MATTHEWS, GLENDA BEVERLEY

MAXIMUM LIKELIHOOD ESTIMATION WHEN MODELLING IN TERMS OF CONSTRAINTS

PhD **UP** **1995**

Maximum likelihood estimation when modelling in terms of constraints

by

Glenda Beverley Matthews

Submitted in partial fulfilment of

the requirements for the degree of

Philosophiae Doctor

in the

Faculty of Science

University of Pretoria

Pretoria

February 1995

PREFACE

The research described in this thesis was carried out in the Department of Statistics, University of Pretoria and Department of Mathematical Statistics, University of Natal from February, 1993 to February, 1995, under the supervision of Professor N.A.S. Crowther.

These studies represent original work by the authoress and have not been submitted in any form to another University. Where use was made of the work of others it has been duly acknowledged in the text.

ACKNOWLEDGEMENTS

Firstly I wish to sincerely thank my supervisor, Professor N.A.S. Crowther for his guidance, advice and for the time set aside to accommodate this work in an already busy schedule.

I also wish to thank Mrs. Jackie Sylaides for her technical assistance in the typing of parts of this thesis, and to Professor L. Troskie for the use of the facilities within his department.

The financial assistance of the Centre for Science and Development towards this research is hereby acknowledged. Opinions in this thesis and conclusions arrived at, are those of the author and are not necessarily to be attributed to the Centre for Science and Development.

I am also indebted to the University of Natal which provided the financial assistance required to complete this research.

Finally I would like to thank my family for their encouragement and support.

ABSTRACT

Maximum likelihood estimation when modelling in terms of constraints

by

G. B. Matthews

Supervisor: **Professor N.A.S. Crowther**
Department of Statistics
University of Pretoria
Pretoria

A maximum likelihood (ML) estimation procedure is developed to find the mean of the exponential family subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, where $\boldsymbol{g}$ is a vector valued function of the mean $\boldsymbol{\mu}$, satisfying the usual regularity constraints. This result forms the basis of an iterative procedure whereby the ML estimates of the mean values of a particular model are found. The constraints on the mean vector may be linear or non-linear. The application of the procedure provides a very flexible method for modelling data either directly in terms of certain constraints or in terms of the implied constraints of the appropriate model. The approach accommodates any choice of model assuming any predetermined distribution of the error terms, provided that the covariance matrix of the error terms can be computed.

This estimation procedure is implemented in estimating the expected frequencies for models suitable for cross-classified data. Models included are the log-linear model, models of marginal homogeneity, symmetry quasi-symmetry and quasi-independence. The estimation procedure also proves useful when formulating certain models in terms of the cross-product ratios instead of the standard parametric versions. Contingency tables containing structural zeros are also investigated and models for such tables are expressed in terms of constraints so that the estimation procedure is once again successfully implemented.

Estimation is also carried out for the class of models called generalized linear models, where the error terms may have a Poisson, binomial, multinomial, normal, gamma or inverse Gaussian distribution. In most cases these models are written in terms of the implied constraints and the ML estimates for the mean values and the model parameters are found. Models considered here are the logit type models, logistic regression, the proportional hazards model as well as other ordinal models.

The problem of estimating a covariance matrix for the multivariate normal distribution, in the presence of certain constraints on the elements, is also solved by implementing the proposed estimation procedure.

From the applications, it is evident that this estimation procedure provides a very natural and flexible method for finding the ML estimates of the parameters for a wide variety of statistical models.

OPSOMMING

Maximum likelihood estimation when modelling in terms of constraints

deur

G. B. Matthews

Promotor: Professor N.A.S. Crowther
Departement Statistiek
Universiteit van Pretoria
Pretoria

'n Maksimumaanneemlikheid beramingsprosedure word ontwikkel om die verwagte waarde van die eksponensiaal familie onder die beperkings $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, waar $\boldsymbol{g}$ 'n vektorwaardige funksie van die gemiddelde $\boldsymbol{\mu}$ is, wat die gewone reëlmatigheids voorwaardes bevredig. Hierdie resultaat vorm die basis van 'n iteratiewe prosedure wat gebruik word om die maksimumaanneemlikheidsberamers (MAB's) van die verwagte waardes van 'n bepaalde model te vind. Die beperkings op die vektor van verwagte waardes kan lineêr of nie-lineêr wees. Die prosedure is 'n baie soepel metode om data te modelleer in terme van die spesifieke beperkings of in terme van die geïmpliseerde beperkings van die betrokke model. Die prosedure omvat enige modelkeuse met enige voorafbepaalde verdeling van die foutterme, mits die kovariansiematriks van die foutterme bepaal kan word.

Die beramingsprosedure word geïmplementeer om die verwagte frekwensies te beraam vir modelle wat geskik is vir kruisgeklassifiseerde data. Modelle hierby ingesluit is die loglineêre model, modelle van marginale homogeniteit, simmetrie, kwasi-simmetrie en kwasi-onafhanklikheid. Die prosedure is ook nuttig wanneer modelle in terme van kruisprodukverhoudings geformuleer word en nie in terme van die gebruiklike parametriese variasies nie. Gebeurlikheidstabelle met strukturele nulle word ook ondersoek en modelle vir sulke tabelle word in terme van beperkings uitgedruk sodat die beramingsprosedure weereens suksesvol toegepas word.

Beraming word ook uitgevoer vir die sogenaamde klas van veralgemeende lineêre modelle, waar die foutterme 'n Poisson-, binomiaal-, mutinomiaal-, normaal-, gamma- of inverse normaalverdeling kan besit. In die meeste gevalle word hierdie modelle in terme van die geïmpliseerde beperkings geskryf en die MAB's vir die verwagte waardes en die parameters in die model word dan gevind. Modelle wat hier beskou word, is die tipe logitmodelle, logistiese regressie, die "proportional hazards"-model sowel as ordinale modelle.

Die probleem van beraming van 'n kovariansiematriks van die meerveranderlike normaalverdeling wanneer daar sekere beperkings op die elemente van die matriks is, word ook opgelos met behulp van die voorgestelde beramingsprosedure.

Uit die toepassings is dit duidelik dat die beramingsprosedure 'n natuurlike en soepel metode is vir die bepaling van die parameters vir 'n wye verskeidenheid van statistiese modelle.

Contents

List of Tables

Chapter 1

INTRODUCTION

Statistical modelling is an important aspect in the analysis of many types of data. In formulating a suitable model for a data set, it is important to estimate the parameters in the model. R.A. Fisher was the first to study and establish optimum properties of the estimates found by maximizing the likelihood function, giving rise to the class of maximum likelihood (ML) estimates. The ML estimates are generally preferred because they possess the attractive properties of consistency, efficiency and asymptotic normality.

Alternative estimates of the parameters have been proposed. Grizzle, Starmer and Koch (1969) present the non-iterative weighted least squares (WLS) procedure to obtain the WLS estimates. The ML estimates can be obtained by an iterative use of the WLS estimate, where the weight matrix changes at each cycle. This procedure is called the iterative reweighted least squares (IRLS) procedure.

Neyman (1949) introduced a class of estimators called minimum chi-squared estimates (MMCS) and showed that they are best asymptotically normal estimators. Neyman shows that when the model holds, the MMCS estimators are asymptotically, as $n \to \infty$, equivalent to the ML estimators.

The iterative proportional fitting (IPF) algorithm, originally introduced by

Deming and Stephan (1940) has been widely used by Haberman (1974, 1978, 1979) and Bishop, Fienberg and Holland (1975) for parameter estimation in loglinear modelling. The method converges to the ML estimates and makes use of the so called sufficient configurations.

A popular method used to find the ML estimates is the Newton-Raphson (NR) procedure, which is employed for solving non-linear equations. The NR algorithm is an iterative procedure which converges to the ML estimates and makes use of a vector of first order partial derivatives and a matrix of second order partial derivatives with respect to the parameters in the model. The broad class of models referred to as generalized linear models, introduced by Nelder and Wedderburn (1972) has become a popular and well used method for modelling data. In this framework the observations y_i are independent, come from an exponential family and a function of their expectation is written as a linear model using an appropriate link function. Agresti (1990) and McCullagh and Nelder (1989) show that when the canonical link is used, the NR method and the Fisher scoring method, which uses the expected value of the second derivative matrix, are identical.

Maximum likelihood estimation subject to constraints is considered by Aitchison and Silvey (1958), in the context of a random variable X, whose distribution function F depends on s parameters $\theta_1, \cdots, \theta_s$, which are not mathematically independent but satisfy r functional relationships, $h_i(\theta_1, \cdots, \theta_s) = 0, \;\; i = 1, \cdots, r; \;\; r < s$. The Lagrangian multiplier method is used to find the ML estimates of the parameters, which are found numerically by a process of iteration. The estimates are shown to have an asymptotic normal distribution. Silvey (1959) further discusses the Lagrangian method and the mathematical conditions for the existence of the ML estimates. Wedderburn (1974) considers generalized linear models where the systematic component of the model is defined by a set of linear constraints. A least squares solution is presented and an application to the problem of marginal homogeneity in a contingency table is discussed.

Haber and Brown (1986) consider a maximum likelihood method for estimating the expected frequencies in a contingency table, when the expected frequencies are subject to certain linear constraints. A multinomial sampling procedure is used and an algorithm involving an iterative procedure is implemented to find the ML estimates for the expected frequencies.

In this thesis we present a maximum likelihood estimation procedure for the mean of the exponential family subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, where $\boldsymbol{g}$ is a vector valued function of $\boldsymbol{\mu}$, satisfying the usual regularity conditions. This result is then applied to find the ML estimates of the mean values expressed in terms of a suitable model. The proposed estimation procedure can be applied to a wide variety of modelling problems. In certain cases, such as in modelling frequency data, under the assumptions of symmetry or marginal homogeneity for a contingency table, the model may be naturally expressed in terms of constraints imposed on the expected frequencies. The proposed estimation procedure provides an appropriate approach for finding the ML estimates of the expected frequencies. In other cases the method can be employed by writing the model to be considered in terms of the implied constraints, which in turn can be written in the form $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$. The major advantage of this method is that it is very flexible for modelling data from different distributions in terms of constraints on the mean vector. These constraints may be linear or non-linear and can be easily formulated either directly or in terms of the implied constraints, as will be illustrated in the applications. Another advantage is that the estimation procedure can be easily implemented numerically using a matrix algebra package.

It is also necessary to emphasize the fact that the proposed estimation procedure provides the exact ML estimates on convergence. These estimates are thus numerically identical to those obtained by the NR and IRLS procedures. Some well known examples are used to show that the estimation procedure gives ML estimates which agree with the ML estimates obtained by standard statistical packages such as GLIM and SAS.

It should also be noted that although statistical packages such as GLIM and SAS provide the ML estimates for the parameters of models fitting into the class of generalized linear models, we show that the proposed estimation procedure can easily be implemented by using the implied constraints for the model. Additional constraints may be added to these models and this enhances the implementation of the proposed ML estimation procedure.

In Chapter 2 the exponential family is introduced and the maximum likelihood estimation procedure is formulated and proved in Proposition 1. In Proposition 2 we find the asymptotic covariance matrix for the ML estimator. Examples are presented to illustrate the procedure for Poisson sampling, multinomial sampling, negative multinomial sampling and the multivariate normal distribution. The Wishart distribution is also considered and an application is that of estimating a covariance matrix in the presence of constraints on some of the elements.

In Chapter 3 we consider the estimation procedure for models for data which are arranged in a contingency table. Measures of goodness of fit are summarized and estimation in the linear and loglinear model is discussed. An approach of using the constraints implied by the odds ratios provides an interesting alternative for estimation in the loglinear model. Models for ordinal data are also reviewed and the estimation procedure for these models is presented.

In Chapter 4 models for square tables are considered and the estimation procedure is also implemented for tables which have structural zeros and for the quasi-independence model. In the case of square tables, the ML estimates are found for the expected frequencies for the models of marginal homogeneity, conditional symmetry, diagonal-parameter-symmetry, and for mobility tables and models for rater agreement.

Chapter 5 presents the estimation procedure for logit models and the variations such as the cumulative logit model, which is used for an ordinal response variable. The proportional hazards model is also discussed and illustrated by means of an example. The estimation procedure for logistic regression and the extreme value distribution is also presented.

Finally in Chapter 6 we consider regression models where the observations may have a Poisson, gamma, normal or inverse Gaussian distribution and the estimation procedure is once more successfully implemented to provide the ML estimates for the parameters.

Programs implementing the estimation procedure for the examples considered in this work, have been written and make use of SAS/IML, which is a procedure of the statistical package SAS. This procedure provides a powerful and flexible interactive matrix language. The IML programs for the examples are included in the Appendix. The example number in the Appendix corresponds to the example number as it appears in the relevant chapter.

Chapter 2

A MAXIMUM LIKELIHOOD ESTIMATION PROCEDURE

The aim of this chapter is to introduce the exponential family and formulate an ML estimation procedure for the mean of the exponential family subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$. The result is then applied to the Poisson, multinomial, negative multinomial, Normal and Wishart distribution.

2.1 THE EXPONENTIAL FAMILY

The exponential family occupies a central position in the theory of the generalized linear model. Earlier work on the exponential family was done by Tweedie (1947) and more recent work has been presented by Barndorff-Nielsen (1978), Brown (1986) and Jorgensen (1986, 1987, 1992). Because of the importance of the exponential family in the work that follows, the probability function and some of the properties of the exponential family will be presented.

Let $\mathbf{X}$ be a $p\times 1$ random vector and $\boldsymbol{\theta}$ be a $p\times 1$ vector of parameters. Jorgensen (1987) defines the exponential dispersion model in the multivariate form as

$$p(\mathbf{x}, \lambda, \boldsymbol{\theta}) = a(\lambda, \mathbf{x}) \exp\{\lambda(\mathbf{x}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta}))\}, \quad \mathbf{x} \in \mathbb{R}^p, \tag{2.1}$$

where a and κ are given functions, $\boldsymbol{\theta} \in \Theta$, where $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \kappa(\boldsymbol{\theta}) < \infty\}$. Furthermore $\lambda > 0$ and $\sigma^2 = \frac{1}{\lambda}$ is called the dispersion parameter.

Barndorff-Nielsen (1978) defines the exponential family by

$$\begin{aligned} p(\mathbf{x},\boldsymbol{\theta}) &= a(\boldsymbol{\theta})b(\mathbf{x})\exp(\mathbf{x}'\boldsymbol{\theta}), \quad \mathbf{x} \in \mathbb{R}^p, \quad \boldsymbol{\theta} \in \mathcal{N} \\ &= b(\mathbf{x})\exp(\mathbf{x}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})) \ , \end{aligned} \tag{2.2}$$

where $\kappa(\boldsymbol{\theta}) = -\ln a(\boldsymbol{\theta})$ and is referred to as the "cumulant generating function" or the "log Laplace transform" , and $\mathcal{N}$ is the natural parameter space for the canonical parameter $\boldsymbol{\theta}$.

We shall use the definition of the exponential family which is given in (2.2), which reduces to (2.1) when $\lambda = 1$.

Suppose $\mathbf{X}$ is a continuous random vector, with probability density function (p.d.f.) belonging to the exponential family. The moment generating function of the exponential family is given by

$$\begin{aligned} \mathbf{M}_{\mathbf{X}}(\mathbf{t}) &= E[e^{\mathbf{t}'\mathbf{X}}] = E[e^{\mathbf{X}'\mathbf{t}}] \\ &= \int\cdots\int \mathbf{b}(\mathbf{x})\exp[\mathbf{x}'(\boldsymbol{\theta}+\mathbf{t}) - \kappa(\boldsymbol{\theta})]d\mathbf{x} \\ &= \exp[-\kappa(\boldsymbol{\theta})]\int\cdots\int \mathbf{b}(\mathbf{x})\exp[\mathbf{x}'(\boldsymbol{\theta}+\mathbf{t})]d\mathbf{x} \\ &= \exp[-\kappa(\boldsymbol{\theta})]\exp[\kappa(\boldsymbol{\theta}+\mathbf{t})] \ , \end{aligned}$$

since from the property of the p.d.f. we have

$$\int\cdots\int \mathbf{b}(\mathbf{x})\exp[\mathbf{x}'\boldsymbol{\theta}]d\mathbf{x} = \exp[\kappa(\boldsymbol{\theta})].$$

Now

$$\begin{aligned} \ln \mathbf{M}_{\mathbf{X}}(\mathbf{t}) &= \kappa(\boldsymbol{\theta}+\mathbf{t}) - \kappa(\boldsymbol{\theta}) \\ &= \kappa(\boldsymbol{\theta}) + \left[\frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\right]'\mathbf{t} + \tfrac{1}{2}\mathbf{t}'\left[\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta})\right]\mathbf{t} + r(\mathbf{t}) - \kappa(\boldsymbol{\theta}) \\ &= \left[\frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\right]'\mathbf{t} + \tfrac{1}{2}\mathbf{t}'\left[\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta})\right]\mathbf{t} + r(\mathbf{t}) \end{aligned}$$

where $\lim_{\mathbf{t}\to\mathbf{0}} \frac{r(\mathbf{t})}{||\mathbf{t}||^2} = 0$. (Magnus and Neudecker (1991) p.101).
Hence the mean vector and covariance matrix of $\mathbf{X}$ are given by

$$E(\mathbf{X}) = \frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \text{ and } \mathrm{Cov}(\mathbf{X}) = \frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\kappa(\boldsymbol{\theta}) = \mathbf{V} \ .$$

The mean value of $\mathbf{X}$, can also be written as $E(\mathbf{X}) = \boldsymbol{\mu} = \boldsymbol{\tau}(\boldsymbol{\theta})$ and $\boldsymbol{\tau}(\boldsymbol{\theta})$ is called the mean value mapping. Barndorff-Nielsen (1978, p.121) shows that $\boldsymbol{\tau}$ is a one-to-one both ways continuously differentiable mapping of $\mathrm{int}(\Theta)$ onto $\Omega = \boldsymbol{\tau}(\mathrm{int}(\Theta))$. Since $\boldsymbol{\tau}$ is one-to-one, the variance covariance matrix $\mathbf{V}$ is positive definite for any $\boldsymbol{\mu}$ in Ω, cf. Barndorff-Nielsen (1978, p.114).

The ML estimate of $\boldsymbol{\mu}$ without any constraints can be found as follows. The log-likelihood function is

$$\ln p(\mathbf{x}, \boldsymbol{\theta}) = \ln b(\mathbf{x}) + \mathbf{x}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta}).$$

Hence

$$\frac{\partial}{\partial\boldsymbol{\theta}} \ln p(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{x} - \frac{\partial}{\partial\boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) \ ,$$

which when set equal to zero, gives solution $\widehat{\boldsymbol{\mu}} = \mathbf{x}$, where $\mathbf{x}$ is a vector observation of $\mathbf{X}$. This is a maximum since

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \ln p(\mathbf{x}, \boldsymbol{\theta}) = -\frac{\partial^2\kappa(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -\mathrm{Cov}(\mathbf{X}) \ ,$$

and the matrix $\mathrm{Cov}(\mathbf{X})$ is positive definite.

Silvey (1959) gives seven assumptions containing the mathematical conditions required to ensure the existence of an ML estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, the vector of unknown parameters appearing in the probability function, when there are certain constraints of the form $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ imposed on $\boldsymbol{\theta}$. Brown (1986) provides a comprehensive discussion of the analytic properties of the exponential family. The properties of the exponential family satisfy the assumptions of Silvey and this ensures the existence of the ML estimate of the mean of the exponential distribution when there are constraints of the form $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$.

2.2 A MAXIMUM LIKELIHOOD ESTIMATION PROCEDURE

An ML estimation procedure for the mean of the exponential distribution subject to constraints, is formulated and proved in Proposition 1 and the asymptotic covariance matrix for the estimator is given in Proposition 2.

Proposition 1

Consider a random vector $\mathbf{X}$, with probability function belonging to the exponential family. Let $\boldsymbol{g}(\boldsymbol{\mu})$ be a continuous vector valued function of $\boldsymbol{\mu}$, for which the first order partial derivatives exist. Let $\mathbf{G}_\mu = \dfrac{\partial \boldsymbol{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ be the derivative of $\boldsymbol{g}(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$ and $\mathbf{G}_x = \left.\dfrac{\partial \boldsymbol{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\boldsymbol{x}}$.

The ML estimate of $\boldsymbol{\mu}$ subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, is given by

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu \mathbf{V})'(\mathbf{G}_x \mathbf{V} \mathbf{G}'_\mu)^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \ . \tag{2.3}$$

This result implies that the ML estimate must, in general, be obtained iteratively.

Proof:

Let $\boldsymbol{\gamma}$ be a vector of Lagrange multipliers. To find the ML estimate of $\boldsymbol{\mu}$ subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, we maximize

$$w(\mathbf{x}; \boldsymbol{\theta}; \boldsymbol{\gamma}) \quad = \quad \ln b(\mathbf{x}) + \mathbf{x}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta}) + \boldsymbol{\gamma}' \boldsymbol{g}(\boldsymbol{\mu}(\boldsymbol{\theta})).$$

Hence we find

$$\frac{\partial}{\partial \boldsymbol{\mu}} w(\mathbf{x}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\theta}} w(\mathbf{x}; \boldsymbol{\theta}; \boldsymbol{\gamma}) \left[\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}}\right] \ .$$

Since we set $\dfrac{\partial}{\partial \boldsymbol{\mu}} w(\mathbf{x}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = \mathbf{0}$ for a maximum, and since $\left[\dfrac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}}\right]$ is invertible for a regular exponential family, we need further only consider $\dfrac{\partial}{\partial \boldsymbol{\theta}} w(\mathbf{x}; \boldsymbol{\theta}; \boldsymbol{\gamma})$.

Thus

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} w(\mathbf{x};\boldsymbol{\theta};\boldsymbol{\gamma}) &= \mathbf{x} - \frac{\partial}{\partial \boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}}\left\{\boldsymbol{\gamma}'\boldsymbol{g}(\boldsymbol{\mu}(\boldsymbol{\theta}))\right\} \\
&= \mathbf{x} - \boldsymbol{\mu} + \left\{\frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{g}(\boldsymbol{\mu}(\boldsymbol{\theta}))\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]\right\}'\boldsymbol{\gamma} \\
&= \mathbf{x} - \boldsymbol{\mu} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\boldsymbol{\gamma}.
\end{aligned}$$

Setting $\frac{\partial}{\partial \boldsymbol{\theta}} w(\mathbf{x};\boldsymbol{\theta};\boldsymbol{\gamma}) = \mathbf{0}$, we get

$$\boldsymbol{\mu} = \mathbf{x} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\boldsymbol{\gamma}\ . \tag{2.4}$$

Using the linear Taylor expansion of $\boldsymbol{g}(\boldsymbol{\mu})$ about $\mathbf{x}$, we get

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{\mu}) &= \boldsymbol{g}\left(\mathbf{x} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\boldsymbol{\gamma}\right) \\
&= \boldsymbol{g}(\mathbf{x}) + \mathbf{G}_{x}\left(\mathbf{x} + \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\boldsymbol{\gamma} - \mathbf{x}\right) + o(||\mathbf{x} - \boldsymbol{\mu}||) \\
&= \boldsymbol{g}(\mathbf{x}) + \mathbf{G}_{x}\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\boldsymbol{\gamma} + o(||\mathbf{x} - \boldsymbol{\mu}||)\ .
\end{aligned}$$

Setting $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ and solving for $\boldsymbol{\gamma}$, gives

$$\boldsymbol{\gamma} = -\left(\mathbf{G}_{x}\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\right)^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||)\ .$$

Substituting $\boldsymbol{\gamma}$ in (2.4) we get

$$\widehat{\boldsymbol{\mu}}_{c} = \mathbf{x} - \left(\mathbf{G}_{\mu}\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]\right)'\left(\mathbf{G}_{x}\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]'\mathbf{G}_{\mu}'\right)^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||)\ .$$

Now

$$\frac{\partial \mu_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left\{\frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i}\right\} = \frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}\ .$$

Hence

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = \left[\frac{\partial \mu_i}{\partial \theta_j}\right] = \left[\frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}\right]$$

and

$$\left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}}\right]' = \left[\frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] = \mathbf{V},$$

from the fact that $\kappa(\boldsymbol{\theta})$ is twice differentiable at an interior point $\boldsymbol{\theta}_0 \in \mathcal{N}$, (Brown (1986) Theorem 2.2), and the Hessian matrix $\mathsf{H}\kappa(\boldsymbol{\theta})$,with ij-th element $\frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$ is symmetric (Magnus and Neudecker (1991) p.105). Therefore

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu \ \mathbf{V})'(\mathbf{G}_x \ \mathbf{V} \ \mathbf{G}_\mu')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \ ,$$

which is the required result.

Remark

We shall assume without loss of generality, that $\mathbf{G}_\mu$ and $\mathbf{G}_x$ have $r \leq p$ linearly independent rows, so that $\mathbf{G}_\mu\mathbf{V}\mathbf{G}_x$ is non-singular and $(\mathbf{G}_\mu\mathbf{V}\mathbf{G}_x)^{-1}$ therefore exists uniquely. If $\mathbf{G}_\mu\mathbf{V}\mathbf{G}_x$ is singular, then this implies that there are linearly dependent rows. In this case a generalized inverse of $\mathbf{G}_\mu\mathbf{V}\mathbf{G}_x$ may be used and $\widehat{\boldsymbol{\mu}}_c$ is an ML estimator.

Estimation

The variance covariance matrix $\mathbf{V}$ could be known, or it could be some function of $\boldsymbol{\mu}$, say $\mathbf{V}_\mu$. The iterative use of the estimation procedure thus depends on the form of $\mathbf{G}_\mu$ and $\mathbf{V}_\mu$. Consider, for example, the estimation procedure for the following cases:

(i) Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\mathbf{D}_\mu$ be a diagonal matrix with the elements of $\boldsymbol{\mu}$ on the principal diagonal. Suppose $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A}\ln(\boldsymbol{\mu})$ and $\mathbf{V} = \mathbf{D}_\mu$, then

$$\mathbf{G}_\mu = \frac{\partial}{\partial \boldsymbol{\mu}}\mathbf{A}\ln(\boldsymbol{\mu}) = \mathbf{A}\mathbf{D}_\mu^{-1} \text{ and } \mathbf{G}_x = \mathbf{A}\mathbf{D}_x^{-1} \ .$$

Thus

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_c &= \mathbf{x} - (\mathbf{A}\mathbf{D}_\mu^{-1}\mathbf{D}_\mu)'(\mathbf{A}\mathbf{D}_x^{-1}\mathbf{D}_\mu\mathbf{D}_\mu^{-1}\mathbf{A}')^{-1}\mathbf{A}\ln(\mathbf{x}) + o(||\mathbf{x}-\boldsymbol{\mu}||) \\ &= \mathbf{x} - \mathbf{A}'(\mathbf{A}\mathbf{D}_x^{-1}\mathbf{A}')^{-1}\mathbf{A}\ln(\mathbf{x}) + o(||\mathbf{x}-\boldsymbol{\mu}||)\ .\end{aligned}$$

Let $\boldsymbol{\mu}^{(r)}$ denote the rth approximation for the ML estimate $\widehat{\boldsymbol{\mu}}_c$. At step r in the iterative process $(r = 0, 1, 2, \ldots)$,

$$\boldsymbol{\mu}^{(r+1)} = \boldsymbol{\mu}^{(r)} - \mathbf{A}'(\mathbf{A}\mathbf{D}_{\mu^{(r)}}^{-1}\mathbf{A}')^{-1}\mathbf{A}\ln(\boldsymbol{\mu}^{(r)}), \tag{2.5}$$

where for $r = 0$, $\boldsymbol{\mu}^{(0)} = \mathbf{x}$, the vector of data. The process then converges to the ML estimate $\widehat{\boldsymbol{\mu}}_c$.

