf{m}(\hat{\boldsymbol{\lambda}})$. Since $L(\mathbf{m}(\boldsymbol{\lambda}), \boldsymbol{\lambda})$ is strictly convex on the convex domain $\mathcal{C}$ defined on page 130, it is straightforward to use a minimization algorithm for finding the MLEs $(\hat{\mathbf{m}}, \hat{\boldsymbol{\lambda}})$, similar to the algorithm used in Section 3.2.2. The difference with the algorithm given in Section 3.2.2 is that here, $\mathbf{m}(\boldsymbol{\lambda})$ is an implicit rather than explicit function of $\boldsymbol{\lambda}$. Any gradient method, for example Newton-Raphson can be used to search for $\hat{\boldsymbol{\lambda}}$. One can start with $\boldsymbol{\lambda} = \mathbf{0}$ (or any other $\boldsymbol{\lambda} \in C$). The procedure is not as efficient as the procedure described in Section 5.1, however, because the implicit function $\mathbf{m}(\boldsymbol{\lambda})$ must be calculated at every iteration. The advantage of the algorithm described here is that convergence can easily be guaranteed.

Appendix C

Results from matrix algebra

Result 1 *For any non-square full column rank matrix* $\mathbf{Q}$ *with orthogonal complement* $\mathbf{R}$*, the following identity holds:*

$$\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}' = \mathbf{I} - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'.$$

Proof. Let $\mathbf{P}_1 = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$ and $\mathbf{P}_2 = \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$. Then $\mathbf{P}_1^2 = \mathbf{P}_1$ and $\mathbf{P}_2^2 = \mathbf{P}_2$. Additionally, $\mathbf{P}_1\mathbf{P}_2 = \mathbf{0}$ and $\mathbf{P}_2\mathbf{P}_1 = \mathbf{0}$ because $\mathbf{Q}'\mathbf{R} = \mathbf{0}$ and $\mathbf{R}'\mathbf{Q} = \mathbf{0}$ (where the "$\mathbf{0}$" matrices have different dimensions). Now $\mathbf{P}_1 - \mathbf{P}_2 = \mathbf{P}_1^2 - \mathbf{P}_2^2 = (\mathbf{P}_1 - \mathbf{P}_2)(\mathbf{P}_1 + \mathbf{P}_2)$ so that

$$(\mathbf{P}_1 + \mathbf{P}_2 - \mathbf{I})(\mathbf{P}_1 - \mathbf{P}_2) = \mathbf{0}. \tag{C.1}$$

Matrix $(\mathbf{P}_1 - \mathbf{P}_2)$ is invertible since there is no non-zero vector $\mathbf{v}$ such that $(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{v} = \mathbf{0}$, because $\mathbf{P}_1\mathbf{v} = \mathbf{P}_2\mathbf{v}$ is impossible ($\mathbf{P}_1\mathbf{v}$ orthogonal to $\mathbf{P}_2\mathbf{v}$). Thus, from (C.1), $\mathbf{P}_1 = \mathbf{I} - \mathbf{P}_2$, which is the desired result. (Proof courtesy to Michel Petitjean and Denis Constales, communicated through Internet.)

□

Result 2 *For any non-square full column rank matrix* $\mathbf{Q}$ *with orthogonal complement* $\mathbf{R}$*, and a nonsingular matrix* $\mathbf{L}$*,*

$$\mathbf{L}\mathbf{Q}(\mathbf{Q}'\mathbf{L}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{L} = \mathbf{L} - \mathbf{R}(\mathbf{R}'\mathbf{L}^{-1}\mathbf{R})^{-1}\mathbf{R}'.$$