(ii) If $\boldsymbol{g}$ is a linear function of $\boldsymbol{\mu}$, say $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\mu}$ and $\mathbf{V} = \mathbf{D}_\mu$, then

$$\mathbf{G}_\mu = \mathbf{A} = \mathbf{G}_x$$

and (2.3) yields

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{A}\mathbf{D}_\mu)'(\mathbf{A}\mathbf{D}_\mu\mathbf{A}')^{-1}\mathbf{A}\mathbf{x}\ .$$

Using the notation of (i), the iterative process at step r $(r = 0, 1, 2, \ldots)$, becomes

$$\boldsymbol{\mu}^{(r+1)} = \mathbf{x} - (\mathbf{A}\mathbf{D}_{\mu^{(r)}})'(\mathbf{A}\mathbf{D}_{\mu^{(r)}}\mathbf{A}')^{-1}\mathbf{A}\mathbf{x}\ , \tag{2.6}$$

where, as in (i), $\boldsymbol{\mu}^{(0)} = \mathbf{x}$. The process then converges to the ML estimate $\widehat{\boldsymbol{\mu}}_c$.

(iii) In some cases, as in this simple illustration, the procedure implies a double iteration, namely over $\mathbf{x}$ and $\boldsymbol{\mu}$. Suppose the elements of $\mathbf{X} : 3 \times 1$ are independent Poisson random variables with parameter vector $\boldsymbol{\mu}$ and that the observed vector $\mathbf{x}' = (80\ ,\ 15\ ,\ 5)$. The model to be fitted is $\mu_i = \alpha\beta^{i-1}$. The model implies the restriction

$$\boldsymbol{g}(\boldsymbol{\mu}) = \mu_1\mu_3 - \mu_2^2 = 0\ .$$

In this case

$$\mathbf{V} = \mathbf{D}_\mu \ , \quad \mathbf{G}_\mu = (\mu_3 \ , \ -2\mu_2 \ , \ \mu_1), \quad \mathbf{G}_\mu \mathbf{D}_\mu = \left(\mu_1\mu_3 \ , \ -2\mu_2^2 \ , \ \mu_1\mu_3\right)$$

and

$$\mathbf{G}_x \mathbf{D}_\mu \mathbf{G}'_\mu = (x_1 + x_3)\mu_1\mu_3 + 4x_2\mu_2^2 \ .$$

In this example the ML estimate of $\boldsymbol{\mu}$ must be found iteratively from

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu \mathbf{D}_\mu)'(\mathbf{G}_x \mathbf{D}_\mu \mathbf{G}'_\mu)^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \ .$$

Iteration takes place over $\mathbf{x}$ and $\boldsymbol{\mu}$ and the procedure runs as follows:

1. Both $\mathbf{x}$ and $\boldsymbol{\mu}$ are initially estimated by the data. Now iterate over $\mathbf{x}$ until convergence is attained. The values for $\widehat{\boldsymbol{\mu}}_c$ during this stage of iteration, are

$$\begin{pmatrix} 78.5263 \\ 16.6579 \\ 3.5263 \end{pmatrix} \text{ and } \begin{pmatrix} 78.5311 \\ 16.6525 \\ 3.5311 \end{pmatrix} .$$

2. The last value $\begin{pmatrix} 78.5311 \\ 16.6525 \\ 3.5311 \end{pmatrix}$ now becomes the second estimate of $\boldsymbol{\mu}$ in $\mathbf{G}_\mu$ and $\mathbf{D}_\mu$ and once again we iterate over $\mathbf{x}$ with initial $\mathbf{x} = \begin{pmatrix} 80 \\ 15 \\ 5 \end{pmatrix}$.

   The values for $\widehat{\mu}_c$ during this set of iterations are

$$\begin{pmatrix} 78.7931 \\ 17.4138 \\ 3.7931 \end{pmatrix} \text{ and } \begin{pmatrix} 78.8218 \\ 17.3564 \\ 3.8218 \end{pmatrix} .$$

3. The vector $\begin{pmatrix} 78.8218 \\ 17.3564 \\ 3.8218 \end{pmatrix}$ now becomes the third estimate of $\boldsymbol{\mu}$ and is replaced in $\mathbf{G}_\mu$ and $\mathbf{D}_\mu$ and iteration begins again over $\mathbf{x}$, with initial $\mathbf{x} = \begin{pmatrix} 80 \\ 15 \\ 5 \end{pmatrix}$, as in 2.

4. The values for $\widehat{\boldsymbol{\mu}}_c$, in the next set of iterations over $\mathbf{x}$, with the new estimate for $\boldsymbol{\mu}$ and initial $\mathbf{x}$ used in steps 2 and 3, are

$$\begin{pmatrix} 78.7931 \\ 17.4138 \\ 3.7931 \end{pmatrix} \text{ and } \begin{pmatrix} 78.8218 \\ 17.3564 \\ 3.8218 \end{pmatrix} .$$

The last vector is the same as that in 3 and we have attained convergence. Hence

$$\widehat{\boldsymbol{\mu}}_c = \begin{pmatrix} 78.8218 \\ 17.3564 \\ 3.8218 \end{pmatrix} .$$

The same ML estimate may be obtained if the function $\boldsymbol{g}(\boldsymbol{\mu})$ is replaced by some other equivalent restriction, for example

$$\boldsymbol{g}(\boldsymbol{\mu}) = \sqrt{\mu_1\mu_3} - \mu_2 = 0 \text{ or } \boldsymbol{g}(\boldsymbol{\mu}) = \ln(\mu_1\mu_3) - 2\ln(\mu_2) = 0 .$$

In the last instance, $\widehat{\boldsymbol{\mu}}_c$ does not depend on $\boldsymbol{\mu}$ and iteration is only over $\mathbf{x}$. The same ML estimate is obtained in the case of multinomial sampling. The only difference is that $\mathbf{V}$ is now replaced by the multinomial covariance matrix.

Proposition 2

The asymptotic covariance matrix of $\hat{\boldsymbol{\mu}}_c$ is given by

$$\boldsymbol{\Sigma}_c = \mathbf{V}_\mu - (\mathbf{G}_\mu \mathbf{V}_\mu)'(\mathbf{G}_\mu \mathbf{V}_\mu \mathbf{G}_\mu')^{-1}\mathbf{G}_\mu \mathbf{V}_\mu \,. \tag{2.7}$$

Proof:

$$\hat{\boldsymbol{\mu}}_c \;=\; \mathbf{x} - (\mathbf{G}_\mu \mathbf{V}_\mu)'(\mathbf{G}_x \mathbf{V}_\mu \mathbf{G}_\mu')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||).$$

Now since $\mathbf{x} \xrightarrow{p} \boldsymbol{\mu}$ and $\mathbf{G}_x \xrightarrow{p} \mathbf{G}_\mu$, $\hat{\boldsymbol{\mu}}_c \xrightarrow{d} (\mathbf{I}\ \mathbf{U})\begin{pmatrix}\mathbf{x}\\ \boldsymbol{g}(\mathbf{x})\end{pmatrix}$, where $\mathbf{U} = -(\mathbf{G}_\mu \mathbf{V}_\mu)'(\mathbf{G}_\mu \mathbf{V}_\mu \mathbf{G}_\mu')^{-1}$. Making use of the "delta method" (Bishop, Fienberg and Holland (1975)), the asymptotic covariance matrix of $\begin{pmatrix}\mathbf{x}\\ \boldsymbol{g}(\mathbf{x})\end{pmatrix}$ is given by

$$\left[\frac{\partial}{\partial \mathbf{x}}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{g}(\mathbf{x})\end{pmatrix}\Bigg|_{\mathbf{x}=\boldsymbol{\mu}}\right] \mathbf{V}_\mu \left[\frac{\partial}{\partial \mathbf{x}}\begin{pmatrix}\mathbf{x}\\ \boldsymbol{g}(\mathbf{x})\end{pmatrix}\Bigg|_{\mathbf{x}=\boldsymbol{\mu}}\right]'$$

$$= \begin{bmatrix}\mathbf{I}\\ \mathbf{G}_\mu\end{bmatrix} \mathbf{V}_\mu[\mathbf{I}\ \ \mathbf{G}'_\mu] = \begin{bmatrix}\mathbf{V}_\mu & (\mathbf{G}_\mu \mathbf{V}_\mu)'\\ \mathbf{G}_\mu \mathbf{V}_\mu & \mathbf{G}_\mu \mathbf{V}_\mu \mathbf{G}'_\mu\end{bmatrix} .$$

Thus $\hat{\boldsymbol{\mu}}_c$ has asymptotic covariance matrix

$$\begin{aligned}\boldsymbol{\Sigma}_c &= (\mathbf{I}\ \mathbf{U})\begin{bmatrix}\mathbf{V}_\mu & (\mathbf{G}_\mu \mathbf{V}_\mu)'\\ \mathbf{G}_\mu \mathbf{V}_\mu & (\mathbf{G}_\mu \mathbf{V}_\mu \mathbf{G}_\mu')\end{bmatrix}\begin{pmatrix}\mathbf{I}\\ \mathbf{U}'\end{pmatrix}\\ &= \mathbf{V}_\mu - (\mathbf{G}_\mu \mathbf{V}_\mu)'(\mathbf{G}_\mu \mathbf{V}_\mu \mathbf{G}_\mu')^{-1}(\mathbf{G}_\mu \mathbf{V}_\mu)\,.\end{aligned}$$

Useful applications of the result in Proposition 1 are considered in the following examples. The results in these examples can be used to find the ML estimates of the parameters in models assuming the appropriate underlying distribution of the observations.

Example 2.1: The Poisson Distribution.

Let $X_i,\ i = 1, \ldots, p$ be independent Poisson variables with $E(X_i) = \mu_i$. The likelihood function is

$$l = exp(-\sum_{i=1}^{p} \mu_i) \prod_{i=1}^{p} \mu_i^{x_i} / \prod_{i=1}^{p} x_i! \,.$$

Thus the log-likelihood function is given by

$$L = \ln(l) = -\sum_{i=1}^{p} \mu_i + \sum_{i=1}^{p} x_i \ln(\mu_i) - \ln \prod_{i=1}^{p} x_i! \,.$$

Hence $\theta_i = \ln(\mu_i)$ and

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\mu_i}\,, \quad i = 1, \ldots, p.$$

Thus

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} = \begin{pmatrix} \frac{1}{\mu_1} & & & \mathbf{0} \\ & \frac{1}{\mu_2} & & \\ & & \ddots & \\ \mathbf{0} & & & \frac{1}{\mu_p} \end{pmatrix}$$

and

$$\left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}}\right)^{-1} = \mathbf{D}_\mu$$

where $\mathbf{D}_\mu$ is a diagonal matrix with the elements of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ on the principal diagonal. Thus from (2.3) the MLE of $\boldsymbol{\mu} = E(\mathbf{X})$ subject to some constraints, $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ is

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu\, \mathbf{D}_\mu)^{'}(\mathbf{G}_x\, \mathbf{D}_\mu \mathbf{G'}_\mu)^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||)\,. \qquad (2.8)$$

This result will be used in Chapter 3 for modelling cross-classified data when assuming a Poisson sampling scheme and for the Poisson regression problem in Chapter 6.

Example 2.2: The Multinomial Distribution.

Let $\mathbf{X} = (X_1, \ldots, X_k)$ have the multinomial distribution with

$$P(\mathbf{X} = \mathbf{x}) = \frac{n!}{x_1! x_2! \cdots x_k!} \pi_1^{x_2} \pi_2^{x_2} \cdots \pi_k^{x_k} \ ,$$

where $x_i \geq 0$; $i = 1, \ldots, k$; $\sum_{i=1}^{k} x_i = n$ and $\sum_{i=1}^{k} \pi_i = 1$. Then

$$P(\mathbf{X} = \mathbf{x}) = \exp[\mathbf{x}'\boldsymbol{\theta} - \kappa(\boldsymbol{\theta})]$$

where

$$\theta_i = \ln(\pi_i) \ \ , \ \ i = 1, \ldots, k$$

and

$$\kappa(\boldsymbol{\theta}) = n \ln \left(\sum_{i=1}^{k} e^{\theta_i} \right) .$$

Brown (1986) points out this exponential family is not full and can be reduced to a minimal family by considering $\mathbf{X}^* = (X_1, \ldots, X_{k-1})$ with parameter

$$\theta_i^* = \theta_i - \theta_k = \ln(\pi_i) - \ln(\pi_k) = \ln \left(\frac{\pi_i}{\pi_k} \right) .$$

The kernel of the log-likelihood function is given by

$$K(\boldsymbol{\pi}) = \sum_{i=1}^{k} x_i \ln(\pi_i) = \mathbf{x}' \ln(\boldsymbol{\pi}) .$$

Consider the constraints $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}$, which do not include the constraint $\sum_{i=1}^{k} \pi_i = \mathbf{1}'\boldsymbol{\pi} = 1$. Let δ and $\boldsymbol{\gamma}$ be Lagrange multipliers, then to find the ML estimate of $\boldsymbol{\pi}$ subject to the constraints above, we need to maximize

$$w(\mathbf{x}, \boldsymbol{\pi}) = \mathbf{x}' \ln(\boldsymbol{\pi}) + \delta(\mathbf{1}'\boldsymbol{\pi} - 1) + \boldsymbol{\gamma}' \boldsymbol{g}(\boldsymbol{\pi})$$

with respect to $\boldsymbol{\pi}$. Thus

$$\frac{\partial}{\partial \boldsymbol{\pi}} w(\mathbf{x}, \boldsymbol{\pi}) = \mathbf{D}_\pi^{-1} \mathbf{x} + \delta \mathbf{1} + \mathbf{G}_\pi' \gamma = \mathbf{0} \ . \tag{2.9}$$

Now premultiply equation (2.9) by $\boldsymbol{\pi}'$ and get

$$\mathbf{1}'\mathbf{x} + \delta + \boldsymbol{\pi}' \mathbf{G}_\pi' \gamma = 0$$

i.e.

$$n + \delta + \boldsymbol{\pi}'\mathbf{G}_\pi'\gamma = 0 \; .$$

Thus

$$\delta = -(n + \boldsymbol{\pi}'\mathbf{G}_\pi'\gamma) \; . \qquad (2.10)$$

Premultiply equation (2.9) by $\mathbf{D}_\pi$ and get

$$\mathbf{x} + \delta\boldsymbol{\pi} + \mathbf{D}_\pi\mathbf{G}_\pi'\boldsymbol{\gamma} = \mathbf{0} \; .$$

Substitute δ found in (2.10), in the latter equation and get

$$\mathbf{x} - \boldsymbol{\pi}(n + \boldsymbol{\pi}'\mathbf{G}_\pi'\gamma) + \mathbf{D}_\pi\mathbf{G}'\boldsymbol{\gamma} = \mathbf{0}$$

i.e.

$$\mathbf{x} - n\boldsymbol{\pi} + (\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{G}_\pi'\gamma = \mathbf{0} \; .$$

Dividing by n gives

$$\mathbf{p} - \boldsymbol{\pi} + \frac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{G}_\pi'\gamma = \mathbf{0}$$

or

$$\boldsymbol{\pi} = \mathbf{p} + \frac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{G}_\pi'\gamma \; . \qquad (2.11)$$

Thus

$$\boldsymbol{g}(\boldsymbol{\pi}) = \boldsymbol{g}\left(\mathbf{p} + \frac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi})'\mathbf{G}_\pi'\gamma\right) \; .$$

Use a Taylor series expansion about $\mathbf{p}$ and get

$$\boldsymbol{g}(\boldsymbol{\pi}) = \boldsymbol{g}(\mathbf{p}) + \mathbf{G}_p\left\{\frac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\right\}\mathbf{G}_\pi'\gamma + o(||\mathbf{p} - \boldsymbol{\pi}||) \; .$$

Setting $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}$, from the set of constraints and solving for $\boldsymbol{\gamma}$ we get

$$\gamma = -\left(\mathbf{G}_p\left\{\frac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\right\}\mathbf{G}_\pi'\right)^{-1}\boldsymbol{g}(\mathbf{p}) + o(||\mathbf{p} - \boldsymbol{\pi}||) \; .$$

Substituting for γ in (2.11) we get

$$\begin{aligned}\hat{\boldsymbol{\pi}}_c &= \mathbf{p} - \left(\mathbf{G}_\pi\left\{\tfrac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\right\}\right)'\left(\mathbf{G}_p\left\{\tfrac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')\right\}\mathbf{G}_\pi'\right)^{-1}\boldsymbol{g}(\mathbf{p}) + o(||\mathbf{p} - \boldsymbol{\pi}||) \\ &= \mathbf{p} - (\mathbf{G}_\pi\boldsymbol{\Sigma}_M)'(\mathbf{G}_p\boldsymbol{\Sigma}_M\mathbf{G}_\pi')^{-1}\boldsymbol{g}(\mathbf{p}) + o(||\mathbf{p} - \boldsymbol{\pi}||) \; ,\end{aligned}$$

where $\mathbf{\Sigma}_M = \frac{1}{n}(\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')$ is the covariance matrix of the multinomial distribution. The above-mentioned result for the multinomial distribution, which is not a full exponential family, has the form of (2.3) , with $\mathbf{V}$ replaced by $\mathbf{\Sigma}_M$ which is singular. Thus (2.3) which was proved for the full regular exponential class may be used with the multinomial covariance matrix.

Example 2.3: The Negative Multinomial Distribution.
Consider $r+1$ mutually exclusive events E_i, $i = 0, 1, \ldots, r$ and let p_i be the probability of an occurrence of E_i in a single trial, then if the independent trials are terminated on obtaining the mth success (including the last trial) of E_0, the probability of obtaining $X_i = x_i$ occurrences of E_i, $i = 1, \ldots, r$ is given by

$$P(X_1 = x_1, \ldots, X_r = x_r) = \frac{(m + \sum_1^r x_i - 1)!}{(m-1)! \prod_1^r x_i!} \left(\prod_{i=1}^r p_i^{x_i} \right) p_0^m ,$$

where $p_0 = 1 - \sum_{i=1}^r p_i$ and m is assumed fixed. This is the probability function of the negative multinomial distribution with parameters m and $\mathbf{p} = (p_1, p_2, \ldots, p_r)'$.

The probability function of the negative multinomial distribution can be written in the form of (2.2) as follows

$$\begin{aligned} p(\mathbf{x}) &= \exp\left[\ln\left\{ \frac{(m + \sum x_i - 1)!}{(m-1)! \prod x_i!} \right\} + \sum_{i=1}^r x_i \ln(p_i) + m \ln(p_0) \right] \\ &= b(\mathbf{x}) \exp\left\{ \mathbf{x}'\boldsymbol{\theta} + m \ln\left(1 - \sum_1^r p_i\right) \right\} \\ &= b(\mathbf{x}) \exp\left\{ \mathbf{x}'\boldsymbol{\theta} + m \ln\left(1 - \sum_1^r e^{\theta_i}\right) \right\} , \end{aligned}$$

where $\theta_i = \ln(p_i)$. It also follows from (2.2) that the so called "cumulant generating function" is given by

$$\kappa(\boldsymbol{\theta}) = -m \ln\left(1 - \sum_1^r e^{\theta_i}\right) .$$

From the properties of the exponential family

$$E(\mathbf{X}) = \boldsymbol{\mu} = \frac{\partial}{\partial \boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) = \left[\frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i}\right],$$

i.e.

$$E(X_i) = \mu_i = \frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i}, \quad i = 1, \ldots, r .$$

Hence

$$\begin{aligned} \frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i} &= m\frac{\partial}{\partial \theta_i}\ln\left(1 - \sum_1^r e^{\theta_i}\right) \\ &= \frac{-m(-e^{\theta_i})}{(1 - \sum_1^r e^{\theta_i})} = \frac{mp_i}{p_0} . \end{aligned}$$

Thus

$$E(X_i) = \frac{mp_i}{p_0}, \quad i = 1, \ldots, r .$$

Furthermore

$$\text{Cov}(\mathbf{X}) = \frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \left\{\frac{\partial^2 \kappa(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\} .$$

Thus

$$\begin{aligned} \text{cov}(X_i, X_j) &= \frac{\partial}{\partial \theta_j}\left(\frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i}\right) = \frac{\partial}{\partial \theta_j}\left[me^{\theta_i}(1 - \sum_1^r e^{\theta_i})^{-1}\right] \\ &= me^{\theta_i}(-1)\left(1 - \sum_1^r e^{\theta_i}\right)^{-2}(-e^{\theta_j}) \\ &= \frac{mp_ip_j}{p_0^2} \quad , \ i \neq j . \end{aligned}$$

Also

$$\begin{aligned} \text{var}(X_i) &= \frac{\partial}{\partial \theta_i}\left(\frac{\partial \kappa(\boldsymbol{\theta})}{\partial \theta_i}\right) \\ &= \frac{\partial}{\partial \theta_i}\left[me^{\theta_i}(1 - \sum_1^r e^{\theta_i})^{-1}\right] \end{aligned}$$

$$= \frac{(1-\sum_1^r e^{\theta_i})me^{\theta_i} - me^{\theta_i}(-e^{\theta_i})}{(1-\sum_1^r e^{\theta_i})^2}$$

$$= \frac{mp_ip_0 + mp_i^2}{p_0^2} .$$

The moments may be summarized in matrix form in the following way.
Let $\mathbf{X}' = (m, X_1, \ldots, X_r) = (m, \mathbf{X}'_{(1)})$ denote the frequency vector (the frequency for the first cell included), where $\mathbf{X}_{(1)} = (X_1, \ldots, X_r)'$. Then

$$\begin{aligned} E(\mathbf{X}') &= \left(m, \frac{mp_1}{p_0}, \ldots, \frac{mp_r}{p_0}\right) = \frac{m}{p_0}(p_0, p_1, \ldots, p_r) \\ &= \frac{m}{p_0}\mathbf{p}'_* \text{ , where } \mathbf{p}'_* = (p_0, p_1, \ldots, p_r) . \end{aligned}$$

Thus

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} m \\ \boldsymbol{\mu}_{(1)} \end{pmatrix} ,$$

where

$$\boldsymbol{\mu}'_{(1)} = \left(\frac{mp_1}{p_0}, \frac{mp_2}{p_0}, \ldots, \frac{mp_r}{p_0}\right) = \frac{m}{p_0}\mathbf{p}' .$$

The covariance matrix of $\mathbf{X}_{(1)} = (X_1, \ldots, X_r)'$ is given by

$$\begin{aligned} \mathrm{Cov}(\mathbf{X}_{(1)}) &= \frac{m}{p_0}\begin{pmatrix} p_1 & & & \mathbf{0} \\ & p_2 & & \\ & & \ddots & \\ \mathbf{0} & & & p_r \end{pmatrix} + \frac{m^2}{mp_0^2}\begin{pmatrix} p_1^2 & p_1p_2 & \cdots & p_1p_r \\ p_2p_1 & p_2^2 & \cdots & p_2p_r \\ \vdots & \vdots & \cdots & \vdots \\ p_rp_1 & p_rp_2 & \cdots & p_r^2 \end{pmatrix} \\ &= \frac{m}{p_0}\mathbf{D}_{p_*} + \frac{m}{p_0^2}\mathbf{p}_*\mathbf{p}'_* \\ &= \mathbf{D}_{\mu_{(1)}} + \frac{1}{m}\boldsymbol{\mu}_{(1)}\boldsymbol{\mu}'_{(1)} = \boldsymbol{\Sigma}_{\mathbf{X}_{(1)}} , \end{aligned}$$

which is the covariance matrix of $X_1, \ldots, X_r$ given by Steyn et al. (1989) p.89). Hence

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{V} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{\mathbf{X}_{(1)}} \end{pmatrix} = \mathbf{D}_\mu + \frac{1}{m}\boldsymbol{\mu}\boldsymbol{\mu}' - (\boldsymbol{\mu},\ \mathbf{0}) - \begin{pmatrix} \boldsymbol{\mu}' \\ \mathbf{0} \end{pmatrix} .$$

We have shown in Proposition 1, that for a random vector belonging to the exponential family

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu \mathbf{V})'(\mathbf{G}_x \mathbf{V} \mathbf{G}_\mu')^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||),$$

is the ML estimate of $\boldsymbol{\mu}$ subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$. This estimation procedure can now be applied to a frequency distribution obtained from a negative multinomial sampling method.

Example 2.4 : The Multivariate Normal Distribution.
Let $\mathbf{X} : p \times 1$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is known. Barndorff-Nielsen (1978) gives the canonical parameters which are functions of $\boldsymbol{\mu}$, as $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.
Thus

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} \text{ and } \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = \boldsymbol{\Sigma}.$$

Let $\mathbf{G}$ denote a matrix of known constants, then the ML estimate of $\boldsymbol{\mu}$ subject to the constraints $\mathbf{G}\boldsymbol{\mu} = \mathbf{0}$, is given by

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}\boldsymbol{\Sigma})'(\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')^{-1}\mathbf{G}\mathbf{x} ,$$

where $\mathbf{x}$ is an ML estimate of $\boldsymbol{\mu}$. This is the result given in Theorem 1 of Crowther and Shaw (1989), which is derived using the conditional distribution of the multivariate normal distribution.

Example 2.5 : The Singular Normal Distribution.
In the case where $\mathbf{X} : p \times 1$ has the singular normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\mathrm{rank}(\boldsymbol{\Sigma}) = k < p$, let $\mathbf{X} = \mathbf{A}\mathbf{Y}$, where $\mathbf{A} : p \times k$, $\mathrm{rank}(\mathbf{A}) = k$ and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$ and $\mathbf{Y} \sim N_k(\boldsymbol{\nu}, \mathbf{I}_k)$. Thus $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\nu}$.

Then from Proposition 1, the ML estimate for $\boldsymbol{\nu}$ given the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ or equivalently $\boldsymbol{g}(\mathbf{A}\boldsymbol{\nu}) = \mathbf{0}$ is

$$\hat{\boldsymbol{\nu}}_c = \mathbf{y} - (\mathbf{GAI}_k)'(\mathbf{GAI}_k\mathbf{A}'\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{Ay}) + o(||\mathbf{y} - \boldsymbol{\nu}||)$$

since

$$\frac{\partial}{\partial \boldsymbol{\nu}}\boldsymbol{g}(\mathbf{A}\boldsymbol{\nu}) = \frac{\partial}{\partial \mathbf{A}\boldsymbol{\nu}}\boldsymbol{g}(\mathbf{A}\boldsymbol{\nu})\frac{\partial \mathbf{A}\boldsymbol{\nu}}{\partial \boldsymbol{\nu}} = \mathbf{GA}$$

Hence

$$\mathbf{A}\hat{\boldsymbol{\nu}}_c = \mathbf{Ay} - \mathbf{A}(\mathbf{GA})'(\mathbf{GAA}'\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{Ay}) + o(||\mathbf{Ay} - \mathbf{A}\boldsymbol{\nu}||)$$

Thus

$$\begin{aligned} \hat{\boldsymbol{\mu}}_c &= \mathbf{x} - \mathbf{AA}'\mathbf{G}'(\mathbf{GAA}'\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \\ &= \mathbf{x} - (\mathbf{G}\boldsymbol{\Sigma})'(\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||)\ . \end{aligned}$$

Example 2.6 : The Wishart Distribution.
Let $\mathbf{A} = \{a_{ij}\}$ be a $p \times p$ random symmetric positive definite matrix. $\mathbf{A}$ is said to have a Wishart distribution, $\mathbf{A} \sim W_p(n, \boldsymbol{\Sigma})$, if the joint p.d.f of the $\frac{1}{2}p(p+1)$ random elements $a_{11}, a_{12}, \ldots, a_{22}, \ldots, a_{pp}$ is given by

$$f(\mathbf{A}; n, \boldsymbol{\Sigma}) = \frac{|\mathbf{A}|^{\frac{1}{2}(n-p-1)}}{\Gamma_p(\frac{1}{2}n)|2\boldsymbol{\Sigma}|^{\frac{1}{2}n}} exp\left(-\frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\mathbf{A}\right), \quad \mathbf{A} > 0.$$

Let $\boldsymbol{\Sigma}^{-1} = \mathbf{B} = \{b_{ij}\}$, then

$$\begin{aligned} tr(\boldsymbol{\Sigma}^{-1}\mathbf{A}) &= tr(\mathbf{BA}) = \sum_i \sum_j a_{ij}b_{ij} \\ &= \sum_i a_{ii}b_{ii} + 2\sum_{i<j}\sum a_{ij}b_{ij}\ . \end{aligned}$$

The Wishart distribution belongs to the exponential family since the p.d.f. can be written in the form

$$f(\mathbf{A}; n, \boldsymbol{\Sigma}) = \frac{|\mathbf{A}|^{\frac{1}{2}(n-p-1)}}{\Gamma_p(\frac{1}{2}n)} exp\left(-\frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\mathbf{A} - \frac{1}{2}n\ln|2\boldsymbol{\Sigma}|\right), \quad \mathbf{A} > 0.$$

The canonical parameters are $\{-\frac{1}{2}b_{ii}\}$, $i = 1, \ldots, p$ and $\{-b_{ij}\}$, $i < j$, since

$$-\frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\mathbf{A}) = \sum_i a_{ii}(-\frac{1}{2}b_{ii}) + \sum_{i<j}\sum a_{ij}(-b_{ij}) \ .$$

A useful application is that of estimating a covariance matrix when there are constraints on the elements $\{\sigma_{ij}\}$.

Let $\mathbf{X}_\alpha$, $\alpha = 1, \ldots, n$ be n independent $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ random vectors. Then the matrix $\mathbf{A} = \sum_{\alpha=1}^{n} \mathbf{X}_\alpha \mathbf{X}'_\alpha$ has the Wishart distribution with n degrees of freedom and covariance matrix $\boldsymbol{\Sigma}$. Furthermore

$$E(\mathbf{A}) = n\boldsymbol{\Sigma} \ \text{ and } \ \text{cov}(a_{ij}, a_{kl}) = n(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}), \ \text{ for } \ i, j, k, l = 1, \ldots, p.$$

Consider the case where $p = 2$ and $\mathbf{X}_\alpha, \alpha = 1, \ldots, n$ are i.i.d. $N_2(\mathbf{0}, \boldsymbol{\Sigma})$. Let $\mathbf{Y} = \frac{1}{n}(a_{11}, a_{22}, a_{12})'$ and $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})'$, then $E(\mathbf{Y}) = \boldsymbol{\sigma}$, and

$$\text{Cov}(\mathbf{Y}) = \frac{1}{n}\begin{pmatrix} 2\sigma_{11} & 2\sigma_{12} & 2\sigma_{11}\sigma_{12} \\ 2\sigma_{12} & 2\sigma_{22} & 2\sigma_{22}\sigma_{12} \\ 2\sigma_{11}\sigma_{12} & 2\sigma_{22}\sigma_{12} & \sigma_{11}\sigma_{22} + \sigma_{12}^2 \end{pmatrix} = \mathbf{V} \ .$$

(i) Consider the case where say $\sigma_{11} = c$ and the ML estimate of $\boldsymbol{\sigma}$, must be found, subject to the above constraint.

Hence $g(\boldsymbol{\sigma}) = \sigma_{11} - c = 0$ and $\mathbf{G}_\sigma = \frac{\partial}{\partial \boldsymbol{\sigma}} g(\boldsymbol{\sigma}) = (1, 0, 0)$.
The ML estimation procedure may be expressed as follows

$$\hat{\boldsymbol{\sigma}}_c = \mathbf{y} - (\mathbf{G}_\sigma \mathbf{V})'(\mathbf{G}_y \mathbf{V} \mathbf{G}'_\sigma)^{-1} \boldsymbol{g}(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\sigma}||).$$

A double iteration is necessary.

If for example $\sigma_{11} = 4$, $n = 100$ and $\frac{1}{n}\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$, then

$$\frac{1}{n}\widehat{\mathbf{A}} = \begin{pmatrix} 4 & 2.667 \\ 2.667 & 3.444 \end{pmatrix}, \ \text{ or } \ \hat{\boldsymbol{\sigma}}_c = (4, 3.444, 2.667)'.$$

(ii) Suppose that the ML estimate of $\boldsymbol{\sigma}$ must be found subject to $\sigma_{11} = \frac{1}{2}\sigma_{22}$. This implies that $g(\boldsymbol{\sigma}) = 2\sigma_{11} - \sigma_{22} = 0$ and $\mathbf{G}_\sigma = (2, -1, 0)$. The form of the estimation procedure in (i) above may be employed to find the ML estimate of $\boldsymbol{\sigma}$ subject to the given constraint. If as in (i)

$$n = 100 \quad \text{and} \quad \frac{1}{n}\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$$

then

$$\hat{\boldsymbol{\sigma}}_c = (2.5, 5, 2)'.$$

(iii) Consider the case where the ML estimate of $\boldsymbol{\sigma}$ must be found subject to $\rho = c$. Thus the function $g(\boldsymbol{\sigma}) = \sigma_{12} - c\sqrt{\sigma_{11}\sigma_{22}} = 0$ and

$$\mathbf{G}_\sigma = \left(-\frac{1}{2}c\sqrt{\sigma_{22}/\sigma_{11}}, -\frac{1}{2}c\sqrt{\sigma_{11}/\sigma_{22}}, 1 \right).$$

If $\rho = \frac{1}{2}$, $n = 100$ and $\frac{1}{n}\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$, then

$$\hat{\boldsymbol{\sigma}}_c = (2.6667, 2.6667, 1.3333).$$

The IML programs for this example appear in the Appendix.

Chapter 3

ESTIMATION FOR MODELS RELATED TO CONTINGENCY TABLES

This chapter presents models suitable for data arranged in a contingency table. The estimation procedure introduced in Chapter 2 is implemented for the estimation of the expected frequencies, where the model under consideration is written in terms of constraints of the form $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, satisfying the conditions of Proposition 1. The sampling procedures giving the underlying probability assumptions for the contingency table are outlined and certain goodness of fit tests are summarized for convenience.

3.1 SAMPLING PROCEDURES

Suppose we observe counts $x_i, i = 1, \ldots, p$ which are the frequencies for the p cells of a contingency table and that the x_i are arranged into a vector $\mathbf{x} = (x_1, \ldots, x_p)'$. These frequencies are considered to be observations of a random variable X_i with mean $E(X_i) = F_i$ for cell i. F_i is called the expected frequency for cell i.

3.1.1 The Poisson Sampling Procedure

If we assume that each X_i is an independent Poisson random variable with mean F_i, then

$$P(X_i = x_i) = \frac{\exp(-F_i)(F_i)^{x_i}}{x_i!} \ , \ x_i = 0, 1, 2, \ldots \ .$$

The Poisson distribution has $E(X_i) = F_i$ and $\text{var}(X_i) = F_i$.
Thus for the Poisson sampling procedure the mean vector and variance covariance matrix of $\mathbf{X}$ are

$$E(\mathbf{X}) = \mathbf{F}$$

and

$$\begin{aligned} \text{Cov}(\mathbf{X}) &= \begin{bmatrix} \text{var}(X_1) & 0 & \cdots & 0 \\ 0 & \text{var}(X_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \text{var}(X_p) \end{bmatrix} \\ &= \mathbf{D}_F \end{aligned}$$

where $\mathbf{D}_F = \text{diag}(\mathbf{F})$ is a diagonal matrix, with the elements of $\mathbf{F}$ on the principal diagonal.

3.1.2 The Multinomial Sampling Procedure

Suppose that n independent observations are taken on p mutually exclusive categories and that X_i counts the number of observations in category i. If the probability that an observation falls in category i, is π_i, $i = 1, \ldots, p$, where $\pi_i \geq 0$ and $\sum_{i=1}^{p} \pi_i = 1$, then we have the multinomial probability distribution, for which

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \frac{n!}{\prod_{i=1}^{p} x_i!} (\pi_1^{x_1} \pi_2^{x_2} \cdots \pi_p^{x_p}) \ .$$

For this distribution

$$E(X_i) = n\pi_i \ , \ \text{var}(X_i) = n\pi_i(1 - \pi_i) \text{ and } \text{cov}(X_i, X_j) = -n\pi_i\pi_j \ \ , \ \ i \neq j \ .$$

The mean vector and covariance matrix for $\mathbf{X} = (X_1, \ldots, X_p)'$ are

$$E(\mathbf{X}) = n\boldsymbol{\pi} = \mathbf{F}$$

and

$$\text{Cov}(\mathbf{X}) = \mathbf{D}_F - \frac{1}{n}\mathbf{F}\mathbf{F}' .$$

3.1.3 Product Multinomial Sampling

Product multinomial sampling is best illustrated using an example. Consider a categorical response variable Y, say income, divided into 3 categories "low" "middle" and "high". Consider two explanatory variables say gender and age, where age is divided into three categories A, B and C. We therefore consider the following $2 \times 3 \times 3$ contingency table.

FREQUENCY TABLE FOR INCOME

Gender	Age	Income		
		Low	Middle	High
Male	A	$\pi_{(1)1}$	$\pi_{(1)2}$	$\pi_{(1)3}$
	B	$\pi_{(2)1}$	$\pi_{(2)2}$	$\pi_{(2)3}$
	C	$\pi_{(3)1}$	$\pi_{(3)2}$	$\pi_{(3)3}$
Female	A	$\pi_{(4)1}$	$\pi_{(4)2}$	$\pi_{(4)3}$
	B	$\pi_{(5)1}$	$\pi_{(5)2}$	$\pi_{(5)3}$
	C	$\pi_{(6)1}$	$\pi_{(6)2}$	$\pi_{(6)3}$

Gender has 2 levels (Male, Female) while age has 3 levels (A,B,C) which together form $2 \times 3 = 6$ so called sub-populations. Each of the six sub-populations follows a multinomial distribution. Thus for the ith sub-population, let $\pi_{(i)j}$ denote the probability of falling in category j, $j = 1, 2, 3$ and let X_{ij} denote the number of individuals in the ith sub-population falling into category j of the response variable, income. Hence for the ith sub-population

$$P[X_{i1} = x_{i1}, X_{i2} = x_{i2}, X_{i3} = x_{i3}] = \frac{n_i!}{\prod_{k=1}^{3} x_{ik}!} \left[(\pi_{(i)1})^{x_{i1}}(\pi_{(i)2})^{x_{i2}}(\pi_{(i)3})^{x_{i3}}\right]$$

where $\sum_{k=1}^{3} x_{ik} = n_i$ and $\sum_{k=1}^{3} \pi_{(i)k} = 1$.

Hence the joint probability distribution of $(X_{11}, X_{12}, X_{13}, \ldots, X_{63})$ is

$$\prod_{i=1}^{6} \left[\frac{n_i!}{\prod_{k=1}^{3} x_{ik}!} (\pi_{(i)1}^{x_{i1}} \pi_{(i)2}^{x_{i2}} \pi_{(i)3}^{x_{i3}}) \right] .$$

3.2 MEASURES OF GOODNESS OF FIT

Suppose that $\{\widehat{m}_i\}$ are the estimated expected frequencies for the contingency table on fitting an appropriate model to the data. The following statistics can be used to test the goodness of fit of a model :

(i) The Pearson Chi-squared Statistic

$$\chi^2 = \sum_{i=1}^{p} (x_i - \widehat{m}_i)^2 / \widehat{m}_i \ . \tag{3.1}$$

(ii) Neyman's Modified Chi-squared Statistic

$$\chi_N^2 = \sum_{i=1}^{p} (x_i - \widehat{m}_i)^2 / x_i \ . \tag{3.2}$$

(iii) The Freeman-Tukey Statistic

$$K_{FT}^2 = 4 \sum_{i=1}^{p} \left(\sqrt{x_i} - \sqrt{\widehat{m}_i} \right)^2 \ . \tag{3.3}$$

(iv) The Likelihood Ratio Statistic (LR)

Let $X_1, X_2, \ldots, X_k$ have the multinomial distribution $M(n; \pi_1, \pi_2, \ldots, \pi_k)$ with probability mass function

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k} ,$$

for $\boldsymbol{\pi} \in \boldsymbol{\Pi} = \{(\pi_1, \ldots, \pi_k) : \pi_i > 0 \text{ and } \sum_{i=1}^{k} \pi_i = 1\}$, $x_i \geq 0$, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} x_i = n$.

Suppose that we wish to test $H_o : \boldsymbol{\pi} = \boldsymbol{\pi}_0$. The likelihood ratio statistic for the test is

$$\boldsymbol{\Lambda} = \frac{L(\boldsymbol{\pi}_0)}{\sup_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \mathbf{L}(\boldsymbol{\pi})} .$$

The ML estimator of $\boldsymbol{\pi}$ is

$$\mathbf{p} = (p_1, p_2, \ldots, p_k) = \left(\frac{x_1}{n}, \frac{x_2}{n}, \ldots, \frac{x_k}{n} \right) .$$

Thus

$$\boldsymbol{\Lambda} = \frac{\dfrac{n!}{x_1! x_2! \cdots x_k!} \pi_{10}^{x_1} \pi_{20}^{x_2} \cdots \pi_{k0}^{x_k}}{\dfrac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}}$$

$$= \prod_{i=1}^{k} \left(\frac{\pi_{i0}}{p_i} \right)^{x_i} .$$

Suppose that under H_o, $\boldsymbol{\pi}_0 = \widehat{\mathbf{m}}/n = (\widehat{m}_1, \ldots, \widehat{m}_k)/n$ where $\widehat{m}_i$ is the expected frequency for the ith cell. Then

$$\boldsymbol{\Lambda} = \prod_{i=1}^{k} \left[\frac{(\widehat{m}_i / n)}{(x_i / n)} \right]^{x_i} = \prod_{i=1}^{k} \left(\frac{\widehat{m}_i}{x_i} \right)^{x_i}$$

and

$$\ln \boldsymbol{\Lambda} = \sum_{i=1}^{k} x_i \ln \left(\frac{\widehat{m}_i}{x_i} \right) .$$

Thus $-2 \ln \boldsymbol{\Lambda} = 2 \sum_{i=1}^{k} x_i \ln \left(\frac{x_i}{\widehat{m}_i} \right)$. This quantity is referred to as the likelihood ratio statistic. Most well known texts use the notation G^2 but we shall use the notation LR, so as to avoid confusion with the matrix **G**. Thus

$$LR = 2\sum_{i=1}^{k} x_i \ln\left(\frac{x_i}{\widehat{m}_i}\right) . \tag{3.4}$$

(BFH (1975) show that LR, K^2_{FT} and the Pearson χ^2- statistic all have an asymptotic χ^2-distribution).

(v) The Akaike Information Criterion.

Another useful criterion used to test the adequacy of a model is the Akaike Information Criterion (AIC), proposed by Akaike (1973). The AIC is related to the likelihood ratio statistic, LR. The AIC for a model having k free parameters to be estimated, is

$$\begin{aligned} AIC(k) &= -2L(\widehat{\theta}_k) + 2k \\ &= -2(\text{maximum loglikelihood of the model}) + 2k \ . \end{aligned}$$

For the multinomial distribution, the log-likelihood is

$$L(\boldsymbol{\pi}) = \ln\left(\frac{n!}{x_1!\cdots x_k!}\right) + \sum_{i=1}^{k} x_i \ln \pi_i$$

and

$$\begin{aligned} L(\widehat{\boldsymbol{\pi}}) &= \ln\left(\frac{n!}{x_1!\cdots x_k!}\right) + \sum_{i=1}^{k} x_i \ln\left(\frac{\widehat{m}_i}{n}\right) \\ &= \sum_{i=1}^{k} x_i \ln\left(\frac{\widehat{m}_i}{x_i}\right) + \sum_{i=1}^{k} x_i \ln(x_i) - \sum_{i=1}^{k} x_i \ln(n) + \ln\left(\frac{n!}{x_1!\cdots x_k!}\right) . \end{aligned}$$

Hence

$$\begin{aligned} -2L(\widehat{\boldsymbol{\pi}}) &= 2\sum_{i=1}^{k} x_i \ln\left(\frac{x_i}{\widehat{m}_i}\right) - 2\left[\sum_{i=1}^{k} x_i \ln(x_i) - n\ln(n) + \ln\left(\frac{n!}{x_1!\cdots x_k!}\right)\right] \\ &= LR + c \ , \end{aligned}$$

where

$$c = -2\left[\sum_{i=1}^{k} x_i \ln(x_i) - n\ln(n) + \ln\left(\frac{n!}{x_1!\cdots x_k!}\right)\right]$$

remains constant for the contingency table.

Hence $AIC(k) = (LR + 2k) + c$ and the quantity $LR + 2k$ in parentheses is often used as the AIC, since it is simply $AIC(k) - c$. To this end let

$$AIC^* = LR + 2k \; . \tag{3.5}$$

A model with a small AIC^*-value is preferable as models with too many parameters are penalized by the AIC-procedure.

(vi) The Deviance.
Nelder and Wedderburn (1972) introduced the so called deviance as a measure of discrepancy. In the context of the generalized linear model, let $\boldsymbol{\mu}$ denote the mean-value parameter and $\boldsymbol{\theta}$ denote the canonical parameter and let ϕ be some dispersion parameter. Let $L(\widehat{\boldsymbol{\mu}}, \phi, \mathbf{y})$ be the log-likelihood maximized over some vector of parameters $\boldsymbol{\beta}$ for a fixed value of ϕ and $\mathbf{L}(\mathbf{y}, \phi, \mathbf{y})$ be the maximum likelihood achievable in the saturated model, then the scaled deviance is defined as

$$D^* = 2[L(\widehat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - L(\mathbf{y}, \phi, \mathbf{y})]/\phi \; . \tag{3.6}$$

If $\phi = 1$, then the deviance is defined as

$$D = 2[L(\widehat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - L(\mathbf{y}, \phi, \mathbf{y})] \; . \tag{3.7}$$

As an example consider the form of the deviance for the Poisson distribution.
Let $Y_1, Y_2, \ldots, Y_n$ be n independent Poisson random variables with $E(Y_i) = \mu_i$. The log-likelihood function is

$$\begin{aligned} L(\boldsymbol{\mu}, \mathbf{y}) &= \ln\left[\prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}\right] \; . \\ &= -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \ln(\mu_i) - \ln\left(\prod_{i=1}^{n} y_i!\right) \; . \end{aligned}$$

Hence the deviance for a model with fitted values $\widehat{\mu_i}$, is

$$\begin{aligned} D &= 2\left[-\sum_{i=1}^{n}\widehat{\mu}_i + \sum_{i=1}^{n} y_i \ln(\widehat{\mu}_i) - \ln\left(\prod_{i=1}^{n} y_i!\right) - \left\{-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \ln(y_i) - \ln\left(\prod_{i=1}^{n} y_i!\right)\right\}\right] \\ &= 2\left[\sum_{i=1}^{n} y_i \ln\left(\frac{\widehat{\mu}_i}{y_i}\right) + \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)\right] . \end{aligned}$$

(vii) The Wald Statistic

Another statistic for testing the goodness of fit is the so called Wald statistic. If the model under consideration is formulated in terms of the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ and $\mathbf{G} = \left[\frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{g}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{x}}$, then the Wald statistic (under Poisson sampling) is

$$W = \boldsymbol{g}'(\mathbf{x})[\mathbf{G}\mathbf{D}_x\mathbf{G}']^{-1}\boldsymbol{g}(\mathbf{x}) , \tag{3.8}$$

where $\mathbf{D}_x = \text{diag}(\mathbf{x})$.

Wald (1943) shows that this statistic has a chi-squared distribution as limiting distribution with degrees of freedom the number of linearly independent constraints specified by $\boldsymbol{g}(\cdot)$.

Remark

Consider the minimum modified chi-squared criterion

$$Q = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{D}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}) .$$

Suppose that the model under consideration is expressed in terms of the implied constraints, say $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$. We now wish to find the value of $\boldsymbol{\mu}$ that minimizes Q subject to the constraints above. Let γ be a vector of Lagrange multipliers, then we must find the solution $\boldsymbol{\mu}$, which minimizes

$$Q^* = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{D}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}) + 2\gamma'\boldsymbol{g}(\boldsymbol{\mu}) .$$

Differentiating Q^* with respect to $\boldsymbol{\mu}$ gives

$$\frac{\partial Q^*}{\partial \boldsymbol{\mu}} = -2\mathbf{D}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}) + 2\mathbf{G}'_{\mu}\gamma .$$

Setting the above expression equal to $\mathbf{0}$, gives

$$(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{D}_x \mathbf{G'}_\mu \boldsymbol{\gamma}$$

i.e.

$$\boldsymbol{\mu} = \mathbf{x} - \mathbf{D}_x \mathbf{G'}_\mu \boldsymbol{\gamma} \; . \tag{3.9}$$

Using a linear Taylor expansion of $\boldsymbol{g}(\boldsymbol{\mu})$ about $\mathbf{x}$, we get

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{\mu}) &= \boldsymbol{g}(\mathbf{x} - \mathbf{D}_x \mathbf{G'}_\mu \boldsymbol{\gamma}) \\
&= \boldsymbol{g}(\mathbf{x}) + \left[\frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{g}(\boldsymbol{\mu}) \bigg|_{\boldsymbol{\mu}=\mathbf{x}} \right] (\mathbf{x} - \mathbf{D}_x \mathbf{G'}_\mu \boldsymbol{\gamma} - \mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \\
&= \boldsymbol{g}(\mathbf{x}) - (\mathbf{G}_x \mathbf{D}_x \mathbf{G'}_\mu) \boldsymbol{\gamma} + o(||\mathbf{x} - \boldsymbol{\mu}||) \; .
\end{aligned}$$

Setting the above equal to zero and solving for $\boldsymbol{\gamma}$, we get

$$\boldsymbol{\gamma} = (\mathbf{G}_x \mathbf{D}_x \mathbf{G'}_\mu)^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \; .$$

Substituting for $\boldsymbol{\gamma}$ in equation (3.9) yields

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu \mathbf{D}_x)' (\mathbf{G}_x \mathbf{D}_x \mathbf{G'}_\mu)^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||) \; . \tag{3.10}$$

This is the minimum modified chi-squared (MMCS) estimator. Neyman (1949) showed that the MMCS estimator belongs to the class of best asymptotic normal or (BAN) estimators, i.e. they are unbiased and asymptotically efficient. If the constraints are linear i.e. $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{G}\boldsymbol{\mu}$, then $\mathbf{G}_\mu = \frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{G}$ and equation (3.10) becomes

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}\mathbf{D}_x)' (\mathbf{G}\mathbf{D}_x \mathbf{G}')^{-1} \mathbf{G}\mathbf{x} \; , \tag{3.11}$$

which is found without iteration. This is the MMCS estimate discussed by Grizzle and Williams (1972) in the case of linear constraints. When the constraints are linear, the minimum of Q is

$$\min Q = (\mathbf{x} - \widehat{\boldsymbol{\mu}}_c)' \mathbf{D}_x^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_c)$$

$$= [(\mathbf{GD}_x)'(\mathbf{GD}_x\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{x})]'\mathbf{D}_x^{-1}[(\mathbf{GD}_x)'(\mathbf{GD}_x\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{x})]$$

$$= (\boldsymbol{g}(\mathbf{x}))'(\mathbf{GD}_x\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{x}) = W \,, \tag{3.12}$$

which is Wald's statistic under Poisson or multinomial sampling. This is another approach to the result of Bhapkar (1966), showing the algebraic equivalence of the MMCS and the Wald statistic for testing a linear hypothesis in categorical data.

3.3 LINEAR MODELLING FOR A CONTINGENCY TABLE

Consider a frequency table with observed count vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ and cell probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_p)'$. Suppose the model is formulated in terms of the constraints $\mathbf{A}\boldsymbol{\pi} = \mathbf{0}$, where $\mathbf{A}$ is a matrix of k linearly independent constraints or alternatively $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, where $\boldsymbol{\mu} = n\boldsymbol{\pi}$. Then in terms of Proposition 1, $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\mu}$, $\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A}$. The ML estimator for the vector of frequencies is

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{GV}_\mu)'(\mathbf{GV}_\mu\mathbf{G}')^{-1}\mathbf{Gx} \,.$$

The ML estimate is found iteratively by using equation (2.6). On convergence we obtain the ML estimates for the frequencies subject to the constraints $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$. In the case of Poisson sampling $\mathbf{V}_\mu = \mathbf{D}_\mu$ and the ML estimate is given by

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{AD}_\mu)'(\mathbf{AD}_\mu\mathbf{A}')^{-1}\mathbf{Ax} \,. \tag{3.13}$$

In the case of multinomial sampling

$$\mathbf{GV}_\mu = \mathbf{A}(\mathbf{D}_\mu - \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\mu}') = \mathbf{AD}_\mu - \mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}'/n = \mathbf{AD}_\mu \,,$$

since $\mathbf{A}\boldsymbol{\mu}$=$\mathbf{0}$, so that the ML estimate is also given by (3.13). Hence the covariance matrix $\mathbf{V}_\mu$, in Proposition 1 may be taken as $\mathbf{D}_\mu$ or $\mathbf{D}_\mu - \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\mu}'$.

A simple application of the estimation procedure is that of symmetry for a frequency table obtained by using one variable in the classification and assuming Poisson sampling.

Example 3.1

Consider the example taken from Haberman (1978) where subjects are asked the question "If you were asked to use one of the four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?" The following table was obtained.

TABLE 3.1: SELF-CLASSIFICATION BY SOCIAL CLASS

Response	Class	Number responding
Lower class	1	72
Working class	2	714
Middle class	3	655
Upper class	4	41

Examining the table we see that the number of respondents in the lower class is comparable with the number in the upper class, while the number in the working class and middle class are comparable.

Now let π_i denote the probability that a response belongs to class i, $i = 1, 2, 3, 4$. The table is symmetrical if $\pi_1 = \pi_4$ and $\pi_2 = \pi_3$, or if $\mu_1 = \mu_4$ and $\mu_2 = \mu_3$, where $\mu_i = n\pi_i$, and n is the total number of respondents. Let $\boldsymbol{\mu}' = (\mu_1, \mu_2, \mu_3, \mu_4)$, then the two restrictions can be written as

$$\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

i.e. $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$.