Proof. Let $\mathbf{S} = \mathbf{L}^{\frac{1}{2}}\mathbf{Q}$ and $\mathbf{T} = \mathbf{L}^{-\frac{1}{2}}\mathbf{R}$. Then, using result 1,

$$\begin{aligned}\mathbf{LQ}(\mathbf{Q}'\mathbf{LQ})^{-1}\mathbf{Q}'\mathbf{L} &= \mathbf{L}^{\frac{1}{2}}\left(\mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\right)\mathbf{L}^{\frac{1}{2}} \\ &= \mathbf{L}^{\frac{1}{2}}\left(\mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\right)\mathbf{L}^{\frac{1}{2}} \\ &= \mathbf{L} - \mathbf{R}(\mathbf{R}'\mathbf{L}^{-1}\mathbf{R})^{-1}\mathbf{R}'.\end{aligned}$$

□

Result 3 *For any non-square full column rank matrix* $\mathbf{Q}$ *with orthogonal complement* $\mathbf{R}$*, and a nonsingular matrix* $\mathbf{L}$*, the matrix*

$$\mathbf{L} - \mathbf{R}(\mathbf{R}'\mathbf{L}^{-1}\mathbf{R})^{-1}\mathbf{R}'$$

is nonnegative definite.

Proof. The result follows immediately from result 2.

□

Result 4 *For a nonsingular matrix* $\mathbf{A}$*, and matrices* $\mathbf{B}$*,* $\mathbf{C}$*, and* $\mathbf{D}$ *such that* $\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}$ *is nonsingular, the following formula holds (Searle, 1982, p. 260)*

$$\begin{pmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix} + \begin{pmatrix}-\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{I}\end{pmatrix}(\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1}\begin{pmatrix}-\mathbf{CA}^{-1} & \mathbf{I}\end{pmatrix}.$$

Appendix D

Homogeneous functions

Definition. A function $f(x_1, \ldots, x_n)$ is homogeneous of degree r (r a whole, possibly negative number) if for every $c > 0$:

$$f(cx_1, \ldots, cx_n) = c^r f(x_1, \ldots, x_n).$$

As an example, $f(x_1, x_2) = x_1/x_2$ is homogeneous of degree zero. A function $f(x_1, x_2, x_3)$ may be said to be homogeneous of degree r with respect to x_1 and x_2 when

$$f(cx_1, cx_2, x_3) = c^r f(x_1, x_2, x_3).$$

A key property of homogeneous functions is given by the following result.

Result 5 *(Euler's formula) Suppose that $f(x_1, \ldots, x_n)$ is homogeneous of degree r and differentiable. Then*

$$\sum_i \frac{\partial f(x_1, \ldots, x_n)}{\partial x_i} x_i = r\, f(x_1, \ldots, x_n).$$

For a proof, see Mas-Colell, Whinston, and Green (1995). For a function that is homogeneous of degree zero, Euler's formula says that

$$\sum_i \frac{\partial f(x_1, \ldots, x_n)}{\partial x_i} x_i = 0.$$

This is, in fact, the result used in this book.

Appendix E

A Mathematica program

A listing of *Mathematica* source code for maximum likelihood fitting of all the models described in this book is presented. The procedure listed below performs the iterative scheme described in Section 5.1. The input which is required consists of a (*Mathematica*) list of observed frequencies *n*, a starting estimate *start*, a link function *g*, a list of measures *zeta*, the Jacobian of zeta *Zt*, a matrix *Mt* such that *Mt.n* produces the observed marginals, and a design matrix *X*. A thorough description of the *Mathematica* language is presented by Wolfram (1996).

The program is not guaranteed to work for all models, though, in practice, few problems were encountered. All models described in the example in Section 4.4 converged quickly. However, if no convergence is obtained for a certain model, the constants defined at the beginning of the program, which regulate the step size, the maximum "error" (which is a measure for the distance from convergence), and the maximum number of iterations may be modified.

The following method is used for choosing a step size. The program starts with *step* = *MaxStepSize*, and keeps halving the value of *step* while the new estimates yield no decrease of the error function *error[m]* defined by (5.8). However, the halving is stopped when *step* < *MinStepSize*. In that case, *step* = *MaxStepSize* is used instead. This is done because it is assumed that, if *MinStepSize* is reached, that the error function does not yield a good indication of convergence, so a "jump" is made using *MaxStepSize*. If the method does not work, other values of *MinStepSize* and *MaxStepSize* may be tried.