These constraints are of the form $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ where $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\mu}$ and

$$\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A} \ .$$

From the preceding discussion the ML estimate for the vector of frequencies is found by iteratively using

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{A}\mathbf{D}_\mu)'(\mathbf{A}\mathbf{D}_\mu\mathbf{A}')^{-1}\mathbf{A}\mathbf{x}$$

as described in the introductory paragraph of this section. The vector of estimated expected frequencies is

$$\widehat{\mathbf{m}}' = (56.5;\ 684.5;\ 684.5;\ 56.5) \ .$$

These estimates agree with the values from the solutions to the equations for finding the ML estimates, where

$$\widehat{m}_1 = \widehat{m}_4 = \frac{1}{2}(x_1 + x_4) = \frac{1}{2}(72 + 41) = 56.5$$

and

$$\widehat{m}_2 = \widehat{m}_3 = \frac{1}{2}(x_2 + x_3) = \frac{1}{2}(714 + 655) = 684.5 \ .$$

The statistic $\chi^2 = 11.05$ based on $df = 2$ with exact p-value 0.0040 indicating that the symmetry model does not provide a satisfactory fit.

The following example also illustrates the use of the estimation procedure for linear modelling of the frequencies in a frequency table.

Example 3.2

Grizzle, Starmer and Koch (1969) present the data recorded on 42 subjects, who were given drugs A, B and C. Some subjects had a favourable response to a single drug, some to two, and some to all three. The patterns of response and the number of subjects showing each pattern are shown in the Table 3.3.

If the three drugs are equally effective then $E(T_1) = E(T_2) = E(T_3)$. This hypothesis can be formulated in terms of the cell probabilities $\{\pi_i\}$ as

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = \pi_1 + \pi_2 + \pi_5 + \pi_6 = \pi_1 + \pi_3 + \pi_5 + \pi_7$$

TABLE 3.2: TABULATION OF RESPONSE TO DRUGS A, B AND C

(1 denotes favourable response, 0 denotes unfavourable response)

Pattern of response					
A	B	C	Number (x_i)	Expected probability	$\widehat{m}_i$
1	1	1	6	π_1	6.0000
1	1	0	16	π_2	10.840
1	0	1	2	π_3	2.6245
1	0	0	4	π_4	3.2312
0	1	1	2	π_5	2.6245
0	1	0	4	π_6	3.2313
0	0	1	6	π_7	11.4477
0	0	0	6	π_8	6.0000
Number favourable			46	1	
28	28	16			
T_1	T_2	T_3			

which simplifies to

$$\pi_3 + \pi_4 - \pi_5 - \pi_6 = 0,$$

and

$$\pi_2 + \pi_6 - \pi_3 - \pi_7 = 0.$$

Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_8)'$, then the hypothesis may be written as $\mathbf{A}\boldsymbol{\pi} = \mathbf{0}$, where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}.$$

Using the iterative procedure of (2.6) we obtain the ML estimates $\widehat{m}_i$ given in the table with $LR = 5.95$ and $\chi^2 = 5.71$ based on 2 degrees of freedom. The chi-squared value of 6.58 recorded by Grizzle et al. (1969) for this problem is based on the method of weighted least squares and the ML estimates for the frequencies under the hypothesis are not found. It is evident that the test of

Grizzle will reject H_o at the 5% level of significance, while the χ^2-test based on the ML estimates will not reject H_o, ($\chi^2_{2,0.95} = 5.99$).
Koch et al. (1977) formulate the hypothesis above as the hypothesis of first order marginal symmetry

$$H_M = \phi_1 = \phi_2 = \phi_3 \ ,$$

where

$$\begin{aligned} \phi_1 &= \pi_1 + \pi_2 + \pi_3 + \pi_4 \ , \\ \phi_2 &= \pi_1 + \pi_2 + \pi_5 + \pi_6 \ , \\ \phi_3 &= \pi_1 + \pi_3 + \pi_5 + \pi_7 \ . \end{aligned}$$

The matrix required to generate these constraints is

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} .$$

In order to construct a function $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, we let

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} .$$

Hence H_M can be written as $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{C}\mathbf{A}_1\boldsymbol{\pi} = \mathbf{0}$,
and

$$\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\pi}}\mathbf{g}(\boldsymbol{\pi}) = \mathbf{C}\mathbf{A}_1 \ ,$$

and the ML estimate for the expected frequencies is

$$\hat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}\mathbf{D}_\mu)'(\mathbf{G}\mathbf{D}_\mu\mathbf{G}')^{-1}\mathbf{G}\mathbf{x}$$

and the iterative procedure given in (2.6) is used to find the ML estimates. The ML estimates for the expected frequencies and the statistics are identical to those previously obtained. The IML program for this example can be found in the Appendix.

Another application is that of adjusting the frequencies in a contingency table to comply with marginal distributions obtained from other sources or to satisfy some theoretical marginal constraints. Little and Wu (1991) consider an $I \times J$ contingency table and compare the estimates of the adjusted cell probabilities for the methods of maximum likelihood under random sampling, least squares, minimum chi-squared and the method of raking. We show how the ML estimates for the adjusted cell probabilities may be obtained using the ML estimation procedure of Proposition 1, utilizing the marginal constraints.

Consider an $I \times J$ contingency table obtained from the cross-classification of the variables A and B. Let π_{ij} denote the probability that $A = i$ and $B = j$ in the target population, and let π_{i+} and π_{+j} denote the known marginal probabilities. Suppose that $p_{ij} = \frac{n_{ij}}{n}$ is the observed sample cell proportion. We now want to find the estimators, $\widehat{\pi}_{ij}$, of π_{ij} by adjusting the sample cell proportions p_{ij} to the known marginal probabilities π_{i+} and π_{+j}, so that

$$\sum_j \widehat{\pi}_{ij} = \pi_{i+}, \quad i = 1, \ldots, I; \tag{3.14}$$

$$\sum_i \widehat{\pi}_{ij} = \pi_{+j}, \quad j = 1, \ldots, J. \tag{3.15}$$

Deming and Stephan (1940) proposed an iterative proportional fitting (IPF) method to a contingency, which is called raking. In the framework of modelling subject to constraints, the adjusted estimates $\widehat{\pi}_{ij}$ can be obtained as follows. The constraints in (3.14) and (3.15) can be written in the form

$$(\mathbf{I}_r \otimes \mathbf{1}')\boldsymbol{\pi} = \mathbf{c}_1$$

and

$$(\mathbf{1} \otimes \mathbf{I}_c)\boldsymbol{\pi} = \mathbf{c}_2$$

which in turn can be written as

$$\mathbf{A}\boldsymbol{\pi} = \mathbf{c}\ .$$

For the problem of known margins

$$\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{A}\boldsymbol{\pi} - \mathbf{c} = \mathbf{0}\ .$$

Hence $\mathbf{G}_\pi = \dfrac{\partial}{\partial \boldsymbol{\pi}} \mathrm{g}(\boldsymbol{\pi}) = \mathbf{A}$. If we consider a multinomial sampling procedure, then

$$\mathrm{Cov}(\mathbf{p}) = \mathbf{V}_\pi = (\mathbf{D}_\pi - \boldsymbol{\pi}\boldsymbol{\pi}')/n \ .$$

The ML estimate for the cell probabilities subject to the marginal restrictions, is given by

$$\hat{\boldsymbol{\pi}}_c = \mathbf{p} - (\mathbf{A}\mathbf{V}_\pi)'(\mathbf{A}\mathbf{V}_\pi\mathbf{A}')^{-1}(\mathbf{A}\mathbf{p} - \mathbf{c}) \ .$$

As an illustration consider the following example.

Example 3.3

Little and Wu (1991) present data from the Second National Health and Nutrition Examination Survey (NHANES II).

TABLE 3.3: NHANES II DATA AND 1980 CENSUS DATA

NHANES II data ($n = 16547$)

		Urban	Rural	
Income	Low	0.3305	0.1955	0.5260
	High	0.3200	0.1540	0.474
		0.6506	0.3495	1.00

1980 Census ($n = 50\ 644\ 862$)

		Urban	Rural	
Income	Low	0.2064	0.1127	0.3191
	High	0.4969	0.1840	0.6809
		0.7033	0.2967	1.00

Table 3.3 shows 2×2 tables of income by urbanity from the survey and the

1980 Census. There is a distinct discrepancy between the NHANES II and Census margins of urbanity and income. It was thus considered necessary to adjust the NHANES II data such that $\pi_{11} + \pi_{12} = 0.3191$. The constraint may be written as

$$(1, \ 1, \ 0, \ 0)\boldsymbol{\pi} = 0.3191 \ .$$

Applying the estimation procedure gives the ML estimate

$$\widehat{\boldsymbol{\pi}}_c = (0.2005, \ 0.1186, \ 0.3200, \ 0.1540)' \ .$$

3.4 LOGLINEAR MODELLING FOR A CONTINGENCY TABLE

Suppose that the frequency of each cell in a contingency table is considered to be an independent observation of a Poisson random variable X_i with mean F_i for cell i. If the frequencies for the cells are conveniently arranged into a vector say $\mathbf{x} = (x_1, \ldots, x_p)'$, the saturated loglinear model can be written as

$$\ln(\mathbf{F}) = \mathbf{A}\boldsymbol{\lambda} \tag{3.16}$$

where $\mathbf{F} = (F_1, F_2, \ldots, F_p)'$ is the mean vector for the Poisson random variables, $\mathbf{A} : p \times p$ is the design matrix and $\boldsymbol{\lambda} : p \times 1$ is the vector of parameters for the mean, main effects and interaction effects. Fitting a lower order model implies fitting a model where certain elements of $\boldsymbol{\lambda}$ are zero. The saturated model of (3.16) can be written as

$$(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \ln(\mathbf{F}) = \boldsymbol{\lambda} \ .$$

The hypothesis that certain linear functions of $\boldsymbol{\lambda}$ are zero, can be written in the form

$$\mathbf{H}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \ln(\mathbf{F}) = \mathbf{H}\boldsymbol{\lambda} = \mathbf{0} \ ,$$

where $\mathbf{H}$ is a matrix specifying the linear functions of $\boldsymbol{\lambda}$ set to zero.
Let $\mathbf{A}_H' = \mathbf{H}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. We now have a function

$$\boldsymbol{g}(\mathbf{F}) = \mathbf{A}_H' \ln(\mathbf{F}) = \mathbf{0} \ .$$

The ML estimate of $\mathbf{F}$ subject to the constraints $\boldsymbol{g}(\mathbf{F}) = \mathbf{A}'_H \ln(\mathbf{F}) = \mathbf{0}$ is given by (2.8), where

$$\mathbf{G}_F = \frac{\partial}{\partial \mathbf{F}} \boldsymbol{g}(\mathbf{F}) = \mathbf{A}'_H \mathbf{D}_F^{-1} .$$

Thus

$$\begin{aligned} \widehat{\mathbf{F}}_c &= \mathbf{x} - (\mathbf{A}'_H \mathbf{D}_F^{-1} \mathbf{D}_F)'(\mathbf{A}'_H \mathbf{D}_x^{-1} \mathbf{D}_F \mathbf{D}_F^{-1} \mathbf{A}_H)^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \\ &= \mathbf{x} - \mathbf{A}_H (\mathbf{A}'_H \mathbf{D}_x^{-1} \mathbf{A}_H)^{-1} \mathbf{A}'_H \ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) . \qquad (3.17) \end{aligned}$$

$\widehat{\mathbf{F}}_c$ must be determined iteratively by iterating over $\mathbf{x}$. The iterative procedure is given by (2.5), which when convergence is attained gives the ML estimate $\widehat{\mathbf{F}}_c$. The ML estimates of the parameters in the loglinear model, are given by

$$\widehat{\boldsymbol{\lambda}}_c = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' \ln(\widehat{\mathbf{F}}_c) . \qquad (3.18)$$

The asymptotic covariance matrix of $\widehat{\mathbf{F}}_c$ is obtained from (2.7) and the estimated asymptotic covariance matrix of $\widehat{\mathbf{F}}_c$ is

$$\widehat{\boldsymbol{\Sigma}}_c = \mathbf{D}_{\widehat{\mathbf{F}}_c} - \mathbf{A}_H (\mathbf{A}'_H \mathbf{D}_{\widehat{\mathbf{F}}_c}^{-1} \mathbf{A}_H)^{-1} \mathbf{A}'_H . \qquad (3.19)$$

The covariance matrix for $\widehat{\boldsymbol{\lambda}}_c$ is

$$\text{Cov}(\widehat{\boldsymbol{\lambda}}_c) = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' \text{Cov}[\ln(\widehat{\mathbf{F}}_c)] \mathbf{A} (\mathbf{A}'\mathbf{A})^{-1},$$

where once again applying the "delta method", the estimated asymptotic covariance matrix

$$\begin{aligned} est.[\text{Cov}(\ln(\widehat{\mathbf{F}}_c))] &= \left(\frac{\partial}{\partial \widehat{\mathbf{F}}_c} \ln(\widehat{\mathbf{F}}_c) \right) \widehat{\boldsymbol{\Sigma}}_c \left(\frac{\partial}{\partial \widehat{\mathbf{F}}_c} \ln(\widehat{\mathbf{F}}_c) \right)' \\ &= \mathbf{D}_{\widehat{\mathbf{F}}_c}^{-1} \widehat{\boldsymbol{\Sigma}}_c \mathbf{D}_{\widehat{\mathbf{F}}_c}^{-1} . \end{aligned}$$

Hence the estimated covariance matrix for $\widehat{\boldsymbol{\lambda}}_c$ is

$$est.[\text{Cov}(\widehat{\boldsymbol{\lambda}}_c)] = (\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' [\mathbf{D}_{\widehat{\mathbf{F}}_c}^{-1} \widehat{\boldsymbol{\Sigma}}_c \mathbf{D}_{\widehat{\mathbf{F}}_c}^{-1}] \mathbf{A} (\mathbf{A}'\mathbf{A})^{-1} . \qquad (3.20)$$

To illustrate the above-mentioned estimation procedure, consider the following applications in loglinear modelling using a Poisson sampling scheme.

(i) **Independence**

Consider a 3×3 contingency table with row variable B and column variable C. The independence model is written as $\ln(F_{ij}) = \mu + \lambda_i^B + \lambda_j^C$, which implies testing

$$H_o : \lambda_{ij}^{BC} = 0 \ , \quad i, j = 1, 2;$$

in the saturated model.

Let $\boldsymbol{\lambda}' = [\mu, \lambda_1^B, \lambda_2^B, \lambda_1^C, \lambda_2^C, \lambda_{11}^{BC}, \lambda_{12}^{BC}, \lambda_{21}^{BC}, \lambda_{22}^{BC}]$.
The design matrix is given by

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \mu & \lambda_1^B & \lambda_2^B & \lambda_1^C & \lambda_2^C & \lambda_{1\,1}^{BC} & \lambda_{1\,2}^{BC} & \lambda_{2\,1}^{BC} & \lambda_{2\,2}^{BC} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array} .$$

Now let

$$\mathbf{H} = [\mathbf{0}_{4\times5} \ , \ \mathbf{I}_4].$$

Then $\mathbf{H}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\ln(\mathbf{F}) = \mathbf{H}\boldsymbol{\lambda} = \mathbf{0}$, gives

$$\begin{bmatrix} \lambda_{11}^{BC} \\ \lambda_{12}^{BC} \\ \lambda_{21}^{BC} \\ \lambda_{22}^{BC} \end{bmatrix} = \mathbf{0},$$

which is the statement under H_o. $\widehat{\mathbf{F}}_c$ and $\widehat{\boldsymbol{\lambda}}_c$ can now be found by using (3.17) and (3.18) with $\mathbf{H}$ above.

Another approach to the independence model is to consider the odds ratios or cross product ratios for the table. Let π_{ij} denote the probability that an observation is classified in cell (i, j). The set of $(I-1)(J-1)$ local odds ratios for an $I \times J$ contingency table is given by

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \;,\; i = 1, \dots, \; I-1; \;\; j = 1, \dots, \; J-1.$$

The independence model in terms of local odds ratios is

$$\theta_{ij} = 1, \;\; \text{for} \;\; i = 1, \dots, I-1, \;\; j = 1, \dots, J-1.$$

Consider the 3×3 table as an illustration. The model of independence in terms of odds ratios is

$$\theta_{11} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1, \quad \theta_{12} = \frac{\pi_{12}\pi_{23}}{\pi_{13}\pi_{22}} = 1,$$

$$\theta_{21} = \frac{\pi_{21}\pi_{32}}{\pi_{22}\pi_{31}} = 1, \quad \theta_{22} = \frac{\pi_{22}\pi_{33}}{\pi_{23}\pi_{32}} = 1.$$

Let F_{ij} denote the expected frequency for cell (i, j). Then the constraints above may be written as

$$\begin{aligned} F_{11}F_{22} - F_{12}F_{21} &= 0, \\ F_{12}F_{23} - F_{13}F_{22} &= 0, \\ F_{21}F_{32} - F_{22}F_{31} &= 0, \\ F_{22}F_{33} - F_{23}F_{32} &= 0. \end{aligned}$$

These constraints are of the form $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$ and the ML estimation procedure of Proposition 1 may be applied.

The matrix $\mathbf{G}_F = \dfrac{\partial}{\partial \mathbf{F}} \boldsymbol{g}(\mathbf{F})$, where

$$\mathbf{G}_F = \begin{matrix} (F_{11} & F_{12} & F_{13} & F_{21} & F_{22} & F_{23} & F_{31} & F_{32} & F_{33}) \end{matrix}$$
$$\begin{bmatrix} F_{22} & -F_{21} & 0 & -F_{12} & F_{11} & 0 & 0 & 0 & 0 \\ 0 & F_{23} & -F_{22} & 0 & -F_{13} & F_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & F_{32} & -F_{31} & 0 & -F_{22} & F_{21} & 0 \\ 0 & 0 & 0 & 0 & F_{33} & -F_{32} & 0 & -F_{23} & F_{22} \end{bmatrix}.$$

If a Poisson sampling procedure is used, then $\mathbf{V} = \mathbf{D}_F$ and the estimation procedure is

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{D}_F)'(\mathbf{G}_x\mathbf{D}_F\mathbf{G}_F)^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||),$$

which will require a double iteration, namely over $\mathbf{x}$ and $\mathbf{F}$ to give the ML estimates of the expected frequencies for the independence model.

Consider the following example, in which the independence model is fitted to a 3×3 contingency table.

Example 3.4
Consider the data of Hedlund (1978), discussed by Agresti (1984). The table presents the relationship between the variables "political ideology" and "political party affiliation". For a sample of voters taken in the 1976 presidential primary of Wisconsin.

TABLE 3.4: POLITICAL IDEOLOGY AND PARTY AFFILIATION

	Political Ideology			
Party Affiliation	Liberal	Moderate	Conservative	Total
Democrat	143(102.05)	156(161.37)	100(135.58)	399
Independent	119(120.21)	210(190.08)	141(159.70)	470
Republican	15 (54.73)	72 (86.55)	127 (72.72)	214

The figures in parentheses are the expected frequencies for the independence model

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^I \ .$$

For the independence model $LR = 105.66$ and $\chi^2 = 102.05$ based on $df = 4$. The fit for the independence model is poor other possible models should be considered to describe the data. Note that $df = 4$ is the number of columns in the matrix $\mathbf{A}_H$. The estimated expected frequencies and parameters are calculated by using the program in the Appendix. This program can be used for higher dimensional tables and the testing of conditional independence, as will be illustrated in Example 3.5.

(ii) **Symmetry**

For symmetry in the 3×3 contingency table, $F_{ij} = F_{ji}$, and in terms of the parameters of the loglinear model, this implies testing

$$H_o : \lambda_{ij}^{BC} = \lambda_{ji}^{BC} \text{ and } \lambda_i^B = \lambda_i^C \quad , \quad i,j = 1,2 \; .$$

Consider

$$\mathbf{H} = \begin{array}{c} \begin{array}{ccccccccc} \mu & \lambda_1^B & \lambda_2^B & \lambda_1^C & \lambda_2^C & \lambda_{11}^{BC} & \lambda_{12}^{BC} & \lambda_{21}^{BC} & \lambda_{22}^{BC} \end{array} \\ \left[\begin{array}{ccccccccc} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{array} \right] \end{array} .$$

Thus H_o can once again be written as

$$\mathbf{H}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \ln(\mathbf{F}) = \mathbf{H}\boldsymbol{\lambda} = \mathbf{0} \; .$$

For quasi-symmetry the hypothesis is

$$H_o : \lambda_{ij}^{BC} = \lambda_{ji}^{BC} \; , \quad i,j = 1,2$$

and hence let $\mathbf{H} = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; -1 \;\; 0]$.
The ML estimate for the expected frequency $\widehat{\mathbf{F}}_c$, is found by iteratively using (3.17) and $\widehat{\boldsymbol{\lambda}}_c$ is given by (3.18).

The symmetry model can also be easily modelled directly in terms of the restrictions $F_{ij} = F_{ji}$. Once again consider a 3×3 contingency table with observed frequencies $\{x_{ij}\}$ and expected frequencies $\{F_{ij}\}$ $i,j = 1,2,3$. The hypothesis of symmetry for the table implies $F_{ij} = F_{ji}$ $\forall i \neq j$, which gives the following constraints

$$F_{12} - F_{21} = 0 \; ,$$
$$F_{13} - F_{31} = 0 \; ,$$
$$F_{23} - F_{32} = 0 \; .$$

These constraints can be written as $\mathbf{AF} = \mathbf{0}$, where

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccccccccc} F_{11} & F_{12} & F_{13} & F_{21} & F_{22} & F_{23} & F_{31} & F_{32} & F_{33} \end{array} \\ \left[\begin{array}{ccccccccc} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{array}\right] \end{array} .$$

Hence $\boldsymbol{g}(\mathbf{F}) = \mathbf{AF} = \mathbf{0}$ and $\mathbf{G} = \dfrac{\partial}{\partial \mathbf{F}}\mathbf{AF} = \mathbf{A}$.

The ML estimation procedure is given by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{AD}_F)'(\mathbf{AD}_F\mathbf{A}')^{-1}\mathbf{Ax} .$$

Since $\mathbf{D}_F$ is unknown and the constraints are linear iteration takes place over $\mathbf{F}$.

The next example illustrates the use of the symmetry model for a 3×3 contingency table.

Example 3.5

Table 3.5 presents the information from a survey investigating the view on the economic and political situation in S.A. at a particular point in time. For the symmetry model we test

$$H_o : \lambda_{ij}^{EP} = \lambda_{ji}^{EP} \text{ and } \lambda_i^E = \lambda_i^P, \; i,j = 1,2.$$

Applying the estimation procedure to the data we get the following parameter estimates and standard normal values.

TABLE 3.5: VIEW ON POLITICAL AND ECONOMIC SITUATION

	Political			
Economic	satisfied	neither	dissatisfied	Total
satisfied	198 (198)	65 (64)	59 (74)	322
neither	63 (64)	79 (79)	66 (71)	208
dissatisfied	89 (74)	76 (71)	272 (272)	437
Total	350	220	397	967

(Figures in parentheses are the estimated expected frequencies for the symmetry model).

parameter	estimate	z-value
μ	4.5239	123.98
λ_1^E	0.0599	1.91
λ_2^E	-0.2602	-7.36
λ_1^P	0.0599	1.91
λ_2^P	-0.2602	-7.36
λ_{11}^{EP}	0.6447	9.69
λ_{12}^{EP}	-0.1647	-2.78
λ_{21}^{EP}	-0.1647	-2.78
λ_{22}^{EP}	0.3660	4.77

For this model $LR = 6.86$ and $\chi^2 = 6.82$ based on $df = 3$. The fit is not adequate so we consider the quasi-symmetry model. For the quasi-symmetry model we test

$$H_o : \lambda_{ij}^{EP} = \lambda_{ji}^{EP},\ i, j = 1, 2.$$

The parameter estimates and standard normal values appear in the following table.

parameter	estimate	z-value
μ	4.5194	123.22
λ_1^E	-0.0052	-0.10
λ_2^E	-0.2854	-5.24
λ_1^P	0.1252	2.50
λ_2^P	-0.2312	-4.32
λ_{11}^{EP}	0.6489	9.72
λ_{12}^{EP}	-0.1629	-2.75
λ_{21}^{EP}	-0.1629	-2.75
λ_{22}^{EP}	0.3667	4.78

For the quasi-symmetry model $LR = 1.04$ and $\chi^2 = 1.04$ based on $df = 1$. The fit for this model is much better. The quasi-symmetry model will be revisited in Chapter 4, where models for square tables are discussed.

The estimation procedure can easily be extended to higher order cross-classifications and the hypotheses of independence and conditional independence are tested by fitting the appropriate models in terms of the implied constraints $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$, as will be illustrated in the following example.

Example 3.6
Consider the data from Haberman (1978) where the following table was compiled from the 1975 General Social Survey. The table gives the gender of the respondent and level of education in determining attitudes toward roles for women. The question asked is whether "women should take care of running their homes and leave running of the country up to men".

Let S denote the sex of the respondent
E denote the education level of the respondent
R denote the response of the respondent.

TABLE 3.6: SUBJECTS CROSS-CLASSIFIED BY ATTITUDE TOWARDS WOMEN STAYING AT HOME, SEX OF RESPONDENT AND EDUCATION OF RESPONDENT

Sex of respondent	Education of respondent (yrs)	Response: Agree	Response: Disagree	Total
Male	$\leq$ 8	72	47	119
	9 – 12	110	196	306
	$\geq$ 13	44	179	223
		226	422	648
Female	$\leq$ 8	86	38	124
	9 – 12	173	283	456
	$\geq$ 13	28	187	215
		287	508	795
		513	930	$N = 1443$

The saturated loglinear model can be written as $\ln(\mathbf{F}) = \mathbf{A}\boldsymbol{\lambda}$ where

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & -1 & -1 & -1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & -1 & -1 & 0 & -1 & 0 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 0 & -1 & 0 & 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 & -1 & 1 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 & 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 & -1 & 1 & 0 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \end{bmatrix}$$

and

$$\boldsymbol{\lambda}' = (\mu, \lambda_1^S, \lambda_1^E, \lambda_2^E, \lambda_{11}^{SE}, \lambda_{12}^{SE}, \lambda_1^R, \lambda_{11}^{SR}, \lambda_{11}^{ER}, \lambda_{21}^{ER}, \lambda_{111}^{SER}, \lambda_{121}^{SER}) \, .$$

Fitting the model of no three-factor effect implies that $\lambda_{111}^{SER} = \lambda_{121}^{SER} = 0$, and the matrix

$$\mathbf{H} = [\mathbf{0}_{2\times 10} \ , \ \mathbf{I}_2].$$

Now consider the model of conditional independence. Using the notation of Agresti (1984), define the set of local odds ratios as

$$\theta_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}, \pi_{i+1,j,k}} \ , \quad 1 \leq i \leq r-1, \ \ 1 \leq j \leq c-1,$$

for variable X, Y and Z.