The program listed below can be obtained by sending an e-mail message to the address

fsw.mto@kub.nl

with the message "bergsma-1" in the header. The program will then be sent to you automatically. A more user friendly program should be available very soon, and can be obtained from the same e-mail address by putting "bergsma-2" in the header.

```
MLE[n_,start_,g_,zeta_,Zt_,Mt_,X_] := Module[

   {MaxStepSize   = 1,
    MinStepSize   = .1,
    MaxError      = 10.^-10,
    MaxIterations = 100,
    v,m,step,error,iterate },

    v[m_,step_] := v[m,step] = Module[
        {Ut,mu,hm,Htm,Hm,lambda},
        Ut     = NullSpace[Transpose[X]];
        mu     = Mt.m;
        hm     = Ut.g[zeta[mu]];
        Htm    = (m*#)&/@ (Ut.(g'[zeta[mu]]*Zt[mu].Mt));
        Hm     = Transpose[Htm];
        lambda = -Inverse[Htm.(1/m*Hm)].(Htm.(n/m-1)+hm);
        Log[m] + step * 1/m * (n - m + Hm.lambda)  ];

    error[m_] := (v[m,1]-Log[m]).(m*(v[m,1]-Log[m]));

    iterate[m_,step_:MaxStepSize] := Module[ {newm},
        newm = Exp[v[m,step]];
        Print[ N[step], "   ", error[newm] ];
        Which[
            step < MinStepSize,     Exp[v[m,MaxStepSize]],
            error[newm] > error[m], iterate[m,step/2],
            True,                   newm ]  ];

    Print["stepsize, error :"];
```

```
FixedPoint[ iterate, start, MaxIterations,
    SameTest -> (error[#]<MaxError&) ]  ]
```

The **t**-functions used for evaluating measures that can be written in "exp-log" notation described in Section 4.3.1 are implemented as follows:

```
t[pi_,{a_,c_,0}]  := pi;
t[pi_,{a_,c_,i_}] :=
    Exp[c[i-1].Log[a[i-1].t[pi,{a,c,i-1}]]];

T[pi_,{a_,c_,0}]  := IdentityMatrix[Length[pi]];
T[pi_,{a_,c_,i_}] := t[pi,{a,c,i}] * c[i-1] .
    (1/a[i-1].t[pi,{a,c,i-1}]*a[i-1].T[pi,{a,c,i-1}])
```

As an example, we demonstrate how, for a 3×3 table, the marginal homogeneity model defined by the constraints

$$\log \frac{m_{1+}}{m_{2+}} = \beta_1 \qquad \log \frac{m_{2+}}{m_{3+}} = \beta_2$$
$$\log \frac{m_{+1}}{m_{+2}} = \beta_1 \qquad \log \frac{m_{+2}}{m_{+3}} = \beta_2$$

can be estimated (note that the constraints reduce to $m_{i+} = m_{+i}$, $i = 1, 2$). The model matrices and the functions $\boldsymbol{\zeta}$ and $\mathbf{Z}$ can be implemented as

```
Mt = {{1,1,1, 0,0,0, 0,0,0},
      {0,0,0, 1,1,1, 0,0,0},
      {0,0,0, 0,0,0, 1,1,1},
      {1,0,0, 1,0,0, 1,0,0},
      {0,1,0, 0,1,0, 0,1,0},
      {0,0,1, 0,0,1, 0,0,1} };

X = {{1,0},{0,1},{1,0},{0,1}};

at[0] = IdentityMatrix[6];
ct[0] = {{1,-1, 0,   0, 0, 0},
         {0, 1,-1,   0, 0, 0},
         {0, 0, 0,   1,-1, 0},
```

```
          {0, 0, 0,   0, 1,-1} };

zeta[mu_] := t[mu,{at,ct,1}];
Zt[mu_]   := T[mu,{at,ct,1}];
```

After specifying the observed frequencies, the estimation process can be started as follows:

```
n  = N[{1,2,3,4,5,6,7,8,9}];
estimates = MLE[n,n,Log,zeta,Zt,Mt,X]
```

where the observed frequencies are taken as starting estimates, and the log link is used. The result is

```
{1., 2.83808, 5.17758, 3.08809, 5., 6.85934, 4.92757,
7.10934, 9.}
```

As a second example, we demonstrate how the model asserting that gamma is zero in a 2×3 table can be fitted. The matrices $\mathbf{A}_i$, $\mathbf{C}_i$, and $\mathbf{E}$ defining gamma are programmed as follows.

```
at[0] = {
     {1,0,0, 0,0,0},
     {0,1,0, 0,0,0},
     {0,0,0, 0,1,1},
     {0,0,0, 0,0,1},
     {0,1,0, 0,0,0},
     {0,0,1, 0,0,0},
     {0,0,0, 1,0,0},
     {0,0,0, 1,1,0} };

ct[0] = {
     {1,0,1,0, 0,0,0,0},
     {0,1,0,1, 0,0,0,0},
     {0,0,0,0, 1,0,1,0},
     {0,0,0,0, 0,1,0,1} };

at[1] = 2 {{1,0},{0,1},{1,1}} . {{1,1,0,0},{0,0,1,1}};

ct[1] = {{1,0,-1},{0,1,-1}};
```

```
et = {{1,-1}};
```

The functions $\boldsymbol{\zeta}$, $\mathbf{Z}'$, and the log-hyperbolic link (see (4.4) can be implemented as

```
zeta[n_]          := et . t[n,{at,ct,2}];
Zt[n_]            := et . T[n,{at,ct,2}];
loghyperbolic[z_] := Log[ (1+z)/(1-z) ];
```

The design matrix $\mathbf{X}$ is the 1×1 zero matrix, since it is tested that gamma is zero. The marginals $\boldsymbol{\mu}$ are the joint frequencies $\mathbf{m}$, so $\mathbf{M}'$ is the 6×6 identity matrix:

```
X = {{0}};
Mt = IdentityMatrix[6];
```

At this point, only the observed vector $\mathbf{n}$ needs to be specified, and the estimation process can begin:

```
n = N[{1,2,3,4,5,6}];
estimates = MLE[n,n,loghyperbolic,zeta,Zt,Mt,X]
```

The result is

```
{1.34087, 2.1377, 2.52143, 3.62284, 4.87149, 6.50566}
```

As we see in a symbolic programming language like *Mathematica*, programs can be written in a very compact way. Similar programs can also be written in *Maple* or *S-plus*. A lot of programming time can be saved by programming in such a language, rather than, for instance, languages such as *Pascal*, *'C'*, or *Fortran*.

References

Agresti, A. (1984). *Analysis of ordinal categorical data.* New York: Wiley.

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

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, *Vol. 7, No. 1*, 131-177.

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Agresti, A., & Lang, J. B. (1993). A proportional odds model with subject-specific effects for repeated ordered responses. *Biometrika*, *80*, 527-534.

Agresti, A., Lipsitz, S., & Lang, J. B. (1992). Comparing marginal distributions of large, sparse contingency tables. *Comp. Stat. and Data Anal.*, *14*, 55-73.

Agresti, A., & Yang, M. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Comput. Stat. Data Anal.*, *5*, 9-21.

Aitchison, A. (1962). Large-sample restricted parametric tests. *J. Roy. Statist. Soc. Ser. B*, *24*, 234-250.

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.*, *29*, 813-828.

Aitchison, J., & Silvey, S. D. (1960). Maximum-likelihood estimation procedures and associated tests of significance. *J. Roy. Statist. Soc. Ser. B*, *22*, 154-171.

Alba, R. D. (1987). Interpreting the parameters of loglinear models. s. long (ed.). *Common problems/proper solutions*, 258-287.

Altham, P. M. E. (1984). Improving the precision of estimation by fitting a model. *J. Roy. Statist. Soc. Ser. B*, *46*, 118-119.

Anscombe, F. (1981). *Computing in statistical science through apl.* New York: Springer-Verlag.