The variables X and Y are conditionally independent at level k of Z if all $(r-1)(c-1)$ of the $\theta_{ij(k)}$ at that fixed level of Z are equal to 1. The variables are said to be conditionally independent given Z (i.e. at every level of Z) if all $l(r-1)(c-1)$ of the $\{\theta_{ij(k)}\}$ are equal to 1, or alternatively if all of the $\{\ln \theta_{ij(k)}\}$ are equal to 0. Hence modelling the log-odds will give a function $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}$, in terms of Proposition 1.

In this example we wish to test whether S and R are conditionally independent given E. For this to be true

$$\begin{aligned}
\theta_{1(1)1} &= \frac{F_{111}F_{212}}{F_{112}F_{211}} = 1 \ , \\
\theta_{1(2)1} &= \frac{F_{121}F_{222}}{F_{122}F_{221}} = 1 \ , \\
\theta_{1(3)1} &= \frac{F_{131}F_{232}}{F_{132}F_{231}} = 1 \ ,
\end{aligned}$$

or alternatively

$$\begin{aligned}
\ln \theta_{1(1)1} &= \ln F_{111} + \ln F_{212} - \ln F_{112} - \ln F_{211} = 0 \ , \\
\ln \theta_{1(2)1} &= \ln F_{121} + \ln F_{222} - \ln F_{122} - \ln F_{221} = 0 \ , \\
\ln \theta_{1(3)1} &= \ln F_{131} + \ln F_{232} - \ln F_{132} - \ln F_{231} = 0 \ .
\end{aligned}$$

This will be true if and only if $\lambda_{11}^{SR} = 0$ and $\lambda_{111}^{SER} = \lambda_{121}^{SER} = 0$. Now let

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

then

$$\mathbf{H}\boldsymbol{\lambda} = \begin{pmatrix} \lambda_{11}^{SR} \\ \lambda_{111}^{SER} \\ \lambda_{121}^{SER} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The estimated expected frequencies can now be found by applying the estimation procedure of (3.17) with **H** as above. Using this estimation procedure to fit the conditional independence model gives the following ML estimates of the parameters and standard normal values.

parameter	estimate	z-value
μ	4.5504	135.57
λ_1^S	-0.0673	-2.29
λ_1^E	-0.4912	-9.57
λ_2^E	0.6452	16.07
λ_{11}^{SE}	0.0467	0.99
λ_{12}^{SE}	-0.1322	-3.64
λ_1^R	-0.2554	-7.63
λ_{11}^{ER}	0.5654	11.03
λ_{21}^{ER}	-0.0077	-0.19

The goodness of fit statistics for the conditional independence model are $LR = 6.02$ and $\chi^2 = 5.99$ based on $df = 3$ with attained significance level 0.11, showing that the fit is reasonable.

The ML estimates for the expected frequencies in the conditional independence model can also be found by using the set of odds ratios as was discussed earlier.

The odds ratios $\theta_{1(j)1} = 1$ for $j = 1, 2, 3$ imply

$$\begin{aligned} F_{111}F_{212} - F_{112}F_{211} &= 0\,, \\ F_{121}F_{222} - F_{122}F_{221} &= 0\,, \\ F_{131}F_{232} - F_{132}F_{231} &= 0\,, \end{aligned}$$

which is of the form $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$. The ML estimates can be found by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{D}_F)'(\mathbf{G}_x\mathbf{D}_F\mathbf{G}_F)^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||),$$

which will require a double iteration, namely over $\mathbf{x}$ and $\mathbf{F}$.

3.5 LOGLINEAR MODELS FOR ORDINAL VARIABLES

In the standard loglinear analysis of categorical data all variables are treated as nominal. Many cross-classification tables include variables such as age, income, opinion to name but some variables, which are ordinal. Loglinear models which utilize the quantitative nature of ordinal variables are recommended for analysis when one or more variables in the cross-classification table is ordinal. Among the advantages for using ordinal methods instead of standard loglinear models are that

(a) ordinal methods have greater power for detecting important alternatives to null hypotheses such as the one of independence.

(b) ordinal methods can use a greater variety of models, most of which have fewer parameters than the standard models for nominal variables.

Agresti (1984) presents a number of models which can be used for ordinal data. Some of these models will be reviewed in the following paragraph and the estimation procedure for these models will be formulated.

3.5.1 Loglinear Models for Two Dimensional Tables

First consider the $r \times c$ table with expected frequencies $\{F_{ij}\}$. The standard loglinear model is

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad ,i = 1, \cdots, r \; ; \; j = 1, \cdots, c \; . \qquad (3.21)$$

$$\text{where} \qquad \sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0.$$

(i) **Uniform Association Model**

Suppose that the row variable X and column variable Y are both ordinal. Assign scores $\{u_i\}$ to the levels of the variable X and scores $\{v_j\}$ to the levels of the variable Y. The most commonly used assignments are the integer scores $\{u_i = i\}$ and $\{v_j = j\}$. A loglinear model which uses the ordering of the rows and columns through the scores is

$$\ln(F_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \beta(u_i - \overline{u})(v_j - \overline{v}) \; , i = 1, \cdots, r \; ; \; j = 1, \cdots, c \; ; \qquad (3.22)$$

and $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$.

The degrees of freedom for testing the goodness of fit, are $df = (r-1)(c-1) - 1 = rc - r - c$. The parameter β describes the association between X and Y. The independence model is obtained if $\beta = 0$. If $\beta < 0$, the interaction term will be positive for cells for which $(u_i - \bar{u}) < 0$ and $(v_j - \bar{v}) > 0$ or $(u_i - \bar{u}) > 0$ and $(v_j - \bar{v}) < 0$, i.e. small X and large Y values or large X and small Y values. If $\beta > 0$, the interaction term will be positive for cells for which $(u_i - \bar{u}) > 0$ and $(v_j - \bar{v}) > 0$ or $(u_i - \bar{u}) < 0$ and $(v_j - \bar{v}) < 0$. For any arbitrary pair of rows $k < \ell$ and columns $m < n$,

$$\ln\left(\frac{F_{km}F_{\ln}}{F_{\ell m}F_{kn}}\right) = \beta(u_\ell - u_k)(v_n - v_m) \; . \qquad (3.23)$$

If $(u_\ell - u_k) = (v_n - v_m) = 1$, then the log-odds ratio is equal to β.

(ii) **The Loglinear Row Effects Model**

Suppose that X is nominal and Y is ordinal with scores $\{v_j\}$. A loglinear model that uses the ordinality of Y is

$$\ln(F_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \tau_i(v_j - \overline{v}) \ , i = 1, \cdots, r \ ; \ j = 1, \cdots, c \ ; \quad (3.24)$$

where $\sum \lambda_i^X = \sum \lambda_j^Y = \sum \tau_i = 0$.

The $\{\tau_i\}$ are parameters, $(r-1)$ of which are linearly independent. For the goodness of fit test, $df = rc - [1 + (r-1) + (c-1) + (r-1)] = (r-1)(c-2)$. The term $\tau_i(v_j - \overline{v})$ describes the interaction term. The $\{\tau_i\}$ are referred to as the row effects. If $\tau_i > 0$ then in row i the interaction term is positive for cells for which $v_j - \bar{v} > 0$. If $\tau_i < 0$ then in row i the interaction term is positive for cells for which $v_j - \bar{v} < 0$.

These models can be extended to higher dimensions. Agresti (1984) discusses a variety of extensions to higher dimensional tables.

The estimation procedure of Proposition 1 can also be used to fit these ordinal models. The ML estimators of the parameters, as well as the expected frequencies can be obtained. An ordinal model can be written in the form $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ has main effects and some covariates as column vectors, instead of the standard interaction terms. Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\mathbf{P}\ln(\mathbf{F}) = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. Thus $\boldsymbol{g}(\mathbf{F}) = \mathbf{P}\ln(\mathbf{F})$, gives the constraints such that $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$ and $\mathbf{G} = \mathbf{P}\mathbf{D}_F^{-1}$. The ML estimate is then

$$\begin{aligned}\widehat{\mathbf{F}}_c &= \mathbf{x} - (\mathbf{P}\mathbf{D}_F^{-1}\mathbf{D}_F)'(\mathbf{P}\mathbf{D}_x^{-1}\mathbf{D}_F\mathbf{D}_F^{-1}\mathbf{P}')^{-1}\mathbf{P}\ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \\ &= \mathbf{x} - \mathbf{P}(\mathbf{P}\mathbf{D}_x^{-1}\mathbf{P})^{-1}\mathbf{P}\ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \ ,\end{aligned}$$

and iteration takes place over $\mathbf{x}$. The iterative procedure is given by (2.5).

The following example illustrates how the estimation procedure is used to fit a row effects model to a contingency table.

Example 3.7

Consider the data presented in Example 3.4. The variable "Political Ideology" is ordinal while the variable "Party Affiliation" is nominal. Assign scores $\{v_j - \bar{v}\} = \{-1, 0, 1\}$ for the categories "Conservative", "Moderate" and "Liberal", respectively. The row effects model for the data is given by

$$\ln(F_{ij}) = \mu + \lambda_i^{PA} + \lambda_j^{PI} + \tau_i(v_j - \bar{v}) \ , \quad i, j = 1, 2, 3;$$

where $\sum_i \lambda_i^{PA} = \sum_j \lambda_j^{PI} = \sum_i \tau_i = 0.$

The τ_i are the row effects for "Party Affiliation". Write the model as $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}' = (\mu \ , \ \lambda_1^{PA} \ , \ \lambda_2^{PA} \ , \ \lambda_1^{PI} \ , \ \lambda_2^{PI} \ , \ \tau_1 \ , \ \tau_2)$ and

$$\mathbf{X} \ = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & -1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & 1 & 0 & 1 & 1 \\ 1 & -1 & -1 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} .$$

Using the estimation procedure with $\boldsymbol{g}(\mathbf{F}) = \mathbf{P}\ln(\mathbf{F})$ as discussed above, the ML estimators $\widehat{\mathbf{F}}_c$ and $\widehat{\boldsymbol{\beta}}$ can be found iteratively. The expected frequencies are given in the Table 3.7. The parameter estimates and standard normal values appear in the following table.

TABLE 3.7: EXPECTED FREQUENCIES FOR ROW EFFECTS MODEL

	Political Ideology		
Party Affiliation	Conservative	Moderate	Liberal
Democrat	136.63	168.73	93.63
Independent	123.79	200.41	145.79
Republican	16.57	68.86	128.57

parameter	value	z-value
μ	4.6203	120.68
λ_1^{PA}	0.2414	5.09
λ_2^{PA}	0.4135	9.00
λ_1^{PI}	-0.4391	-7.95
λ_2^{PI}	0.2666	6.22
τ_1	-0.4947	-7.98
τ_2	-0.2240	-3.81

The parameter $\tau_3 = -(\tau_1 + \tau_2) = 0.7187$. For this model $LR = 2.81$ and $\chi^2 = 2.80$ based on $df = 2$. The model thus provides an adequate fit for the data. Agresti (1984) gives a discussion on the interpretation of the parameters. The program for the above-mentioned example appears in the Appendix, with a brief explanation on the use of the program.

3.5.2 Ordinal Loglinear Models For Higher Dimensional Tables

Consider another example of fitting an ordinal loglinear model to a higher dimensional table.

Example 3.8

In a survey on socio-political change in South Africa of October 1987, people were asked their opinion as to how effectively the unrest situation had been handled. The results are summarized in Table 3.8.

TABLE 3.8: OBSERVED FREQUENCIES AND EXPECTED FREQUENCIES FOR CUMULATIVE LOGIT MODEL FOR OPINION ON THE UNREST SITUATION

		Opinion			
Language	Age	Ineffective	Neither effective nor ineffective	Fairly effective	Very effective
Afrikaans	18-24	5	6	53	16
	25-34	7	15	115	25
	35-64	13	20	197	96
English and other	18-24	7	3	33	7
	25-34	8	9	65	23
	35-64	13	13	149	56

Here the variable language is nominal, while age and opinion are ordinal. Let L denote language, A denote age and O denote opinion. Consider the following model

$$\ln(F_{ijk}) = \mu + \lambda_i^L + \lambda_j^A + \lambda_k^O + \tau_i^{LO}(w_k - \bar{w}) + \beta^{AO}(v_j - \bar{v})(w_k - \bar{w}) \ , \quad (3.25)$$

where $i = 1,2$; $j = 1,2,3$; $k = 1,2,3,4$; and $\sum_i \lambda_i^L = \sum_j \lambda_j^A = \sum_k \lambda_k^O Z = \sum_i \tau_i^{LO} = 0$.

The allocation of scores $\{v_j\}$ and $\{w_k\}$ for age and opinion are as follows: $\{v_j\} = \{1,2,3\}$ and $\{w_k\} = \{1,2,3,4\}$ so that $\{v_j - \bar{v}\} = \{-1;0;1\}$ and $\{w_k - \bar{w}\} = \{-1.5;-0.5;0.5;1.5\}$.

The term $\tau_i^{LO}(w_k - \bar{w})$ describes the language $\times$ opinion interaction, while

$\beta^{AO}(v_j - \bar{v})(w_k - \bar{w})$ describes the age $\times$ opinion interaction. We refer to $i(w_k - \bar{w})$, ($i = 1$ for Afrikaans and $i = -1$ for English) and $(v_j - \bar{v})(w_k - \bar{w})$ in the respective terms, as the covariates for the above-mentioned interactions. The interaction $\tau_i^{LA}(v_j - \bar{v})$ for language $\times$ age, is not significant and has been omitted from the model.

The loglinear model can be also be written $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{F}$ is the vector of frequencies, $\mathbf{X}$ is the corresponding design matrix and $\boldsymbol{\beta}$ is the vector of parameters. The covariates mentioned above are entered as columns in the design matrix and the parameters are then estimated. The column vectors for the covariates, to be included in the design matrix $\mathbf{X}$ are as follows

level (i, j, k)	$(v_j - \bar{v})(w_k - \bar{w})$	$i(w_k - \bar{w})$	$(v_j - \bar{v})$
111	1.5	-1.5	-1
112	0.5	-0.5	-1
113	-0.5	0.5	-1
114	-1.5	1.5	-1
121	0	-1.5	0
122	0	-0.5	0
123	0	0.5	0
124	0	1.5	0
131	-1.5	-1.5	1
132	-0.5	-0.5	1
133	0.5	0.5	1
134	1.5	1.5	1

level (i,j,k)	$(v_j - \bar{v})\,(w_k - \bar{w})$	$i\,(w_k - \bar{w})$	$(v_j - \bar{v})$
211	1.5	1.5	-1
212	0.5	0.5	-1
213	-0.5	-0.5	-1
214	-1.5	-1.5	-1
221	0	1.5	0
222	0	0.5	0
223	0	-0.5	0
224	0	-1.5	0
231	-1.5	1.5	1
232	-0.5	0.5	1
232	0.5	-0.5	1
234	1.5	-1.5	1

The covariate $(v_j - \bar{v})$ will be used later for an ordinal main effect for "age", namely $\beta^A\,(v_j - \bar{v})$. The ML estimates for the parameters for the model in (3.25) are given in the following table.

parameter	estimate	z-value
μ	3.049158	57.046211
λ_1^L	0.157676	3.827043
λ_1^A	-0.625857	-9.103568
λ_2^A	0.002297	0.043882
λ_1^O	-0.906354	-8.090560
λ_2^O	-0.748264	-7.332036
λ_3^O	1.389461	24.022251
β^{AO}	0.202855	3.337615
τ^{LO}	0.064780	1.424733

For this model $LR = 11.94$ and $\chi^2 = 11.50$ based on $df = 15$. The Akaike Information Criteria, $AIC^* = LR + 2p = 11.94 + 2(9) = 29.94$. The model fits the data well. The "language opinion" interaction is not significant and can be omitted from the model. It is interesting to note that the age main effect exhibits a linear trend, namely -0.625857, 0.002297 and 0.623560. Thus the age main effect $\lambda_j^Y, j = 1, 2, 3$; can be replaced by the term $\beta^A (v_j - \bar{v})$. We thus explain the age main effect by using one parameter, β^A, instead of λ_1^A and λ_2^A.

We now fit the following model to the data

$$\ln(F_{ijk}) = \mu + \lambda_i^L + \beta^A(v_j - \bar{v}) + \lambda_k^O + \beta^{AO}(v_j - \bar{v})(w_k - \bar{w})$$

$$i = 1, 2\ ; \quad j = 1, 2, 3; \quad k = 1, 2, 3, 4\ ; \text{and} \ \sum_i \lambda_i^L = \sum_k \lambda_k^O = 0\ .$$

Applying the estimation procedure, we get the ML estimates for the parameters. These appear in the following table.

parameter	estimate	z-value
μ	3.045718	56.992641
λ_1^L	0.193142	5.855991
λ_1^O	-0.919581	-8.224010
λ_2^O	-0.755430	-7.414750
λ_3^O	1.392513	24.081166
β^{AO}	0.202623	3.348394
β^A	0.623992	11.559688

The model fits the data well. The likelihood ratio statistic, $LR = 13.97$ and $\chi^2 = 13.74$ based on $df = 17$. The $AIC^* = 25.97$, is smaller than that of the previous model, hence we would prefer the latter model. The parameter $\widehat{\beta}^{AO} = 0.202623$ is significant and describes the age $\times$ opinion interaction. The log-odds ratio for any pair of adjacent rows and adjacent columns is $\widehat{\beta}^{AO} =$

0.202623. The interpretation of $\widehat{\beta}^{AO}$ is that for a particular row j, of the age variable Y, the interaction term is a linear function of opinion, O through the scores $\{w_k\}$ with slope $\widehat{\beta}^{AO}\,(w_k - \bar{w}) = 0.202623\,(w_k - \bar{w})$.

For the first row of the age variable (18 - 24 years), the interaction term is given by

$$0.2026\,(v_j - \bar{v})\,(w_k - \bar{w}) = -0.2026\,(w_k - \bar{w}) \quad \text{since} \quad (v_j - \bar{v}) = -1 \ .$$

Across the opinion categories, for which $\{w_k - \bar{w}\} = \{-1.5; -0.5; 0.5; 1.5\}$, the interaction term has the following contribution to $\ln(F_{ijk})$

$$\begin{array}{lll} -0.2026(-1.5) & = 0.3039 & \text{for } w_k - \bar{w} = -1.5 \ , \\ -0.2026(-0.5) & = 0.1013 & \text{for } w_k - \bar{w} = -0.5 \ , \\ -0.2026(0.5) & = -0.1013 & \text{for } w_k - \bar{w} = 0.5 \ , \\ -0.2026(1.5) & = -0.3039 & \text{for } w_k - \bar{w} = 1.5 \end{array} .$$

Thus for the age category 18 - 24 years having the opinion "ineffective", the expected frequency is exp(0.3039) = 1.35 times the effect due to the mean and main effects for the cell.

[Note that $F_{ijk} = \exp\left\{\mu + \lambda_i^L + \beta^A(v_j - \bar{v}) + \lambda_k^O\right\} \exp\left\{\beta^{AO}\,(v_j - \bar{v})\,(w_k - \bar{w})\right\}$].

Similarly for the age category 35 - 64 years, the interaction term is

$$0.2026\,(w_k - \bar{w}) \qquad \text{since} \qquad v_j - \bar{v} = 1$$

Thus for age category 35 - 64 years with opinion "very effective" the expected frequency is 35% higher than the effect due to the mean and main effects for the cell. We can thus conclude that the younger age group is of the opinion that the unrest tends to be "ineffectively" handled, while the older group is more likely to think that the unrest is being handled "very effectively".
This example will be discussed again in Chapter 5 using logit models .

3.5.3 Orthogonal Polynomials

Orthogonal polynomial scores can also be used as a method for describing data which are ordinal. Consider the following example.

Example 3.9

Using the data set in Example 3.8 and considering only the variables, age and opinion, we have the following cross-classification.

TABLE 3.9: FREQUENCY TABLE FOR UNREST DATA

Age	Opinion			
	Ineffective	Neither	Fairly effective	Very effective
18-24	12 (12.38)	9 (11.91)	86 (83.30)	23 (22.41)
25-34	15 (17.12)	24 (20.17)	180 (172.78)	48 (56.93)
35-64	26 (23.50)	33 (33.92)	346 (355.90)	152 (143.66)

The linear by linear association model is given by

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^O + \beta(u_i - \bar{u})(v_j - \bar{v}), \quad i = 1,2,3; \; j = 1,2,3,4.$$

The $\{u_i - \bar{u}\}$ values assigned to the age categories are $\{-1, 0, 1\}$ while the $\{v_j - \bar{v}\}$ values for the opinion categories are $\{-1.5, -0.5, 0.5, 1.5\}$. Fitting this model to the data gives the following ML estimates for the parameters.

parameter	estimate	z-value
μ	3.7572	70.62
λ_1^A	-0.6259	-9.10
λ_2^A	0.0023	0.04
λ_1^O	-0.9195	-8.22
λ_2^O	-0.7554	-7.41
λ_3^O	1.3925	24.08
β	0.2029	3.34

The estimated frequencies appear in parentheses in the cross-classification table. The likelihood ratio statistic $LR = 4.67$ and $\chi^2 = 4.57$ based on $df = 2$.

If we now consider a model using orthogonal polynomial coefficients, we have for $l = 3$ and 4,

$l = 3$	
ψ_1	ψ_2
-1	1
0	-2
1	1

$l = 4$		
ψ_1	ψ_2	ψ_3
-3	1	-1
-1	-1	3
1	-1	-3
3	1	1

Hence replacing main effects by orthogonal polynomial scores we have the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & 1 & -3 & 1 & -1 & 3 \\ 1 & -1 & 1 & -1 & -1 & 3 & 1 \\ 1 & -1 & 1 & 1 & -1 & -3 & -1 \\ 1 & -1 & 1 & 3 & 1 & 1 & -3 \\ 1 & 0 & -2 & -3 & 1 & -1 & 0 \\ 1 & 0 & -2 & -1 & -1 & 3 & 0 \\ 1 & 0 & -2 & 1 & -1 & -3 & 0 \\ 1 & 0 & -2 & 3 & 1 & 1 & 0 \\ 1 & 1 & 1 & -3 & 1 & -1 & -3 \\ 1 & 1 & 1 & -1 & -1 & 3 & -1 \\ 1 & 1 & 1 & 1 & -1 & -3 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 & 3 \end{bmatrix}.$$

The last column is the result of multiplying the elements of columns 2 and 4 of $\mathbf{X}$ which give the linear effects in the age and opinion variables. Applying the estimation procedure gives the following parameter estimates.

parameter	estimate	z-value
μ	3.7572	70.62
γ_1^A	0.6247	11.07
γ_2^A	-0.0011	-0.04
γ_1^O	0.2877	11.11
γ_2^O	-0.3186	-6.30
γ_3^O	-0.2621	-12.55
α	0.1014	3.34

Although the parameters differ from the linear by linear association model, the expected frequencies are identical and hence also LR and χ^2. The reason for this is that the columns of the design matrices of the two models generate the same vector space.

3.6 ASSOCIATION MODELS

Goodman (1979a) and Clogg (1982) discuss association models for an $I \times J$ cross-classification table having ordered categories. The $(I-1)(J-1)$ odds ratios for the table are used in formulating suitable models. For 2×2 tables formed from the adjacent rows i and $i+1$, the columns j and $j+1$, let θ_{ij} denote the corresponding odds ratio where

$$\theta_{ij} = \frac{F_{ij}F_{i+1,j+1}}{F_{i,j+1}F_{i+1,j}} \ , \quad i = 1,\ldots,I-1 \ ; \ j = 1,\ldots,J-1. \tag{3.26}$$

The model of statistical independence between the row and column categories may be expressed as

$$\theta_{ij} = 1, \quad \text{for} \quad i = 1,\ldots,I-1 \ ; \ j = 1,\ldots,J-1. \tag{3.27}$$

This is sometimes called the null association model.

If the odds ratios satisfy

$$\theta_{ij} = \theta, \quad \text{for} \quad i = 1,\ldots,I-1 \ ; \ j = 1,\ldots,J-1; \tag{3.28}$$

then we have the uniform association model.

The row-effect association model is

$$\theta_{ij} = \theta_{i.} \ , \quad i = 1, \ldots, I-1 \ ; \ j = 1, \ldots, J-1, \tag{3.29}$$

where $\theta_{i.}$ is unspecified.

The column-effect association model is

$$\theta_{ij} = \theta_{.j} \ , \quad i = 1, \ldots, I-1 \ ; \ j = 1, \ldots, J-1. \tag{3.30}$$

A model which includes the effects of the rows and columns for the θ_{ij} is

$$\theta_{ij} = \theta_{i.}\theta_{.j} \ , \quad i = 1, \ldots, I-1 \ ; \ j = 1, \ldots, J-1. \tag{3.31}$$

The estimated frequencies for these models can be found by using the procedure of Proposition 1 and expressing the conditions given in the various models in terms of the constraints $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$.

To illustrate the procedure, consider a 4×6 table. For this table there are $3 \times 5 = 15$ odds ratios, θ_{ij}, $i = 1, \ldots, 3$; $j = 1, \ldots, 5$ as defined earlier. If the expected frequencies for the cells are arranged into a vector $\mathbf{F} = (F_{11}, F_{12}, \ldots, F_{64})'$, then the log-odds ratios may be found by the expression $\mathbf{A}\ln(\mathbf{F})$, where if

$$\mathbf{B} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

and $\mathbf{O}$ is a 5×6 matrix of zeros, then

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{B} \end{bmatrix} .$$

The model in equation (3.28) may be written as $\ln(\theta_{ij}) = \ln(\theta) = \mu$ and hence the uniform association model may be written in the form

$$\mathbf{A}\ln(\mathbf{F}) = \mathbf{1}_{15}\mu \ .$$

Let

$$\mathbf{P} = \mathbf{I}_{15} - \mathbf{1}_{15}(\mathbf{1}'_{15}\mathbf{1}_{15})^{-1}\mathbf{1}'_{15},$$

then

$$\mathbf{PA}\ln(\mathbf{F}) = \mathbf{0} \ , \text{ thus } \boldsymbol{g}(\mathbf{F}) = \mathbf{PA}\ln(\mathbf{F}) = \mathbf{0}$$

and $\mathbf{G}_F = \mathbf{PAD}_F^{-1}$.

From Proposition 1, the ML estimate for the vector of expected frequencies is

$$\begin{aligned} \widehat{\mathbf{F}}_c &= \mathbf{x} - (\mathbf{G}_F\mathbf{D}_F)'(\mathbf{G}_x\mathbf{D}_F\mathbf{G}'_F)^{-1}\mathbf{PA}\ln(\mathbf{x}) + o(||\mathbf{x}-\mathbf{F}||) \\ &= \mathbf{x} - (\mathbf{PA})'(\mathbf{PAD}_x^{-1}\mathbf{A}'\mathbf{P})^{-1}\mathbf{PA}\ln(\mathbf{x}) + o(||\mathbf{x}-\mathbf{F}||). \qquad (3.32) \end{aligned}$$

Iteration takes place over $\mathbf{x}$.