Becker, M. P. (1994). Analysis of repeated categorical measurements using models for marginal distributions: an application to trends in attitudes on legalized abortion. In P. V. Marsden (Ed.), *Sociological methodology.* Oxford, UK: Blackwell.

Becker, M. P., & Balagtas, C. C. (1991). A non-loglinear model for binary crossover data. *Unpublished technical report.*

Bennett, B. M. (1967). Tests of hypotheses concerning matched samples. *J. R. Statist. Soc. Ser. B*, *29*, 468-474.

Berkson, J. (1978). In dispraise of the exact test. *J. Statist. Plan. Infer.*, *2*, 27-42.

Bertsekas, D. P. (1982). *Constrained optimization and lagrange multiplier methods.* New York: Academic Press.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Amer. Statis. Assoc.*, *61*, 228-235.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B*, *25*, 220-233.

Bishop, Y. V. V., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B*, *43*, 41-45.

Carlin, J. B., & Louis, T. A. (1996). *Bayes and empirical bayes methods for data analysis.* London: Chapman & Hall.

Caussinus, H. (1965). Contribution à l'analyse statistique des tablaux de correlation. *Annales de la faculté des sciences de l'université de toulouse*, *29*, 77-182.

Clayton, D. G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika*, *61*, 525-531.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables.* Thousand Oaks, Ca: Sage.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Psychological Bulletin*, *70*, 213-220.

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B*, *49*, 1-39.

Croon, M. A., Bergsma, W. P., & Hagenaars, J. A. (2000). Analyzing change in categorical variables by generalized log-linear models. *Sociol. Methods & Res.*, *Vol. 29 No 2*, 195–229.

Csiszár, I. (1975). i-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, *3*, 146-158.

Csiszár, I. (1989). A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *The Annals of Statistics*, *17*, 1409–1413.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, *42*, 909-917.

Darroch, J. N., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Stat.*, *43*, 1470-1480.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B*, *39*, 1-38.

Dennis, J. E., & Schnabel, R. B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations.* Englewood Cliffs, NJ: Prentice-Hall.

Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford: Oxford University Press.

Duncan, O. D. (1979). *Testing key hypotheses in panel analysis. in k. f. schuessler (ed.),* sociological methodology 1980 *p. 279-289.* San Fransisco: Jossey-Bass.

Duncan, O. D. (1981). *Two faces of panel analysis: Parallels with comparative cross-sectional analysis and time-lagged association. in s. leinhard (ed.),* sociological methodology 1981 *p. 281-318.* San Fransisco: Jossey-Bass.

Fienberg, S. (1980). *The analysis of cross-classified categorical data.* Cambridge, Mass.: MIT Press.

Firth, D. (1989). Marginal homogeneity and the superposition of latin squares. *Biometrika*, *76*, 179-182.

Fisher, R. A. (1934). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., & Laird, N. M. (1993). A likelihood-based method for analyzing longitudinal binary responses. *Biometrika*, *80*, 141-151.

Fitzmaurice, G. M., Laird, N. M., & Rotnitsky, A. G. (1993). Regression models for discrete longitudinal responses (with discussion). *Statistical Science*, *8*, 284-309.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.*, *33*, 613-619.

Fleiss, J. L., & Everett, B. S. (1971). Comparing the marginal totals of square contingency tables. *Psychol. Bulletin*, *72*, 323-327.

Forthofer, R. N., & Koch, G. G. (1973). An analysis for compounded functions of categorical data. *Biometrics*, *29*, 143-157.

Fryer, J. G. (1971). On the homogeneity of marginal distributions of a multidimensional contingency table. *J. Roy. Statist. Soc. Ser. A*, *134*, 368-371.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Glonek, G. J. N., Darroch, J. N., & Speed, T. P. (1988). On the existence of maximum likelihood estimators for hierarchical loglinear models. *Scand. J. Statist.*, *15*, 187-193.

Glonek, G. J. N., & McCullagh, P. (1995). Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B*, *57*, 533-546.

Gokhale, D. V. (1973). Iterative maximum likelihood estimation for discrete distributions. *Sankhya*, *B 35*, 293-298.

Goodman, L. A. (1973a). Causal analysis of panel studies and other kinds of surveys. *American Journal of Sociology*, *78*, 1153-1191.

Goodman, L. A. (1973b). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, *60*, 179-192.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statis. Assoc.*, *74*, 537-552.

Goodman, L. A. (1984). *The analysis of cross-classified data having ordered categories.* Cambridge, Mass.: Harvard University Press.

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications.* New York: Springer-Verlag.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B*, *46*, 149-192.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, *25*, 489-504.

Haber, M. (1985). Maximum likelihood methods for linear and loglinear models in categorical data. *Comput. Statist. Data anal.*, *3*, 1-10.

Haber, M., & Brown, M. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *J. Amer. Statis. Assoc.*, *81*, 477-482.

Haberman, S. J. (1973). The analysis of residuals in cross-classification tables. *Biometrics*, *29*, 205-220.

Haberman, S. J. (1974). *The analysis of frequency data.* Chicago: University of Chicago Press.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected counts. *Ann. Stat.*, *5*, 815-841.

Haberman, S. J. (1978). *Analysis of qualitative data. vol.1, introduction topics.* New York: Academic Press.

Haberman, S. J. (1979). *Analysis of qualitative data. vol.2, new developments.* New York: Academic Press.

Hagenaars, J. A. (1990). *Categorical longitudinal data.* Newbury Park: Sage.

Hagenaars, J. A. (1992). Analyzing categorical longitudinal data using marginal homogeneity models. *Statistica Applicata*, *4*, 763-771.

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1967). *Inequalities.* Cambridge: Cambridge University Press.

Ireland, C. T., Ku, H. H., & Kullback, S. (1968). Minimum discrimination information estimation. *Biometrics*, *24*, 707-713.

Jennrich, R. I., & Moore, R. H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proc. Statistical Computing Section, Amer. Statist. Assoc.*, 57-65.

Jørgensen, B. (1984). The delta algorithm. *Int. Stat. Review*, *52*, 283-300.

Kendall, M. G. (1945). The treatment of ties in rank problems. *Biometrika*, *33*, 239-251.

Koehler, K. (1986). Goodness-of-fit tests for loglinear models in sparse contingency tables. *J. Amer. Statis. Assoc.*, *81*, 483-493.

Koehler, K., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statis. Assoc.*, *75*, 336-344.

Krauth, J. (1985). A comparison of tests for marginal homogeneity in square contingency tables. *Biom. J.*, *27*, 3-15.

Kritzer, H. M. (1977). Analyzing measures of association derived from contingency tables. *Sociol. Methods Res.*, *5*, 35-50.

Kullback, S. (1959). *Information theory and statistics.* New York: Wiley.

Laird, N. M. (1991). Topics in likelihood-based methods for longitudinal data analyis. *Statistica Sinica*, *1*, 33-50.

Landis, J. R., & Koch, G. G. (1979). The analysis of categorical data in longitudinal studies of behavioral development. *In J. R. Nesselroade and P. B. Baltes (eds.): Longitudinal Reasearch in the Study of Behavior and Development*, 233-262.

Lang, J. B. (1996a). Maximum likelihood methods for a generalized class of log-linear models. *Ann. Stat.*, *24*, 726-752.

Lang, J. B. (1996b). On the comparison of multinomial and poisson log-linear models. *J. Roy. Statist. Soc. Ser. B*, *58*, 253-266.

Lang, J. B. (1996c). On the partitioning of goodness-of-fit statistics for multivariate categorical response models. *J. Am. Stat. Assoc.*, *91*, 1017-1023.

Lang, J. B., & Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *J. Am. Stat. Ass.*, *89*, 625-632.

Larntz, K. (1978). Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *J. Am. Stat. Assoc.*, *73*, 253-263.

Lauritzen, S. L. (1996). *Graphical models.* Oxford: Clarendon Press.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13-22.

Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. Roy. Statist. Soc. Ser. B*, *54*, 3-40.