The row-effect association model in (3.29) may be written as

$$\ln(\theta_{ij}) = \ln\theta_{i.} = \alpha_i \ , \ \ i = 1,\ldots,I-1 \ ; \ j = 1,\ldots,J-1.$$

This model may be written in the form

$$\mathbf{A}\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta} \ ,$$

where $\mathbf{X} = \mathbf{I}_3@\mathbf{1}_5$ and $\boldsymbol{\beta} = (\alpha_1, \alpha_2, \alpha_3)'$. Let

$$\mathbf{P} = \mathbf{I}_{15} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ . \qquad (3.33)$$

Then as before

$$\boldsymbol{g}(\mathbf{F}) = \mathbf{PA}\ln(\mathbf{F}) = \mathbf{0} \ \text{ and } \ \mathbf{G}_F = \mathbf{PAD}_F^{-1}$$

and the ML estimates for the frequencies can be found by applying (3.32) with **P** given in equation (3.33).

The column-effect association model may be written as

$$\ln(\theta_{ij}) = \ln\theta_{.j} = \gamma_j \quad , \quad i = 1, \ldots, I-1 \; ; \; j = 1, \ldots, J-1.$$

This model may now be written as

$$\mathbf{A}\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta} \quad , \quad \text{where } \boldsymbol{\beta} = (\gamma_1, \ldots, \gamma_5)'$$

and $\mathbf{X} = \mathbf{1}_3 @ \mathbf{I}_5$ and the ML estimates may be found by using equation (3.32) with the latter **X** used for **P** in equation (3.33).

The model in equation (3.31) may be written as

$$\ln(\theta_{ij}) = \ln\theta_{i.} + \ln\theta_{.j} = \alpha_i + \gamma_j \quad , \quad i = 1, \ldots, I-1 \; ; \; j = 1, \ldots, J-1$$

which can also be written in the form

$$\mathbf{A}\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta} \; , \; \text{where } \boldsymbol{\beta} = (\alpha_1, \alpha_2, \alpha_3, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)'$$

and

$$\mathbf{X} = [\mathbf{I}_3 @ \mathbf{1}_5 \; , \; \mathbf{1}_3 @ \mathbf{I}_5].$$

The ML estimates may once again be found by using equation (3.32) with **P** defined using the latter **X**.

Example 3.10
Consider the data in Table 3.10, which was analysed by Goodman (1979).

If the procedure described earlier is used to fit the RC model in (3.31) to the data, then the figures in parentheses are the estimated expected frequencies for the model. The likelihood ratio statistic $LR = 3.045$ and $\chi^2 = 3.057$. The degrees of freedom $df = (I-2)(J-2) = 4 \times 2 = 8$. The program for the example may be found in the Appendix.

TABLE 3.10: CROSS-CLASSIFICATION OF SUBJECTS ACCORDING TO THEIR MENTAL HEALTH AND PARENTS' SOCIOECONOMIC STATUS

Mental Health Status	Parents' Socioeconomic Status					
	A	B	C	D	E	F
Well	64 (63.8)	57 (59.2)	57 (56.8)	72 (68.3)	36 (36.4)	21 (22.6)
Mild	94 (94.4)	94 (91.9)	105 (106.4)	141 (143.5)	97 (94.1)	71 (71.7)
Moderate	58 (57.8)	54 (51.6)	65 (62.8)	77 (83.2)	54 (58.6)	54 (47.9)
Impaired	46 (46.0)	40(42.3)	60 (61.0)	94 (89.0)	78 (75.9)	71 (74.8)

It is also possible to construct models for the $I \times J$ contingency table in terms of odds. Goodman (1983) proposes a class of models based on the log-odds for cross-classifications where the categories of the response variable can be ordered.

For adjacent categories j and $j+1$ of the variable B, define the odds

$$\Omega_{ij} = \frac{F_{ij}}{F_{i,j+1}} \ , \ i = 1, \ldots, I \ ; \ j = 1, \ldots, J-1. \tag{3.34}$$

Let $\mathbf{\Psi}_{ij} = \ln \Omega_{ij}$, then Goodman (1983) defines three models for the log-odds $\mathbf{\Psi}_{ij}$, namely the null log-odds model

$$\mathbf{\Psi}_{ij} = 0 \ , \ i = 1, \ldots, I \ ; \ j = 1, \ldots, J-1; \tag{3.35}$$

the uniform log-odds model

$$\mathbf{\Psi}_{ij} = \Psi_i^A \ , \ i = 1, \ldots, I \ ; \ j = 1, \ldots, J-1; \tag{3.36}$$

where Ψ_i^A are unspecified, and the parallel log-odds model

$$\mathbf{\Psi}_{ij} = \Psi_i^A + \gamma_j^B \ , \ i = 1, \ldots, I \ ; \ j = 1, \ldots, J-1; \tag{3.37}$$

where the Ψ_j^B satisfy an appropriate constraint.

The expected frequencies for these models can also be found by applying the estimation procedure of Proposition 1.

Consider a 3×3 contingency table and suppose the model in equation (3.37) is to be fitted to the data. There are $2 \times 3 = 6$ odds for the table, and the model may be written in the form

$$\mathbf{A}\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} ,$$

$$\boldsymbol{\beta} = (\Psi_1^A, \Psi_2^A, \Psi_3^A, \gamma_1^B, \gamma_2^B)' \ ,$$

and

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} .$$

Now let $\mathbf{P} = \mathbf{I}_6 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\boldsymbol{g}(\mathbf{F}) = \mathbf{PA}\ln(\mathbf{F}) = \mathbf{0}$ and $\mathbf{G}_F = \mathbf{PAD}_F^{-1}$. The ML estimation procedure takes the form of that for the odds-ratio in (3.32) earlier.

Chapter 4

SQUARE TABLES, QUASI-INDEPENDENCE AND STRUCTURAL ZEROS

In this chapter models suitable for square tables will be discussed. The symmetry and quasi-symmetry models have already been considered in §3.4. The marginal homogeneity problem, conditional symmetry model, the rater agreement problem and various diagonal symmetry models will be discussed. The ML estimation procedure also proves very useful for modelling data where there are structural zeros and data for which the quasi-independence model is suitable.

4.1 SQUARE TABLES

Often cross-classifications of data result in tables which are square and a number of models have been proposed to model such data.

4.1.1 Marginal Homogeneity

Consider the problem of comparing two marginal distributions of a square table. Marginal homogeneity states that $\pi_{i+} = \pi_{+i}$, $i = 1, \ldots, r$.

Example 4.1

Consider the well known "unaided vision data set" of Bhapkar (1966) also discussed by Bishop Fienberg and Holland (1975). The observed frequencies and expected frequencies (in braces) for the marginal homogeneity model are shown in Table 4.1.

TABLE 4.1: UNAIDED DISTANCE VISION FOR 7477 WOMEN AGED 30-39

	Left eye				
Right eye	Highest grade	Second grade	Third grade	Lowest grade	Total
Highest grade	1520 (1520)	266 (252.48)	124 (111.84)	66 (56.97)	1976
Second grade	234 (247.24)	1512 (1512)	432 (409.42)	78 (70.59)	2256
Third grade	117 (131.27)	362 (383.13)	1772 (1772)	205 (195.26)	2456
Lowest grade	36 (42.79)	82 (91.63)	179 (188.40)	492 (492)	789
Total	1970	2222	2507	841	7477

The hypothesis of marginal homogeneity is $H_o : \pi_{i+} = \pi_{+i}$, $i = 1, 2, 3, 4$.

Let $\boldsymbol{\pi}' = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}, \pi_{21}, \ldots, \pi_{24}, \ldots, \pi_{41}, \ldots, \pi_{44})$.
Notice that $\pi_{1+} = \pi_{+1}$ gives

$$\pi_{11} + \pi_{12} + \pi_{13} + \pi_{14} = \pi_{11} + \pi_{21} + \pi_{31} + \pi_{41}$$
$$\text{i.e. } \pi_{12} + \pi_{13} + \pi_{14} - \pi_{21} - \pi_{31} - \pi_{41} = 0$$

which can be written as

$$(0\ 1\ 1\ 1\ \ -1\ 0\ 0\ 0\ \ -1\ 0\ 0\ 0\ \ -1\ 0\ 0\ 0)\boldsymbol{\pi} = 0.$$

Furthermore H_o can be written as $\mathbf{A}\boldsymbol{\pi} = \mathbf{0}$, where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

Since $\mathbf{A}$ is singular, any 3 rows of $\mathbf{A}$ may be used to represent H_o and thus 3 degrees of freedom are associated with H_o.

The constraints may be written as $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{A}\boldsymbol{\pi} = \mathbf{0}$ or $\boldsymbol{g}(\mathbf{F}) = \mathbf{AF} = \mathbf{0}$. Hence

$$\mathbf{G} = \frac{\partial}{\partial \mathbf{F}} \boldsymbol{g}(\mathbf{F}) = \mathbf{A} .$$

Thus the estimated expected frequency is found iteratively using (2.7), where at step r in the iterative procedure ($r = 0, 1, 2 \ldots$),

$$\boldsymbol{\mu}^{(r+1)} = \mathbf{x} - (\mathbf{AD}_{\mu^{(r)}})'(\mathbf{AD}_{\mu^{(r)}}\mathbf{A}')^{-1}\mathbf{Ax} ,$$

where, $\boldsymbol{\mu}^{(0)} = \mathbf{x}$ is the original frequency vector. Applying the above estimation procedure gives ML estimates of the expected frequencies, shown in parentheses in Table III, with $\chi^2 = 11.97$ and $LR = 11.99$ with $df = 3$.

The program for this example can be found in the Appendix.

4.1.2 Conditional Symmetry

For ordered classifications, when symmetry does not hold, often $\pi_{ij} > \pi_{ji}$ for all $i < j$, or $\pi_{ij} < \pi_{ji}$ for all $i < j$. A generalization of symmetry that has this property is

$$\ln(F_{ij}) = \mu + \lambda_i + \lambda_j + \lambda_{ij} + \tau I(i < j) \qquad (4.1)$$

where all $\lambda_{ij} = \lambda_{ji}$ and where $I(\cdot)$ is the indicator function. The corresponding logit model is

$$\ln(F_{ij}/F_{ji}) = \tau \ \text{ for } i < j \qquad (4.2)$$

or

$$\ln(F_{ij}) - \ln(F_{ji}) = \tau \ \text{ for } i < j \ . \qquad (4.3)$$

This model can be formulated as $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$ as follows. Consider a 4×4 table with

$$\mathbf{F} = (F_{11}, F_{12}, \ldots, F_{14}, F_{21}, \ldots, F_{24}, F_{31}, \ldots, F_{34}, F_{41}, \ldots, F_{44}) \ .$$

The model in (4.3) can be written as

$$\mathbf{C} \ln(\mathbf{F}) = \mathbf{1}\tau \text{ where } \mathbf{1}' = (1,1,1,1,1,1) \tag{4.4}$$

and $\mathbf{C}$ is the matrix

F_{11}	F_{12}	F_{13}	F_{14}	F_{21}	F_{22}	F_{23}	F_{24}	F_{31}	F_{32}	F_{33}	F_{34}	F_{41}	F_{42}	F_{43}	F_{44}
0	1	0	0	−1	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	−1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	−1	0	0	0
0	0	0	0	0	0	1	0	0	−1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	−1	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	−1	0

(4.5)

Let $\mathbf{X} = \mathbf{1}_6$ and $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The implied constraints for the model are

$$\boldsymbol{g}(\mathbf{F}) = \mathbf{P}\mathbf{C} \ln(\mathbf{F}) = \mathbf{0} \ ,$$

or

$$\mathbf{K} \ln(\mathbf{F}) = \mathbf{0}, \text{ where } \mathbf{K} = \mathbf{P}\mathbf{C}.$$

Furthermore

$$\mathbf{G} = \frac{\partial}{\partial \mathbf{F}} \boldsymbol{g}(\mathbf{F}) = \mathbf{K}\mathbf{D}_F^{-1} \text{ and } \mathbf{V} = \mathbf{D}_F \ .$$

Thus the ML estimates for the expected frequencies are given by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - \mathbf{K}'(\mathbf{K}\mathbf{D}_x^{-1}\mathbf{K}')^{-1}\mathbf{K} \ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \ ,$$

where $\mathbf{D}_x = \text{diag}(\mathbf{x})$ and iteration takes place over $\mathbf{x}$.

Example 4.2

Table 4.2, taken from Breslow (1982), and discussed by Agresti (1990) p.364 compares 80 esophageal cancer patients with 80 matched control subjects. The response is the number of beverages reported drunk at "burning hot" temperatures. The analysis is to establish whether cases tended to drink more beverages hot than did controls.

TABLE 4.2: NUMBER OF BEVERAGES DRUNK AT BURNING HOT TEMPERATURES, FOR ESOPHAGEAL CANCER CASE-CONTROL PAIRS

	Control			
Case	0	1	2	3
0	31	5	5	0
1	12	1	0	0
2	14	1	2	1
3	6	1	1	0

The ordinal nature of the data may be used in the construction of an appropriate model. If the symmetry model is fitted to the data, then we observe nonnegative residuals below the main diagonal, which indicates a systematic lack of fit. The conditional symmetry model of (4.1) gives a much better fit, with $\chi^2 = 3.6$ based on $df = 5$. The ML estimate of τ is -1.1584 and $\frac{\widehat{F}_{ij}}{\widehat{F}_{ji}} = \exp(\tau) = 0.314$. The probability that a control drank k more beverages burning hot than did the case is estimated to be 0.314 times the probability that the case drank k more beverages burning hot than the control, for $k = 1, 2, 3$. The program for this example can be found in the Appendix.

4.1.3 The Diagonals-Parameter Model

Goodman (1972, 1979b) presents models in which diagonals that are equidistant from the main diagonal exhibit similar patterns for expected cell frequencies. The diagonals-parameter symmetry model is of the form

$$F_{ij} = F_{ji}\delta_{j-i}\ , \quad i < j, \quad i = 1, \ldots, r\ . \tag{4.6}$$

The parameters of this model are thus $\delta_1, \delta_2, \ldots, \delta_{r-1}$. This model can be written as

$$\ln(F_{ij}) - \ln(F_{ji}) = \ln(\delta_{j-i})$$

or

$$\mathbf{C}\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\alpha}\ ,$$

where $\alpha_k = \ln(\delta_k)$, and $\mathbf{C}$ and $\mathbf{X}$ are appropriately constructed.

Consider for example, a 4×4 contingency table as discussed under the conditional symmetry model. Then $\mathbf{C}$ is the matrix in equation (4.5), and

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}\ , \quad \boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \alpha_3)\ .$$

Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The model can then be formulated in terms of the corresponding constraint function $\boldsymbol{g}(\mathbf{F}) = \mathbf{PC}\ln(\mathbf{F}) = \mathbf{0}$, as for the conditional symmetry model and the estimation procedure takes on the same form, namely

$$\widehat{\mathbf{F}}_c = \mathbf{x} - \mathbf{K}'(\mathbf{K}\mathbf{D}_x^{-1}\mathbf{K}')^{-1}\mathbf{K}\ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||)$$

where $\mathbf{K} = \mathbf{PC}$, and iteration takes place over $\mathbf{x}$.

Example 4.3
Consider the data of Example 4.1. Fit the diagonals-parameter symmetry model of equation (4.6) to the data.

The ML estimates of the frequencies are given in Table 4.3.

TABLE 4.3: UNAIDED DISTANCE VISION FOR 7477 WOMEN AGED 30-39

Right Eye Grade	Left Eye Grade			
	Best	Second	Third	Worst
Best	1520.00	269.07	121.40	66.00
Second	230.93	1512.00	427.28	80.60
Third	119.60	366.72	1772.00	206.65
Worst	36.00	79.40	177.35	492.00

The ML estimates for the parameters are

$$\widehat{\alpha}_1 = 0.15286, \quad \widehat{\alpha}_2 = 0.01496 \quad \text{and} \quad \widehat{\alpha}_3 = 0.60614,$$
$$\widehat{\delta}_1 = 1.1652\ , \quad \widehat{\delta}_2 = 1.0151 \quad \text{and} \quad \widehat{\delta}_3 = 1.8333\ .$$

The estimate $\widehat{\delta}_k$ is the estimated odds that an observation falls in cell (i, j) satisfying $j - i = k$, instead of in a cell satisfying $j - i = -k$, $k = 1, 2, 3$.

This model fits the data extremely well, $LR = 0.50$ and $\chi^2 = 0.50$ based on $df = 3$. The p-value for both tests being 0.9194.
The program for this example can also be found in the Appendix.

4.1.4 Models for Mobility Tables

Square tables are often used to describe the mobility pattern of a population. In a mobility table it is often the case that the probability of observing an individual in a cell, is smaller the further away it is located from the principal diagonal. A model that will describe the pattern is

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \delta|i - j|\ , \quad \delta < 0\ . \tag{4.7}$$

This model gives a smaller expected frequency in cell (i,j) the larger the value of $|i-j|$, i.e. the further the cell is from the principal diagonal.

There is however often a distinct tendency against changing from a group (cell) in a mobility table. This will imply that the observed frequencies for cells on the principal diagonal of the table will be large and the model in (4.7) will often be inadequate for the data. A model which will better describe the above-mentioned trend is

$$\ln(F_{ij}) = \begin{cases} \mu + \lambda_i^A + \lambda_j^B + \delta|i-j| & , \quad i \neq j \\ \mu + \lambda_i^A + \lambda_j^B + \alpha_i & , \quad i = j \end{cases} \tag{4.8}$$

The estimation procedure can once again be implemented for parameter estimation. The model in (4.8) can be written as

$$\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta} ,$$

where $\mathbf{X}$ is the design matrix for the mean, main effects and covariates and

$$\boldsymbol{\beta}' = (\mu, \lambda_1^A, \ldots, \lambda_{r-1}^A, \lambda_1^B, \ldots, \lambda_{r-1}^B, \delta, \alpha_1, \ldots, \alpha_r) .$$

Example 4.4

Haberman (1974b) gives a cross classification table of husband's and wive's highest degree attained. The data are given in the Table 4.4.

If we fit the model in (4.3), then the column for the covariate, i.e. $|i-j|$ is $\mathbf{c}_1$, where

$$\mathbf{c}_1' = (0\ 1\ 2\ 3\ 1\ 0\ 1\ 2\ 2\ 1\ 0\ 1\ 3\ 2\ 1\ 0),$$

and the other columns in $\mathbf{X}$ are the columns for the mean and main effects. If we fit the model in (4.4), then the matrix of the covariates for the parameters $\delta, \alpha_1, \alpha_2, \alpha_3$ and α_4, is $\mathbf{C}_2$, where

$$\mathbf{C}_2' = \begin{bmatrix} 0 & 1 & 2 & 3 & 1 & 0 & 1 & 2 & 2 & 1 & 0 & 1 & 3 & 2 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

TABLE 4.4: MARRIED RESPONDENTS IN 1974 GENERAL SOCIAL SURVEY CROSS-CLASSIFIED BY HIGHEST DEGREES ATTAINED

Husband's highest degree	Wife's highest degree: Less than high school diploma	High school diploma junior college degree	Bachelor's degree	Graduate degree	Total
Less than high school diploma	259 (250.95)[a] (259)[b]	123 (123.86) (121.92)	2 (8.64) (2.67)	0 (0.54) (0.41)	384
High school diploma or junior college degree	82 (78.96) (83.08)	370 (381.73) (370)	30 (26.63) (31.13)	7 (1.67) (4.79)	489
Bachelor's degree	5 (10.91) (3.44)	59 (52.77) (58.76)	34 (36.06) (34.)	4 (2.26) (5.80)	102
Graduate degree	2 (7.16) (2.48)	41 (34.64) (42.33)	29 (23.67) (27.20)	8 (14.53) (8)	80
Total	348	593	95	19	1055

[(a)(b): the figures in parentheses are the expected frequencies for the models in (4.7) and (4.8) respectively.]

Both models (4.3) and (4.4) can be written in the form $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ has as columns the relevant main effects and covariates. Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then the function $\boldsymbol{g}(\mathbf{F}) = \mathbf{P}\ln(\mathbf{F}) = \mathbf{0}$, is the corresponding constraint function for the model. The estimation procedure is once more given by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - \mathbf{P}(\mathbf{P}\mathbf{D}_x^{-1}\mathbf{P})^{-1}\mathbf{P}\ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \ ,$$

and iteration is over $\mathbf{x}$.

For model (4.7) where the diagonal elements are not fixed, the likelihood ratio statistic $LR = 39.97$, $\chi^2 = 38.26$ and $df = 8$. The fit is very poor. If model (4.8) is considered where the principal diagonal elements are fixed, then $LR = 3.48$ and $\chi^2 = 3.19$ with $df = 4$ and there is a significant improvement in the fit of the model. For model (4.8), $\widehat{\delta} = -1.6888$ with normalized value $z = -8.27$, showing the significance of the parameter.

4.1.5 Models for Agreement Among Raters

Landis and Koch (1977), Tanner and Young (1985) and Agresti (1988) address the problem of agreement between ratings done on an ordinal scale and propose various models for the cross-classification which can be displayed as a square contingency table when considering the joint ratings of the two raters. A model suggested is

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \delta(i,j) \tag{4.9}$$

where

$$\delta(i,j) = \begin{cases} \delta & , \ \text{if} \ \ i = j \\ 0 & , \ \text{otherwise} \ . \end{cases}$$

The parameter δ included for the cells on the principal diagonal represents agreement beyond what is expected by chance. Another model is

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \delta(i,j) \tag{4.10}$$

where

$$\delta(i,j) = \begin{cases} \delta_i & , \ \ i = j, \ \ i = 1, \ldots, r \\ 0 & , \ \ \text{otherwise} \ . \end{cases}$$

The δ_i are for differences by response category.

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j \tag{4.11}$$

where $u_1 < u_2 < \cdots < u_r$ are fixed scores. This is the linear-by-linear association model previously discussed.

An extension of (4.7) where a perfect fit is imposed on the principal diagonal cells, is

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta(i,j) \tag{4.12}$$

where

$$\delta(i,j) = \begin{cases} \delta_i & , \quad i = j, \quad i = 1, \ldots, r, \\ 0 & , \quad \text{otherwise} . \end{cases}$$

The above-mentioned models can all be written in the form

$$\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta} \ ,$$

where the design matrix will include columns for the mean, main effects and any covariate appearing in the model. The estimation procedure of Proposition 1 can thus be implemented for any of these models where the implied constraint function

$$\boldsymbol{g}(\mathbf{F}) = \mathbf{P}\ln(\mathbf{F}) = \mathbf{0} \ , \text{where} \ \ \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \ .$$

The estimation procedure is given by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - \mathbf{P}(\mathbf{P}\mathbf{D}_x^{-1}\mathbf{P})^{-1}\mathbf{P}\ln(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \ ,$$

where iteration takes place over $\mathbf{x}$.

Example 4.5

Consider the data from Agresti (1988) taken from Landis and Koch (1977). Table 4.5 displays diagnoses of multiple sclerosis for two neurologists who classified patients in two sites, Winnipeg and New Orleans. The diagnostic classes are (1) certain multiple sclerosis (2) probable multiple sclerosis (3) possible multiple sclerosis (4) doubtful, unlikely, or definitely not multiple sclerosis.

TABLE 4.5: DIAGNOSTIC CLASSIFICATIONS REGARDING MULTIPLE SCLEROSIS

New Orleans neurologist	Winnipeg neurologist							
	Winnipeg patients				New Orleans patients			
	1	2	3	4	1	2	3	4
1	38	5	0	1	5	3	0	0
2	33	11	3	0	3	11	4	0
3	10	14	5	6	2	13	3	4
4	3	7	3	10	1	2	4	14

Let S: site, $R1$: rating by New Orleans neurologist, $R2$: rating by Winnipeg neurologist.

Agresti fits the model

$$\ln(F_{ij}) = \mu + \lambda_i^{R1} + \lambda_j^{R2} + \beta u_i u_j + \delta(i,j)$$

with $\delta(i,j)$ defined under model (4.9) for the Winnipeg patients and the New Orleans patients separately and finds

Site	$\widehat{\beta}$	$\widehat{\delta}$	LR	df
Winnipeg	0.804	-0.028	9.4	7
New Orleans	1.041	0.028	8.8	7

The similar results for the patients suggest a single model for the $4 \times 4 \times 2$

cross-classification. Thus consider the following model

$$\ln(F_{ijk}) = \mu+\lambda_i^S+\lambda_j^{R1}+\lambda_k^{R2}+\lambda_{ij}^{SR1}+\lambda_{ik}^{SR2}+\beta u_j u_k+\delta(j,k), \quad i=1,2, \quad j,k=1,\ldots,4$$

where $u_i, u_j = 1, \ldots, 4$ and

$$\delta(j,k) = \begin{cases} 1 & , \text{ if } j=k \\ 0 & , \text{ otherwise .} \end{cases}$$

The matrix for the covariates is $\mathbf{C}$, where

$$\mathbf{C}' = \begin{bmatrix} 1 & 2 & 3 & 4 & 2 & 4 & 6 & 8 & 3 & 6 & 9 & 12 & 4 & 8 & 12 & 16 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} :\beta \\ :\delta \end{matrix}$$

The model can be written as $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ has the columns for the mean, main effects, site $\times$ neurologist interaction and the columns of $\mathbf{C}$. The estimation procedure follows as described earlier and the ML-estimates are $\widehat{\delta} = 0.0198$ and $\widehat{\beta} = 0.8548$ with $LR = 19.04$ and $\chi^2 = 24.70$ based on $df = 16$.

Since δ is not significant, remove it from the model. The ML estimate for β is then $\widehat{\beta} = 0.8612$ and $LR = 19.03$ based on $df = 17$ with p-value 0.3268 indicating an adequate fit.

4.1.6 Quasi-symmetry

In Chapter 3 estimates for the expected frequencies under the quasi-symmetry (QS) model were found by expressing the constraints in terms of the *parameters* in the loglinear model. It is also possible to formulate the constraints of the QS model in terms of constraints on the *frequencies* using odds ratios. From Agresti (1990) equation (10.14), the QS model can be written in the multiplicative form

$$\pi_{ij} = \alpha_i\beta_j\gamma_{ij}, \quad \text{where } \gamma_{ij} = \gamma_{ji} \ \forall \ i,j, \text{ and } \alpha_i, \beta_j, \gamma_{ij} > 0.$$

Quasi-symmetry can also be written in terms of symmetry of odds ratios. The quasi-symmetry model holds if and only if for all integers i, j, k and l between 1 and R

$$\frac{\pi_{ik}\pi_{jl}}{\pi_{il}\pi_{jk}} = \frac{\pi_{ki}\pi_{lj}}{\pi_{li}\pi_{kj}} \,. \tag{4.13}$$

Now fixing one cell, by taking $j = l = R$ and forming all odds ratios, equation (4.13) becomes

$$\frac{\pi_{ik}\pi_{RR}}{\pi_{iR}\pi_{Rk}} = \frac{\pi_{ki}\pi_{RR}}{\pi_{Ri}\pi_{kR}} \text{ or } \pi_{ik}\pi_{kR}\pi_{Ri} = \pi_{ki}\pi_{iR}\pi_{Rk} \ ,$$

(which is the result of Agresti (1990) exercise 10.26). These constraints can now be written in the form $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}$ and the proposed estimation procedure may be used to find the ML estimates of the expected frequencies for the quasi-symmetry model.