Lindsey, J. K. (1993). *Models for repeated measurements.* Oxford: Oxford University Press.

Lipsitz, S. R., Laird, N. M., & Harrington, D. P. (1990). Finding the design matrix for the marginal homogeneity model. *Biometrika*, *77*, 353-358.

Madansky, A. (1963). Tests of homogeneity for correlated samples. *J. Amer. Statis. Assoc.*, *58*, 97-119.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory.* Oxford: Oxford University Press.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. Ser. B*, *42*, 109-142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman and Hall.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrica*, *12*, 153-157.

Molenberghs, G., & Lesaffre, E. (1994). Marginal modelling of multivariate categorical data. *J. Am. Statist. Assoc.*, *89*, 633-644.

Neyman, J. (1949). *Contributions to the theory of the* χ^2 *test. in* proceedings of the first berkeley symposium on mathematical statistics and probability*, ed. by j. neyman.* Berkeley: University of California Press.

Pierce, D. A., & Schafer, D. W. (1986). Residuals in generalized linear models. *J. Amer. Statis. Assoc.*, *81*, 977-983.

Pratt, J. W. (1981). Concavity of the log likelihood. *J. Amer. Statis. Assoc.*, *76*, 103-106.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference.* Chichester: Wiley.

Schollenberger, J., Agresti, A., & Wackerly, D. (1979). Measuring association and modelling relationships between interval and ordinal variables. *Proc. Soc. Statist. Sec., ASA*, 624-626.

Searle, S. R. (1982). *Matrix algebra useful for statistics.* New York: Wiley.

Semenya, K., & Koch, G. G. (1980). *Compound function and linear model methods for the multivariate analysis of ordinal categorical data.* Mimeo Series No. 1323. Chapel Hill: University of North Carolina Institute of Statistics.

Silvey, S. D. (1959). The lagrange multiplier test. *Ann. Math. Stat.*, *30*, 389-407.

Simon, G. (1973). Additivity of information in exponential family probability laws. *J. Amer. Statis. Assoc.*, *68*, 478-482.

Sobel, M. E. (1988). Some models for the multi-way contingency table with a one-to-one correspondence among categories. *In C. C. Clogg (ed.): Sociological Methodology 1988*, 165-192.

Sobel, M. E. (1995). *The analysis of contingency tables, in "handbook of statistical modelling for the social and behavioural sciences", g. arminger, c. c. clogg, and m. e. sobel (eds.).* New York: Plenum Press.

Sobel, M. E., Hout, M., & Duncan, O. D. (1985). Exchange, structure and symmetry in occupational mobility. *American Journal of Sociology*, *91*, 359–372.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *Amer. Sociol. Review*, *27*, 799-811.

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry*, *17*, 83-87.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, *42*, 412-416.

Vermunt, J. K. (1997). *Log-linear models for event histories.* Thousand Oaks: Sage.

Wahrendorf, J. (1980). Inference in contingency tables with ordered categories using plackett's coefficient of association for bivariate distributions. *Biometrika*, *67*, 15-21.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, *54*, 426-482.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the gaussian method. *Biometrika*, *61*, 439-447.

Wolfram, S. (1996). *The mathematica book (3rd edition).* Cambridge: Wolfram Media/Cambridge University Press.

Samenvatting

De laatste dertig jaar hebben enorme ontwikkelingen op het gebied van methoden voor de analyse van categorische data plaatsgevonden. Categorische data worden veelvuldig gebruikt in sociaal-wetenschappelijk en biomedisch onderzoek. Het loglineaire model neemt in deze ontwikkelingen een centrale plaats in. Een beperking van het loglineaire model is echter dat het niet geschikt is voor het toetsen van hypothesen betreffende de relatie tussen (afhankelijke) *marginale* verdelingen. Onder de marginale verdeling van een of meer variabelen wordt verstaan de kansverdeling van die variabele ongeacht de overige onderzoeksvariabelen. Een eenvoudig voorbeeld van een hypothese betreffende marginalen levert het marginale-homogeniteitsmodel, dat bruikbaar is om te toetsen of gecorreleerde categorische variabelen een identieke verdeling hebben. Vanwege de afhankelijkheden tussen de waarnemingen kunnen standaard chi-kwadraat toetsen niet gebruikt worden, en moet men andere methoden toepassen. In dit boek wordt een algemene methode gepresenteerd voor het toetsen van een brede klasse van modellen betreffende marginale verdelingen.

De indeling van dit boek is als volgt. Hoofdstukken 1 tot en met 3 vormen het inleidende deel. In hoofdstuk 1 wordt de probleemstelling uitvoerig geschetst. Het loglineaire model wordt beschreven in hoofdstuk 2. Een aantal basisconcepten die gebruikt worden in de rest van het boek wordt hier gepresenteerd. Hoofdstuk 3 bevat een overzicht van de belangrijkste literatuur die betrekking heeft op marginale-homogeniteitsmodellen. Verschillende toetsings- en schattingsmethoden worden beschreven.

Hoofdstukken 4 en 5 vormen de kern van het boek. In hoofdstuk 4 wordt het *marginale model* gepresenteerd, een zeer algemeen model dat bruikbaar is voor de analyse van marginale verdelingen. Het marginale model generaliseert het loglineaire en marginale-homogeniteitsmodel van

het tweede en derde hoofdstuk. Zo kan men met behulp van dit model hypothesen over de verandering van associatie tussen twee variabelen toetsen, bijvoorbeeld: wordt de associatie tussen partijvoorkeur en voorkeur voor minister president sterker naarmate de verkiezingen naderen? Associatie kan op verschillende manieren gemeten worden. In de loglineaire traditie worden odds-ratios gebruikt. Het marginale model kan ook gebruikt worden voor andere associatiematen, zoals bijvoorbeeld gamma of Kendall's tau. Ook is een vergelijking van overeenstemmingsmaten, zoals bijvoorbeeld kappa, mogelijk. Behalve aan een uitleg van het marginale model, wordt in hoofdstuk 4 ook aandacht besteed aan de implementatie op een computer.

In het laatste hoofdstuk wordt een eenvoudig en efficiënt algoritme beschreven voor het schatten van celfrequenties met behulp van de methode van de meest aannemelijke schatters. Dit algoritme blijkt in de praktijk zeer goed te werken, en kan gebruikt worden voor kruistabellen met enkele miljoenen cellen. Een voordeel van het algoritme is ook dat het eenvoudig te programmeren is (voor een programmabeschrijving zie Appendix E) en dat het voor een zeer brede klasse van modellen gebruikt kan worden. Verder wordt in dit hoofdstuk een klasse van loglineaire modellen voor marginalen beschreven waarvoor de aannemelijkheidsfunctie een uniek maximum heeft gegeven de modelrestricties. Dit simplificeert het schatten van celfrequenties aanzienlijk omdat, indien een lokaal maximum is gevonden, men er zeker van kan zijn dat dit maximum ook globaal is.

Verder wordt een korte beschrijving gegeven van de gewogen kleinste kwadraten methode en van de zogeheten *generalized estimating equations* benadering. Hoewel veel statistici de voorkeur geven aan de methode van de meest aannemelijke schatters, kunnen deze alternatieve methoden voordelen bieden bij de analyse van zeer grote kruistabellen, die ontstaan als veel variabelen gebruikt worden.

Wanneer de schatters voor een marginaal model gevonden zijn, kunnen de hypotheses met behulp van standaard chi-kwadraat toetsen getoetst worden. Resultaten van Lang (1996c) over het partitioneren van toetsen worden gegeneraliseerd. Bovendien wordt een generalisatie van de zogeheten *adjusted residuals* (Haberman, 1973; Lang, 1996a) gepresenteerd. Aangetoond wordt hoe conditionele adjusted residuals en adjusted residuals voor verschillende maten berekend kunnen worden. Tot slot worden de asymptotische verdelingen van de verschillende schatters in dit hoofdstuk beschreven.