As an illustration, consider a 4×4 table. For the quasi-symmetry model to hold, the following constraints must hold

$$\begin{array}{lcl} \pi_{12}\pi_{24}\pi_{41} = \pi_{21}\pi_{14}\pi_{42} & & F_{12}F_{24}F_{41} - F_{21}F_{14}F_{42} = 0 \\ \pi_{13}\pi_{34}\pi_{41} = \pi_{31}\pi_{14}\pi_{43} & \text{or} & F_{13}F_{34}F_{41} - F_{31}F_{14}F_{43} = 0 \\ \pi_{23}\pi_{34}\pi_{42} = \pi_{32}\pi_{24}\pi_{43} & & F_{23}F_{34}F_{42} - F_{32}F_{24}F_{43} = 0 \ , \end{array}$$

which is of the form $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$ and $\mathbf{G}_F = \dfrac{\partial}{\partial \mathbf{F}}\boldsymbol{g}(F)$ is a 3×16 matrix, with

$$\begin{array}{lll} \mathbf{G}_F(1,2) = F_{24}F_{41}, & \mathbf{G}_F(1,4) = -F_{21}F_{42}, & \mathbf{G}_F(1,5) = -F_{14}F_{42}, \\ \mathbf{G}_F(1,8) = F_{12}F_{41}, & \mathbf{G}_F(1,13) = F_{12}F_{24}, & \mathbf{G}_F(1,14) = -F_{21}F_{14}, \end{array}$$

$$\begin{array}{lll} \mathbf{G}_F(2,3) = F_{34}F_{41}, & \mathbf{G}_F(2,4) = -F_{31}F_{43}, & \mathbf{G}_F(2,9) = -F_{14}F_{43}, \\ \mathbf{G}_F(2,12) = F_{13}F_{41}, & \mathbf{G}_F(2,13) = F_{13}F_{34}, & \mathbf{G}_F(2,15) = -F_{31}F_{14}, \end{array}$$

$$\begin{array}{lll} \mathbf{G}_F(3,7) = F_{34}F_{42}, & \mathbf{G}_F(3,8) = -F_{32}F_{43}, & \mathbf{G}_F(3,10) = -F_{24}F_{43}, \\ \mathbf{G}_F(3,12) = F_{23}F_{42}, & \mathbf{G}_F(3,14) = F_{23}F_{34}, & \mathbf{G}_F(3,15) = -F_{32}F_{24}. \end{array}$$

All other $\mathbf{G}_F(i,j)$ are zero.

The ML estimates of the expected frequencies are found by using

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{V})'(\mathbf{G}_x\mathbf{V}\mathbf{G}_F')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \ .$$

$\mathbf{V}$ may be taken as $\mathbf{D}_F$ for a Poisson sampling procedure. A double iteration over $\mathbf{x}$ and $\mathbf{F}$ is necessary. The QS model may be considered for the following square table.

Example 4.6
Consider the migration example of Agresti (1990) where the table compares region of residence in 1985 with 1980 for a sample of U.S. residents and fit the QS model to the data.

TABLE 4.6: MIGRATION FROM 1980 TO 1985

Residence in 1980	Residence in 1985			
	North East	Mid West	South	West
North East	11607	100 (95.7)	366 (370.4)	124 (123.8)
Mid West	87 (91.2)	13677	515 (501.7)	302 (311.1)
South	172 (167.6)	225 (238.3)	17819	270 (261.1)
West	63 (63.2)	176 (166.9)	286 (294.9)	10192

(Figures in parentheses are the ML estimates of the expected frequencies for the QS model).

The likelihood ratio statistic $LR = 2.99$ and $\chi^2 = 2.98$ with $df = 3$.

4.2 QUASI-INDEPENDENCE

4.2.1 Quasi-Independence in the $I \times J$ Table

Using the notation of Goodman (1994), let the odds ratio $\theta_{ij,rs}$ be defined as follows

$$\theta_{ij,rs} = \frac{\pi_{ij}\pi_{rs}}{\pi_{is}\pi_{rj}} \quad \text{for } i < r \text{ and } j < s. \tag{4.14}$$

Instead of considering all possible odds ratios for the $I \times J$ table, consider the situation where a given subset S of the cells (i, j) in the table, is of interest.

The concept of quasi-independence (QI) is stated as

$$\pi_{ij} = \alpha_i \beta_j, \quad \forall \text{ cells } (i,j) \text{ in } S.$$

From equation (4.14) the quasi-independence model implies that

$$\theta_{ij,rs} = \frac{\alpha_i \beta_j \alpha_r \beta_s}{\alpha_i \beta_s \alpha_r \beta_j} = 1, \quad \forall \ T_{ij,rs} \text{ in } T$$

where T is the set of 2×2 tables $T_{ij,rs}$ in which cells (i,j), (r,s), (i,s) and (r,j) are in S.

The conditions above can be written in terms of constraints $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}$ and the proposed ML estimation procedure can be employed to find the ML estimates of the expected frequencies in the QI model.

Example 4.7
Consider the migration example of Example 4.6 which compares region of residence in 1985 with 1980 for a sample of U.S. residents. Consider whether, for people who moved, residence in 1985 is independent of region in 1980. We investigate independence for cells not on the principal diagonal, i.e. a subtable S in T.

Let $\mathbf{F} = (F_1, F_2, \ldots, F_{16})'$ denote the expected frequencies arranged as a column vector. For the quasi-independence model there are 5 cross product ratios, which are each set equal to one. These are given by

$$\frac{F_5 F_{12}}{F_8 F_9} = 1, \quad \frac{F_3 F_8}{F_4 F_7} = 1, \quad \frac{F_2 F_{15}}{F_3 F_{14}} = 1, \quad \frac{F_5 F_{15}}{F_7 F_{13}} = 1 \text{ and } \frac{F_9 F_{14}}{F_{10} F_{13}} = 1.$$

These constraints can now be expressed in the form $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$, where

$$\begin{aligned}
g_1(\mathbf{F}) &= F_5 F_{12} - F_8 F_9 = 0, \\
g_2(\mathbf{F}) &= F_3 F_8 - F_4 F_7 = 0, \\
g_3(\mathbf{F}) &= F_2 F_{15} - F_3 F_{14} = 0, \\
g_4(\mathbf{F}) &= F_5 F_{15} - F_7 F_{13} = 0, \\
g_5(\mathbf{F}) &= F_9 F_{14} - F_{10} F_{13} = 0.
\end{aligned}$$

TABLE 4.7: MIGRATION FROM 1980 TO 1985

Residence in 1980	Residence in 1985			
	North East	Mid West	South	West
North East	11607	100 (126.6)	366 (312.9)	124 (150.5)
Mid West	87 (117.4)	13677	515 (531.1)	302 (255.5)
South	172 (133.2)	225 (243.8)	17819	270 (290.0)
West	63 (71.4)	176 (130.6)	286 (323.0)	10192

(Figures in parentheses are the ML estimates of the expected frequencies for the QI model).

The matrix $\mathbf{G}_F = \dfrac{\partial}{\partial \mathbf{F}} \boldsymbol{g}(\mathbf{F})$ is a 5×16 matrix, with

$$
\begin{array}{llll}
\mathbf{G}_F(1,5) = F_{12}, & \mathbf{G}_F(1,8) = -F_9, & \mathbf{G}_F(1,9) = -F_8, & \mathbf{G}_F(1,12) = F_5 \\
\mathbf{G}_F(2,3) = F_8, & \mathbf{G}_F(2,4) = -F_7, & \mathbf{G}_F(2,7) = -F_4, & \mathbf{G}_F(2,8) = F_3 \\
\mathbf{G}_F(3,2) = F_{15}, & \mathbf{G}_F(3,3) = -F_{14}, & \mathbf{G}_F(3,14) = -F_3, & \mathbf{G}_F(3,15) = F_2 \\
\mathbf{G}_F(4,5) = F_{15}, & \mathbf{G}_F(4,7) = -F_{13}, & \mathbf{G}_F(4,13) = -F_7, & \mathbf{G}_F(4,15) = F_5 \\
\mathbf{G}_F(5,9) = F_{14}, & \mathbf{G}_F(5,10) = -F_{13}, & \mathbf{G}_F(5,13) = -F_{10}, & \mathbf{G}_F(5,14) = F_9
\end{array}
$$

All other $\mathbf{G}_F(i,j)$ are zero.

The ML estimation procedure

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{V})'(\mathbf{G}_x\mathbf{V}\mathbf{G}_F')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||)$$

is implemented. If a Poisson sampling procedure is considered, then $\mathbf{V} = \mathbf{D}_F$ and a double iteration is required for convergence to the ML estimates of the expected frequencies for the QI model. The likelihood ratio, $LR = 69.51$ with $df = 5$.

Another approach to quasi-independence, is to express the QI model as a loglinear model, where

$$\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B, \quad \text{for} \quad i \neq j \ ,$$

i.e. quasi-independence for cells other than the cells on the principal diagonal.

For this model $\widehat{F}_{ii} = x_{ii}$, $i = 1, \ldots, 4$ and these constraints can be imposed through the matrix

$$\mathbf{C}_1' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} .$$

The quasi-independence model can be written as $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, where the design matrix $\mathbf{X}$ has columns $\mathbf{1}_{16}$, the main effects for the variables A and B and the columns of $\mathbf{C}_1$. Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then $\mathbf{P}\ln(\mathbf{F}) = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, and we thus have a function $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$ with $\mathbf{G}_F = \mathbf{P}\mathbf{D}_F^{-1}$. The estimation procedure is given by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{V})'(\mathbf{G}_x\mathbf{V}\mathbf{G}_F')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||)$$

and iteration is over $\mathbf{x}$.

4.2.2 Quasi Independence in a $R \times R$ Triangular Table

Goodman (1994) presents an explicit expression for the ML estimate of the frequencies expected in a $R \times R$ triangular contingency table under the quasi-independence model. For the $R \times R$ triangular table, with π_{ij} denoting the probability that an observation will fall in the ith row and jth column of the table, the QI model states that

$$\begin{aligned} \pi_{ij} &= \alpha_i \beta_j, \quad \text{for} \quad i \leq j, \\ &= 0, \qquad \text{for} \quad i > j, \end{aligned}$$

with $\alpha_i > 0$ and $\beta_j > 0$, for $i = 1, \ldots, R$, and $j = 1, \ldots, R$. Without loss of generality consider the table with positive probabilities in the upper-right triangle.

The problem of QI can also be easily modelled in terms of all possible odds ratios for the triangular table set equal to one, which can be expressed in terms of constraints on the cell probabilities.

Example 4.8

Consider the 5×5 contingency table taken from BFH (1975) wherein the initial and final condition of stroke patients is considered.

TABLE 4.8: INITIAL AND FINAL CONDITION OF STROKE PATIENTS

Initial	Final Condition				
condition	A	B	C	D	E
E	11 (15.66)	23 (21.92)	12 (11.93)	15 (11.48)	8 (8.0)
D	9 (6.16)	10 (8.63)	4 (4.69)	1 (4.52)	0 (0)
C	6 (4.43)	4 (6.20)	4 (3.37)	0 (0)	0 (0)
B	4 (3.75)	5 (5.25)	0 (0)	0 (0)	0 (0)
A	5 (5.0)	0 (0)	0 (0)	0 (0)	0 (0)

(Figures in parentheses are the ML estimates of the expected frequencies for the QI model).

For convenience arrange the cell probabilities π_{ij} as a vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{25})'$, where $\pi_1 = \pi_{11}$, $\pi_2 = \pi_{12}, \ldots, \pi_{25} = \pi_{55}$.

The QI model can be expressed in terms of 6 odds ratios (cross product ratios), each of which must be equal to one for independence to hold for the non-zero cells. The constraints are

$$\frac{\pi_1\pi_7}{\pi_2\pi_6}=1,\ \frac{\pi_1\pi_8}{\pi_3\pi_6}=1,\ \frac{\pi_1\pi_9}{\pi_4\pi_6}=1,\ \frac{\pi_1\pi_{12}}{\pi_2\pi_{11}}=1,\ \frac{\pi_1\pi_{13}}{\pi_3\pi_{11}}=1,\ \frac{\pi_1\pi_{17}}{\pi_2\pi_{16}}=1.$$

These constraints in turn can now be expressed in the form $\boldsymbol{g}(\boldsymbol{\pi})=\mathbf{0}$, where

$$\begin{aligned}
g_1(\boldsymbol{\pi}) &= \pi_1\pi_7-\pi_2\pi_6=0,\\
g_2(\boldsymbol{\pi}) &= \pi_1\pi_8-\pi_3\pi_6=0,\\
g_3(\boldsymbol{\pi}) &= \pi_1\pi_9-\pi_4\pi_6=0,\\
g_4(\boldsymbol{\pi}) &= \pi_1\pi_{12}-\pi_2\pi_{11}=0,\\
g_5(\boldsymbol{\pi}) &= \pi_1\pi_{13}-\pi_3\pi_{11}=0,\\
g_6(\boldsymbol{\pi}) &= \pi_1\pi_{17}-\pi_2\pi_{16}=0\ .
\end{aligned}$$

The matrix $\mathbf{G}_\pi=\dfrac{\partial}{\partial\boldsymbol{\pi}}\boldsymbol{g}(\boldsymbol{\pi})$ is a 6×25 matrix, with

$$\mathbf{G}_\pi(1,1)=\pi_7,\quad \mathbf{G}_\pi(1,2)=-\pi_6,\quad \mathbf{G}_\pi(1,6)=-\pi_2,\quad \mathbf{G}_\pi(1,7)=\pi_1\ ,$$

$$\mathbf{G}_\pi(2,1)=\pi_8,\quad \mathbf{G}_\pi(2,3)=-\pi_6,\quad \mathbf{G}_\pi(2,6)=-\pi_3,\quad \mathbf{G}_\pi(2,8)=\pi_1\ ,$$

$$\mathbf{G}_\pi(3,1)=\pi_9,\quad \mathbf{G}_\pi(3,4)=-\pi_6,\quad \mathbf{G}_\pi(3,6)=-\pi_4,\quad \mathbf{G}_\pi(3,9)=\pi_1\ ,$$

$$\mathbf{G}_\pi(4,1)=\pi_{12},\quad \mathbf{G}_\pi(4,2)=-\pi_{11},\quad \mathbf{G}_\pi(4,11)=-\pi_2,\quad \mathbf{G}_\pi(4,12)=\pi_1\ ,$$

$$\mathbf{G}_\pi(5,1)=\pi_{13},\quad \mathbf{G}_\pi(5,3)=-\pi_{11},\quad \mathbf{G}_\pi(5,11)=-\pi_3,\quad \mathbf{G}_\pi(5,13)=\pi_1\ ,$$

$$\mathbf{G}_\pi(6,1)=\pi_{17},\quad \mathbf{G}_\pi(6,2)=-\pi_{16},\quad \mathbf{G}_\pi(6,16)=-\pi_2,\quad \mathbf{G}_\pi(6,17)=\pi_1.$$

All other entries $\mathbf{G}_\pi(i,j)$ are equal to zero.

Let $\mathbf{F}=(F_1,\ldots,F_{25})'$ denote the vector of expected frequencies, then for example

$$\frac{\pi_1\pi_7}{\pi_2\pi_6}\text{ may be written as }\frac{(n\pi_1)(n\pi_7)}{(n\pi_2)(n\pi_6)}=1\text{ or }\frac{F_1F_7}{F_2F_6}=1$$

and similarly for the other cross product ratios. The constraints $\boldsymbol{g}(\boldsymbol{\pi}) = \mathbf{0}$ may be written in the form $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$ and $\mathbf{G}_F = \dfrac{\partial}{\partial \mathbf{F}} \boldsymbol{g}(\mathbf{F})$ is the 6×25 matrix with elements of $\mathbf{G}_\pi$ replaced by the corresponding entry in terms of the expected frequencies, i.e. $\mathbf{G}_F(1,1) = F_7$ and so on.

The ML estimation procedure

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{V})'(\mathbf{G}_x\mathbf{V}\mathbf{G}_F')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||)$$

may now be used to find the ML estimates of the expected frequencies in the QI model. The matrix $\mathbf{V}$ may be taken as $\mathbf{D}_F$ for Poisson sampling or $\mathbf{D}_F - \frac{1}{n}\mathbf{F}\mathbf{F}'$ for multinomial sampling. A double iteration is required and the procedure converges to the ML estimates. The likelihood ratio statistic, $LR = 9.60$ with $df = 6$. Note that df = number of constraints in the function $\boldsymbol{g}(\mathbf{F})$.

4.3 STRUCTURAL ZEROS

If a contingency table has a cell for which it is theoretically impossible to have an observation, then we have a so called *structural zero*. For such a cell the ML estimate of the frequency must necessarily also be zero. This can be done by imposing a constraint on any cell having a structural zero. These constraints can be written to comply with the conditions of Proposition 1 and estimation for models for the contingency table will be possible. Bishop, Fienberg and Holland (1975) as well as Haberman (1973, 1974a) examine the influence of empty cells on the existence and uniqueness of ML estimates for loglinear models for incomplete tables. The constraint method provides ML estimates for expected frequencies for tables where the ML estimates can or cannot be written in closed form. To illustrate the principle, consider the following hypothetical example.

Example 4.9
Consider the following incomplete table.

We wish to fit the model of independence to the non-zero cells and the cells

TABLE 4.9: FREQUENCY TABLE WITH STRUCTURAL ZEROS

	Variable B		
Variable A	64	70	11
	83	95	0
	0	0	32

with structural zeros must have zero expected frequencies. Thus impose the constraints on the zero frequency cells through

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

In the design matrix $\mathbf{X}$, incorporate a column of ones for the mean along with the columns for the main effects for the variables A and B, and concatenate these columns with the columns of $\mathbf{C}$. Write the model as $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$ and let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then $\mathbf{P}\ln(\mathbf{F}) = \mathbf{PX}\boldsymbol{\beta} = \mathbf{0}$, and we thus have a function $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$. The estimation procedure can once more be applied to obtain the ML estimates of the expected frequencies. The program for the example can be found in the Appendix. The estimation procedure gives the following ML estimates for the expected frequencies for the quasi-independence model.

	Variable B		
	63.1346	70.8654	11.0000
Variable A	83.8654	94.1346	0.0000
	0.0000	0.0000	32.0000

For this model $LR = 0.0393$ and $\chi^2 = 0.0393$ with $df = 1$.

Note that for this table the ML estimates can be obtained by partitioning the table into the subtable

64	70
83	95

The ML estimates for the frequencies can now be found by using the marginal totals in the usual way. For example

$$\widehat{F}_{11} = \frac{134 \times 147}{312} = 63.1346 \ .$$

Bishop, Fienberg and Holland (1975) also discuss the so-called "Block Stairway Incomplete Table" and give formulae for the ML estimates for the non-zero expected frequencies for the quasi-independence model. They also comment on the fact that the iterative proportional procedure does not always converge exactly to the ML estimates for certain incomplete tables. The constraint method, however, does give the exact ML estimates as will be illustrated by the following example taken from BFH.

Example 4.10

Consider the following example, the details of the experiment can be found in BFH p.200.

The loglinear model for the quasi-independence for the incomplete table can be written as

$$\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$$

TABLE 4.10: BLOCK STAIRWAY TABLE FORM

Female Type	Male Type			
	$A'B$	$A'B'$	AB'	AB
Female with a Y-Chromosome	—	1029	2240	1413
Female with Proximal segment of the Translocation	346	548	1287	—

where

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & -1 & -1 & 0 & 1 \end{bmatrix}$$

The last two columns of $\mathbf{X}$ are the constraints necessary to impose zero expected frequencies for the two cells having structural zeros. Using the same technique as in Example 4.8, we have a function $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$, and the ML estimates for the expected frequencies can be found. These are given in the following table (segment)

0.0000	1010.0339	2258.9661	1413.0000
346.0000	566.9661	1268.0339	0.0000

These are the exact values that will be obtained by using the closed formula given in BFH equation (5.2-48) on page 199. The likelihood ratio $LR = 1.44$ and $\chi^2 = 1.43$ with $df = 1$.

Quasi-independence for this incomplete table can also be expressed in terms of the only possible odds ratio for the table, namely

$$\theta = \frac{F_{12}F_{23}}{F_{22}F_{13}} = 1 \text{ or } F_{12}F_{23} - F_{22}F_{13} = 0 \; .$$

This is of the form $g(\mathbf{F}) = 0$ and it follows that $\mathbf{G}_F = (0, F_{23}, -F_{22}, 0, 0, -F_{13}, F_{12}, 0)$. The ML estimates for the expected frequencies under independence for the incomplete table are given by

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{D}_F)'(\mathbf{G}_x\mathbf{D}_F\mathbf{G}_F)^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||),$$

which will require a double iteration, namely over $\mathbf{x}$ and $\mathbf{F}$.

The programs for this example appear in the Appendix.

The next example will illustrate another problem that needs consideration when dealing with incomplete contingency tables.

Example 4.11
Consider the following example taken from Everitt (1977) p.109. The table presents the cross classification of 291 teenagers according to Age (A), Sex (B) and Health Problems (C).

TABLE 4.11: CROSS-CLASSIFICATION OF HEALTH OF TEENAGERS

		Sex			
		Males		Females	
Age		12-15	16-17	12-15	16-17
Health Problem	S	4	2	9	7
	M	—	—	4	8
	H	42	7	19	10
	N	57	20	71	31

S:	Sex reproduction	H:	How healthy I am
M:	Menstrual problems	N:	Nothing

Since males were naturally not affected by menstrual problems, certain cells are structural zeros.

Consider fitting the model

$$\ln(F_{ijk}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

to the data. The constraints for the two structural zeros are

$$\mathbf{C}_1' = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The design matrix $\mathbf{X}$ for the above mentioned model, written as $\ln(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$, will have 13 columns for the 13 unknown parameters in the model. If we consider $\mathbf{Z} = [\mathbf{X},\ \mathbf{C}_1]$ to impose the structural zero constraints, then $\mathbf{Z}$ will be singular. By applying a singular-value-decomposition to $\mathbf{Z}$, we will find 14 non-zero singular values. Hence the degrees of freedom for the model, $df = 16 - 14 = 2$. Now by using the eigenvectors corresponding to the non-zero singular values as the column vectors of the matrix say, $\mathbf{U}$, we can write the model as $\ln(\mathbf{F}) = \mathbf{U}\boldsymbol{\alpha}$ and by taking $\mathbf{P} = \mathbf{I} - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$, we have $\boldsymbol{g}(\mathbf{F}) = \mathbf{P}\ln(\mathbf{F}) = \mathbf{0}$ and once more the estimation procedure of Proposition 1 will yield the ML estimates of the expected frequencies. The ML estimates for the expected frequencies for the preceding table are as follows.

4.03	1.97	8.97	7.03
0.00	0.00	4.00	8.00
39.81	9.18	21.19	7.81
59.16	17.84	68.84	33.16

The likelihood ratio statistic, $LR = 2.026$ and $\chi^2 = 2.030$ based on $df = 2$.

Chapter 5

LOGIT MODELS AND LOGISTIC REGRESSION

This chapter deals with a cross-classification, where one variable is considered to be a response variable and the other variables are explanatory variables. Logit models are suitable for such data. The standard logit, cumulative logit model, adjacent categories logit model and continuation ratio logit model will be discussed and written in terms of the implied constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$. The proportional hazards model proposed by McCullagh (1980), which is an appropriate model for data where the response variable is exponential, is also discussed. The estimation procedure of Proposition 1 will then be implemented for parameter estimation in these models.

Estimation for models suitable for binary data is discussed and the logistic regression model and extreme value model are presented in terms of the implied constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$.

5.1 THE LOGIT MODEL

Suppose that C is a response variable with k categories and that variables A and B are explanatory variables with r and c categories respectively. Suppose that the last category of variable C is chosen as the reference category, then a standard logit model including main effects for A and B is

$$\ln\left(\frac{F_{ijl}}{F_{ijk}}\right) = \mu + \lambda_i^A + \lambda_j^B, \quad l = 1,\ldots,k-1. \tag{5.1}$$

The standard logit model can be written in the form

$$\mathbf{K}\ \ln(\mathbf{F}) = \mathbf{A}\boldsymbol{\lambda}, \tag{5.2}$$

where $\mathbf{A}$ is the full design matrix, $\boldsymbol{\lambda}$ is the vector of parameters and the matrix $\mathbf{K}$ is appropriately constructed.

The $s = r \times c$ category combinations of A and B form the so called populations. Arrange the cross-classification as follows.

	VARIABLE C				
Population	1	2	$\cdots$	$k-1$	k
1	F_{11}	F_{12}	$\cdots$	$F_{1,k-1}$	F_{1k}
2	F_{21}	F_{22}	$\cdots$	$F_{2,k-1}$	F_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	F_{s1}	F_{s2}	$\cdots$	$F_{s,k-1}$	F_{sk}

Let F_{ij} now denote the frequency in the ith population and the jth category of the response variable C. For the 1st population let $\mathbf{F}_1' = (F_{11}, F_{12}, \ldots, F_{1k})$. The logits for the 1st population are

$$\ln\left(\frac{F_{11}}{F_{1k}}\right),\ \ln\left(\frac{F_{12}}{F_{1k}}\right),\ldots,\ln\left(\frac{F_{1,k-1}}{F_{1k}}\right).$$

Let $\mathbf{L}_1$ be the vector of these logits, then

$$\mathbf{L}_1 = \mathbf{C}\ \ln(\mathbf{F}_1),$$

where

$$\mathbf{C}_{(k-1)\times k} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} . \tag{5.3}$$

Repeat for the other populations and let

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \vdots \\ \mathbf{L}_s \end{pmatrix},$$

$$\mathbf{K} = \mathbf{C} \otimes \mathbf{I_s},$$

$$\mathbf{F}' = (F_{11}, F_{12}, \ldots, F_{1k}, F_{s1}, \ldots, F_{sk}).$$

Then all the logits can be written as the vector

$$\mathbf{L} = \mathbf{K}\ \ln(\mathbf{F})\ . \tag{5.4}$$

In order to set certain parameters equal to zero, write the model in the form

$$\mathbf{H}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{K}\ln(\mathbf{F}) = \mathbf{H}\boldsymbol{\lambda} = \mathbf{0}\ ,$$

where $\mathbf{H}$ specifies the functions of $\boldsymbol{\lambda}$ set to zero. Let $\mathbf{A}_H^{'} = \mathbf{H}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ as in Chapter 3. Then

$$\boldsymbol{g}(\mathbf{F}) = \mathbf{A}_H^{'}\ \mathbf{K}\ \ln(\mathbf{F}) = \mathbf{0}. \tag{5.5}$$

From (5.5) $\mathbf{G}_F = \dfrac{\partial}{\partial \mathbf{F}}\, \boldsymbol{g}(\mathbf{F}) = \mathbf{A}_H^{'}\ \mathbf{K}\ \mathbf{D}_F^{-1}$.

Use $\mathbf{V} = \mathbf{D}_F$ for Poisson sampling and the estimation procedure becomes

$$\begin{aligned} \widehat{\mathbf{F}}_c &= \mathbf{x} - (\mathbf{G}_F\mathbf{D}_F)'(\mathbf{G}_x\mathbf{D}_F\mathbf{G}_F^{'})^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x}-\mathbf{F}||) \\ &= \mathbf{x} - (\mathbf{K}'\mathbf{A}_H)(\mathbf{A}_H^{'}\ \mathbf{K}\mathbf{D}_x^{-1}\mathbf{K}'\mathbf{A}_H)^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x}-\mathbf{F}||) \end{aligned} \tag{5.6}$$

where iteration takes place over $\mathbf{x}$.

5.2 THE CUMULATIVE LOGIT MODEL

Suppose that the response variable is ordinal with say k categories. Models utilizing the ordinal nature of the response variable are constructed by defining alternative logits. Consider explanatory variables A and B with r and c levels respectively and response variable C with k levels.
Let $\pi_1, \pi_2, \ldots, \pi_k$ be the response probabilities at a certain combination of levels of explanatory variables.

The cumulative logit at cut point j is defined as:

$$L_j = \ln\left[\frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_k}\right] , \quad j = 1, \ldots, k-1. \tag{5.7}$$

Agresti (1984) suggests a number of cumulative logit models some of which are given below. McCullagh (1980) also discusses a number of cumulative link models and presents an iterative routine for calculating the ML estimates of the parameters. The routine expresses the response probabilities in the likelihood function in terms of cumulative probabilities, and applies a Fisher scoring algorithm.

5.2.1 Cumulative Logit Models For Two-way Tables

(i) Ordinal-Ordinal Tables

Suppose that the row and column variables are ordinal. Let j be a fixed cut point for the cumulative logit. The jth cumulative logit in row i is

$$L_{j(i)} = \ln\left[\frac{F_{i1} + \cdots + F_{ij}}{F_{i,j+1} + \cdots + F_{ik}}\right] , \quad i = 1, \ldots, r \, . \tag{5.8}$$

Suppose we assign scores $\{u_i\}$ to the row variable. A linear model for the j-th cumulative logit values $\{L_{j(1)}, L_{j(2)}, \ldots, L_{j(r)}\}$ is

$$L_{j(i)} = \mu_j + \beta_j(u_i - \bar{u}) \, , \quad i = 1, \ldots, r. \tag{5.9}$$

This model holds for each of the cut points j, where $j = 1, 2, \ldots, k-1$. The model in (5.9) is in fact the usual logit model, where for cut point j, an $r \times 2$ table is formed by combining the first j categories to form the first category and combining the last $c-j$ categories to form the second category.

If in (5.9) $\beta_1 = \beta_2 = \cdots = \beta_{k-1} = \beta$, then this model simplifies to

$$L_{j(i)} = \mu_j + \beta(u_i - \bar{u}) \ , \quad 1 \leq i \leq r \ , \quad 1 \leq j \leq k-1. \qquad (5.10)$$

(ii) Ordinal-Nominal Tables

Suppose that A is nominal and B is ordinal. A cumulative logit model having row effects, each of which is identical for the $k-1$ ways of forming the logits is

$$L_{j(i)} = \mu_j + \tau_i \ , \quad 1 \leq i \leq r \ , \quad 1 \leq j \leq k-1 \ , \qquad (5.11)$$

$$\text{where} \qquad \sum_i \tau_i = 0 \ .$$

Note that the ith row effect τ_i is the same for all $k-1$ ways of forming the logits, i.e. τ_i and not τ_{ij}.

5.2.2 Cumulative Logit Models For Higher Dimensions

Let A and B be explanatory variables and C an ordinal response variable with k categories. Suppose that there are $r \times c$ subpopulations formed by the category combinations of X and Y. At levels i of X and j of Y and for cut point ℓ,

$$L_{\ell(ij)} = \ln\left[\frac{F_{ij1} + \cdots + F_{ij\ell}}{F_{ij,\ell+1} + \cdots + F_{ijk}}\right] \ , \quad \ell = 1, \ldots k-1 \ . \qquad (5.12)$$

Consider the following examples of cumulative logit models.

(a) If A and B are nominal, a possible cumulative logit model is

$$L_{\ell(ij)} = \mu_\ell + \tau_i^A + \delta_j^B \ , \ \text{where} \ \sum_i \tau_i^A = \sum_j \delta_j^B = 0. \qquad (5.13)$$

(b) If A and B are ordinal, a possible cumulative logit model is

$$L_{\ell(ij)} = \mu_\ell + \beta^A(u_i - \bar{u}) + \beta^B(v_j - \bar{v}). \tag{5.14}$$

(c) If A is nominal and B is ordinal, a possible cumulative logit model is

$$L_{\ell(ij)} = \mu_\ell + \tau_i^A + \beta^B(v_j - \bar{v}). \tag{5.15}$$

The above-mentioned cumulative logit models have homogeneous row effects. It is possible to start with non-homogeneous effects and then establish whether homogeneous effects are valid.

All of the above-mentioned cumulative logit models can be written in the form

$$\ln(\mathbf{C}_1\mathbf{F}) - \ln(\mathbf{C}_2\mathbf{F}) = \mathbf{X}\boldsymbol{\beta} \tag{5.16}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are appropriately constructed, $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is the vector of unknown parameters. The design matrix $\mathbf{X}$ has main effects, interactions and any covariates to be considered. In order to apply the estimation procedure, we require a function $\boldsymbol{g}(\mathbf{F}) = \mathbf{0}$. To this end let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and perform a singular-value-decomposition if there are column dependencies. Now let $\mathbf{A}_H = [\mathbf{P}\ ,\ \mathbf{X}\mathbf{G}_1']$, where $\mathbf{X}\mathbf{G}_1'$ can be used to impose further constraints on the parameters in the logit model. For example, one might have a homogeneous model for the main effects of A. Then $\mathbf{X}\mathbf{G}_1'$ will be used to equate the parameters for each logit. This aspect of the model formulation allows a great amount of flexibility in the type of model that is constructed. This area can be an advantage over the established packages such as SAS and GLIM, where it may be necessary to use macros to cater for more diverse models. The function $\boldsymbol{g}(\mathbf{F}) = \mathbf{A}_H'[\ln(\mathbf{C}_1\mathbf{F}) - \ln(\mathbf{C}_2\mathbf{F})]$ is the function of implied constraints for the model. Furthermore

$$\mathbf{G}_F = \frac{\partial \boldsymbol{g}(\mathbf{F})}{\partial \mathbf{F}} = \mathbf{A}_H'[\mathbf{D}_{\mathbf{C}_1\mathbf{F}}^{-1}\mathbf{C}_1 - \mathbf{D}_{\mathbf{C}_2\mathbf{F}}^{-1}\mathbf{C}_2]$$

where $\mathbf{D}_{\mathbf{C}_1\mathbf{F}} = \text{diag}(\mathbf{C}_1\mathbf{F})$ and $\mathbf{D}_{\mathbf{C}_2\mathbf{F}} = \text{diag}(\mathbf{C}_2\mathbf{F})$ and the estimation procedure is

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F\mathbf{V})'(\mathbf{G}_x\mathbf{V}\mathbf{G}_F')^{-1}\boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \tag{5.17}$$

where $\mathbf{V}$ is the variance-covariance matrix for product multinomial sampling. $\mathbf{D}_F$ may be used in place of $\mathbf{V}$, since for a contingency table with say s populations and c response categories, the likelihood function may be written as

$$l = \prod_{i=1}^{s} \left[\frac{n_i!}{\prod_{j=1}^{c} x_{ij}!} \pi_{i1}^{x_{i1}} \pi_{i2}^{x_{i2}} \cdots \pi_{ic}^{x_{ic}} \right]$$

and the canonical parameter

$$\begin{aligned} \theta_{ij} &= \ln \pi_{ij} \quad \text{(see Example 2.2)} \\ &= \ln \left(\frac{\mu_{ij}}{n_i} \right) , \end{aligned}$$

where $\mu_{ij} = n_i \pi_{ij}$.

Thus

$$\frac{\partial \theta_{ij}}{\partial \mu_{ij}} = \frac{1}{\mu_{ij}}$$

and

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = \mathbf{D}_\mu \, .$$

Hence, from Proposition 1

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{x} - (\mathbf{G}_\mu \mathbf{D}_\mu)'(\mathbf{G}_x \mathbf{D}_\mu \mathbf{G}_\mu')^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \boldsymbol{\mu}||)$$

or in terms of the frequencies

$$\widehat{\mathbf{F}}_c = \mathbf{x} - (\mathbf{G}_F \mathbf{D}_F)'(\mathbf{G}_x \mathbf{D}_F \mathbf{G}_F')^{-1} \boldsymbol{g}(\mathbf{x}) + o(||\mathbf{x} - \mathbf{F}||) \qquad (5.18)$$

where iteration takes place over $\mathbf{x}$ and $\mathbf{F}$.

Example 5.1

Consider the data of Example 3.7. The opinion variable may be considered as an ordinal response variable with 4 categories. There are 2×3 so called populations, i.e.

Afrikaans	Age	18-24
Afrikaans	Age	25-34
Afrikaans	Age	35-64
English	Age	18-24
English	Age	25-34
English	Age	35-64

Consider the contingency table as an $r \times c$ array where F_{ij} is for the ith population and jth category of the response variable.

For the ith population, the cumulative logits are

$$\ln\left(\frac{F_{11}}{F_{12}+F_{13}+F_{14}}\right), \ln\left(\frac{F_{11}+F_{12}}{F_{13}+F_{14}}\right), \ln\left(\frac{F_{11}+F_{12}+F_{13}}{F_{14}}\right).$$

Arrange these 3 expressions in a vector, say $\mathbf{L}_1$, then $\mathbf{L}_1 = \ln(\mathbf{C}_1\mathbf{F}_1) - \ln(\mathbf{C}_2\mathbf{F}_1)$, where

$$\mathbf{C}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}, \quad \mathbf{C}_2 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.19}$$

and $\mathbf{F}_1' = (F_{11}, F_{12}, F_{13}, F_{14})$.

Repeating the principle for the 6 populations, the $3 \times 6 = 18$ logits can be jointly written as

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \vdots \\ \mathbf{L}_6 \end{pmatrix} = \ln(\mathbf{K}_1\mathbf{F}) - \ln(\mathbf{K}_2\mathbf{F})$$

where $\mathbf{K}_1 = \mathbf{C}_1 \otimes \mathbf{I}_6$ and $\mathbf{K}_2 = \mathbf{C}_2 \otimes \mathbf{I}_6$. The Kronecker product is defined as in Searle (1982) equation (27), section 10.7, and

$$\mathbf{F}' = (F_{11}, F_{12}, \ldots, F_{14}, F_{21}, \ldots, F_{24}, \ldots, F_{64}).$$

The cumulative logit vector **L** can now be written in the model form

$$\mathbf{L} = \ln(\mathbf{K}_1\mathbf{F}) - \ln(\mathbf{K}_2\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}$$

as mentioned in (5.16).

Example 5.1 (continued) :
In a survey on socio-political change in South Africa of October 1987 people were asked their opinion as to how effectively the unrest situation had been handled. The results are summarized in the following contingency table.

TABLE 5.1: OBSERVED FREQUENCIES AND EXPECTED FREQUENCIES FOR CUMULATIVE LOGIT MODEL FOR OPINION ON THE UNREST SITUATION

		Opinion			
Language	Age	Ineffective	Neither effective nor ineffective	Fairly effective	Very effective
Afrikaans	18-24	5 (5.33)	6 (8.25)	53 (53.25)	16 (13.17)
	25-34	7 (7.95)	15 (12.90)	115 (106.43)	25 (34.72)
	35-64	13 (11.72)	20 (19.73)	197 (205.21)	96 (89.35)
English and other	18-24	7 (5.49)	3 (4.58)	33 (32.48)	7 (7.45)
	25-34	8 (8.60)	9 (7.59)	65 (68.33)	23 (20.48)
	35-64	13 (13.98)	13 (12.91)	149 (146.11)	56 (58.00)

(Figures in parentheses are the expected frequencies for the cumulative logit model $L_{\ell(ij)} = \mu_\ell + \tau_{i\ell}^L + \beta^A(v_j - \bar{v})$, $\ell = 1,2,3$).

Consider "opinion" as the response variable, with "language" a nominal explanatory variable and "age" an ordinal explanatory variable. The non-homogeneous model for the data is

$$L_{\ell(ij)} = \mu_\ell + \tau_{i\ell}^L + \beta_\ell^A(v_j - \bar{v}) \ , \quad \ell = 1,2,3.$$

When this model is fitted to the data, it is found that the $\widehat{\beta}_k^A$ are similar in magnitude ($\widehat{\beta}_1^A = -0.3656, \widehat{\beta}_2^A = -0.2747$ and $\widehat{\beta}_3^A = -0.3555$), hence the following model is fitted to the data

$$L_{\ell(ij)} = \mu_\ell + \tau_{i\ell}^L + \beta^A (v_j - \bar{v}) \ , \quad \ell = 1, 2, 3.$$

The parameter estimates and their standard normal values are given in the following table.

parameter	estimate	z-value
μ_1	-2.6908	-18.59
τ_{11}^L	-0.2737	-1.93
β^A	-0.3249	-3.29
μ_2	-1.8072	-17.74
τ_{12}^L	-0.1052	-1.08
μ_3	1.3584	14.59
τ_{13}^L	-0.0594	-0.75

The cumulative logit model fits the data well. The likelihood ratio statistic, $LR = 8.35$ with $df = 11$.

Discussion of the cumulative logit output.

The cumulative logit model fits the data well. The likelihood ratio statistic, $LR = 8.35$ with $df = 11$, $AIC^* = LR + 2p = 8.35 + 2(7) = 22.35$.
For cut point $\ell = 1$, $\widehat{\mu}_1 = -2.6908$, which being negative indicates an overall tendency that fewer people favour the opinion "ineffective" rather than the other three categories combined. For the Afrikaans speaking people aged 18-24 years, the odds of having the opinion "ineffective" rather than the other three categories combined, is 0.0714, while the above-mentioned odds for the English speaking people aged 18-24 years is 0.1234. The odds for the Afrikaans speaking people aged 35-64 years is 0.0373, while the odds for the English speaking people is 0.0644.

Consider the cut point $\ell = 3$. For the Afrikaans speaking people aged 18-24 years the odds of having the opinion "ineffective or neither effective nor ineffective or fairly effective" rather than "very effective" is 5.0725. The above-mentioned odds for Afrikaans speaking people aged 35-64 years is 2.6487, or alternatively the odds of "very effective" rather than the other three categories combined are $1/5.0725 = 0.1971$ and $1/2.6487 = 0.3775$. Similarly the odds of "very effective" rather than the other three categories combined, for the English speaking people aged 18-24 years and 35-64 years are $1/5.7128 = 0.1750$ and $1/2.9830 = 0.3352$. Thus showing that the older people have a higher odds of the opinion "very effective" rather than the other three categories combined, than the younger age group.

On the grounds of the AIC-criterion, the above-mentioned model will be preferred to the model suggested in Example 3.7 for which $AIC^* = 29.94$. As mentioned earlier, the model with the smaller AIC^*- value is preferable. The program for fitting the cumulative logit models appears in the Appendix.

5.3 ADJACENT CATEGORIES AND CONTINUATION RATIOS LOGITS

(i) The adjacent categories logit is defined as:

$$L_j = \ln\left[\frac{\pi_{j+1}}{\pi_j}\right] , \quad j = 1, \ldots, k-1. \tag{5.20}$$

(ii) The continuation ratios logit is defined as:

$$L_j = \ln\left[\frac{\pi_{j+1}}{\pi_1 + \cdots + \pi_j}\right] , \quad j = 1, \ldots, k-1. \tag{5.21}$$

or

$$L_j^* = \ln\left[\frac{\pi_j}{\pi_{j+1} + \cdots + \pi_k}\right] , \quad j = 1, \ldots, k-1.$$

These models can also be written in the form of equation (5.16), where $\mathbf{C}_1$ and $\mathbf{C}_2$ are appropriately defined. For Example 5.1, the matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ (for a population) for (i) are:

$$\mathbf{C}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{C}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

and for (ii) the matrices are:

$$\mathbf{C}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{C}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

In order to specify the model over all s populations use $\mathbf{C}_1 \otimes \mathbf{I}_s$ and $\mathbf{C}_2 \otimes \mathbf{I}_s$ as $\mathbf{C}_1$ and $\mathbf{C}_2$ in (5.16) and use (5.18) with the double iteration to find the ML estimates for the expected frequencies for the model.

5.4 PROPORTIONAL HAZARDS MODEL

Suppose the underlying distribution for the response variable Y is exponential, then McCullagh (1980) suggests the response function

$$\ln[-\ln(1 - F_j(\mathbf{x}))]$$

where $\{F_j(\mathbf{x}), j = 1, \ldots, c\}$ denotes the distribution function of Y when vector variable $\mathbf{X}$ takes on the value $\mathbf{x}$. The linear model

$$\ln[-\ln(1 - F_j(\mathbf{x}))] = \beta_{0j} + \beta'\mathbf{x}\,, \quad 1 \le j \le c-1 \tag{5.22}$$

is called the proportional hazards model. The Fisher scoring iterative algorithm is generally used to find the ML estimates of the parameters. The method presented in Proposition 1 also provides a useful and flexible procedure for finding the ML estimates, by writing the model in terms of the implied constraints. This can be done as follows.

Estimation

Let $l_j = \ln[-\ln(1 - F_j(\mathbf{x}))]$. The saturated model for the complementary log-log model may be written as $\boldsymbol{l} = \mathbf{A}\boldsymbol{\lambda}$, where $\mathbf{A}$ is the appropriate design matrix and $\boldsymbol{\lambda}$ is the vector of parameters. The reduced model to be considered may be written as

$$\mathbf{A}'_H\boldsymbol{l} = \mathbf{A}'_H\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}\ .$$

The vector $\boldsymbol{l}$ may be written in the form

$$\boldsymbol{l} = \ln(-\ln(\mathbf{C}_1\boldsymbol{\pi}))$$

where

$$\mathbf{C}_1 = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

and $\mathbf{C}_1\boldsymbol{\pi}$ then gives the survival probabilities. Hence

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\pi}) &= \mathbf{A}'_H \ln(-\ln(\mathbf{C}_1\boldsymbol{\pi})) \\ &= \mathbf{A}'_H \ln(\mathbf{h}(\boldsymbol{\pi})) \end{aligned}$$

where $\mathbf{h}(\boldsymbol{\pi}) = -\ln(\mathbf{C}_1\boldsymbol{\pi})$.

Now

$$\begin{aligned} \mathbf{G}_\pi &= \frac{\partial}{\partial\boldsymbol{\pi}}\boldsymbol{g}(\boldsymbol{\pi}) \\ &= \frac{\partial}{\partial\boldsymbol{\pi}}\mathbf{A}'_H \ln(\mathbf{h}(\boldsymbol{\pi})) \\ &= \mathbf{A}'_H\mathbf{D}^{-1}_{\mathbf{h}(\boldsymbol{\pi})}\left[\frac{\partial}{\partial\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi})\right] \\ &= \mathbf{A}'_H\mathbf{D}^{-1}_{\mathbf{h}(\boldsymbol{\pi})}\left[-\frac{\partial}{\partial\boldsymbol{\pi}}\ln(\mathbf{C}_1\boldsymbol{\pi})\right] \\ &= -\mathbf{A}'_H\mathbf{D}^{-1}_{\mathbf{h}(\boldsymbol{\pi})}[\mathbf{D}^{-1}_{\mathbf{C}_1\boldsymbol{\pi}}\cdot\mathbf{C}_1] \end{aligned}$$

where $\mathbf{D}_{\mathbf{h}(\boldsymbol{\pi})}$ is a diagonal matrix with the elements of $\mathbf{h}(\boldsymbol{\pi})$ on the principal diagonal and $\mathbf{D}_{\mathbf{C}_1\boldsymbol{\pi}}$ is a diagonal matrix with the elements of $\mathbf{C}_1\boldsymbol{\pi}$ on the principal diagonal.

Let $\boldsymbol{\pi}' = (\boldsymbol{\pi}'_{(1)}, \boldsymbol{\pi}'_{(2)}, \ldots, \boldsymbol{\pi}'_{(s)})$, where $\boldsymbol{\pi}'_{(i)} = (\boldsymbol{\pi}_{(i)1}, \boldsymbol{\pi}_{(i)2}, \ldots, \boldsymbol{\pi}_{(i)c})$ denotes the ith subpopulation probabilities. Then

$$\Sigma_{\boldsymbol{\pi}} = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V}_s \end{pmatrix}$$

where

$$\mathbf{V}_i = \frac{1}{n_i}(\mathbf{D}_{\boldsymbol{\pi}(i)} - \boldsymbol{\pi}_{(i)}\boldsymbol{\pi}'_{(i)}), \quad i = 1, \ldots, s \,.$$

The estimation procedure for the expected cell probabilities is

$$\widehat{\boldsymbol{\pi}}_c = \mathbf{p} - (\mathbf{G}_\pi \Sigma_{\boldsymbol{\pi}})'(\mathbf{G}_p \Sigma_{\boldsymbol{\pi}} \mathbf{G}'_\pi)^{-1}\boldsymbol{g}(\mathbf{p}) + o(||\mathbf{p} - \boldsymbol{\pi}||) \,.$$

A double iteration is necessary namely over $\mathbf{p}$ and $\boldsymbol{\pi}$.

Example 5.2

The following data from the HSRC show the distribution of income of white male graduates in S.A. Note: Income in thousands of 1981 Rands.

Consider the model

$$\ln[-\ln(1 - F_{ijk})] = \mu_k + \lambda_i^Y + \lambda_j^S$$

where the subscripts i, j and k refer to the year, work sector and income category respectively, and F_{ijk} is the cumulative relative frequency in subgroup (i, j) earning an amount less than or equal to the amount specified as upper limit of income category k. For example, if F^*_{ijk} denotes the sample analogue for F_{ijk}, then

$$F^*_{213} = \frac{5397 + 2333 + 3505}{5397 + 2333 + 3505 + 1506 + 645}$$

If we fit the model to the data, the ML estimates are as follows:

TABLE 5.2: INCOME DISTRIBUTION FOR WHITE MALE GRADUATES IN SA

YEAR	1975		1981		1987	
SECTOR	Pub.	Priv.	Pub.	Priv.	Pub.	Priv.
INCOME						
0-15	2309	2050	5397	4125	3315	2737
15-18	2138	1360	2333	1626	2044	1263
18-24	4011	2528	3505	3313	3472	2530
24-30	2794	1967	1506	2032	1105	1848
≤ 30	784	2475	645	2264	477	2397

parameter	estimate
μ_1	-1.1430
μ_2	-0.5914
μ_3	0.2113
μ_4	0.7497
λ_1^Y	-0.1812
λ_2^Y	0.1539
λ_1^S	0.2590

For any specific sector j

$$\ln[-\ln(1-\widehat{F}_{2jk})] - \ln[-\ln(1-\widehat{F}_{ijk})]$$

$$= \widehat{\lambda}_2^Y - \widehat{\lambda}_1^Y = 0.1539 - (-0.1812) = 0.3351 \ .$$

Hence

$$\begin{aligned} \ln(1-\widehat{F}_{2jk}) &= \ln(1-\widehat{F}_{1jk})\exp(0.3351) \\ &= 1.3981\ln(1-\widehat{F}_{1jk}) \end{aligned}$$

or

$$(1-\widehat{F}_{2jk}) = (1-\widehat{F}_{1jk})^{1.3981} \ .$$

This means that for example that if 60% of the graduates earned more than a specified amount in 1975, then $(0.60)^{1.3981} = 0.49$ or 49% of the graduates earned more than the same amount in 1981. Similarly

$$(1 - \widehat{F}_{3jk}) = (1 - \widehat{F}_{2jk})^{0.8811} .$$

Thus, if for example, 60% of the graduates earned more than a specified amount in 1981, then $(0.60)^{0.8811} = 0.64$ or 64% of the graduates earned more than the same amount in 1987.

For any specific year i

$$\ln[-\ln(1 - \widehat{F}_{i2k})] - \ln[-\ln(1 - \widehat{F}_{i1k})]$$

$$= \widehat{\lambda}_2^S - \widehat{\lambda}_1^S = -0.2590 - (0.2590) = -0.5180 .$$

Hence

$$\begin{aligned} \ln(1 - \widehat{F}_{i2k}) &= \ln(1 - \widehat{F}_{i1k}) \exp(-0.5180) \\ &= 0.5957 \ln(1 - \widehat{F}_{i1k}) \end{aligned}$$

or

$$(1 - \widehat{F}_{i2k}) = (1 - \widehat{F}_{i1k})^{0.5957} .$$

Thus, if for example, 60% of the graduates in the public sector earned more than a specified amount in any year then $(0.60)^{0.5957} = 0.74$ or 74% of the graduates in the private sector earned more than the same amount. (The program for the proportional hazards model may be found in the Appendix).

5.5 MODELS FOR A BINARY RESPONSE VARIABLE

5.5.1 The Logistic Regression Model

Let $Y_i, i = 1, \ldots, k$ be independent random variables with Y_i binomially distributed with parameters n_i and π_i, i.e. $Y_i \sim b(n_i, \pi_i)$. We may observe the following frequency distribution for the k independent binomial distributions:

	Subgroups			
	1	2	$\cdots$	k
Successes	y_1	y_2	$\cdots$	y_k
Failures	$n_1 - y_1$	$n_2 - y_2$	$\cdots$	$n_k - y_k$

Suppose that the covariates $X_1, X_2, \ldots, X_m$ are observed and that at occasion i, $\mathbf{x}'_i = (x_{1i}, x_{2i}, \ldots, x_{mi})$ and y_i is the number of successes in the n_i trials, $i = 1, \ldots, k$.

The logistic regression model is written as

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} \tag{5.23}$$

$$= (1, \mathbf{x}'_i)\boldsymbol{\beta}\ , \quad \text{where } \boldsymbol{\beta}' = (\beta_0, \ldots, \beta_m)$$

$$\text{or} \quad \pi_i = \frac{e^{(1,\mathbf{x}'_i)\boldsymbol{\beta}}}{1 + e^{(1,\mathbf{x}'_i)\boldsymbol{\beta}}}\ . \tag{5.24}$$

For the binomial variables Y_i, $i = 1, \ldots, k$,

$$E(Y_i) = n_i \pi_i \ \text{ and } \ \text{var}(Y_i) = n_i \pi_i (1 - \pi_i).$$

The likelihood function for $Y_1, Y_2, \ldots, Y_k$ is

$$l(\mathbf{y}; \boldsymbol{\pi}) = \prod_{i=1}^{k} P(Y_i = y_i) = \prod_{i=1}^{k} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}\ .$$

Thus

$$\ln l(\mathbf{y}; \boldsymbol{\pi}) = \ln \prod_{i=1}^{k} \binom{n_i}{y_i} + \sum_{i=1}^{k} y_i \ln \pi_i + \sum_{i=1}^{k} (n_i - y_i) \ln(1 - \pi_i)$$

$$= \ln \sum_{i=1}^{k} \binom{n_i}{y_i} + \sum_{i=1}^{k} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{k} n_i \ln(1 - \pi_i)$$

Hence the canonical parameter is $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, $i = 1, \ldots, k$ or $\theta_i = \ln\left(\frac{\mu_i}{n_i - \mu_i}\right)$, where $\mu_i = n_i\pi_i$. Thus

$$\begin{aligned}\frac{\partial\theta_i}{\partial\mu_i} &= \frac{n_i}{\mu_i(n_i-\mu_i)} \\ &= \frac{1}{\mu_i\left(1-\frac{\mu_i}{n_i}\right)} \\ &= \frac{1}{n_i\pi_i(1-\pi_i)} \qquad (5.25)\end{aligned}$$

and

$$\frac{\partial\mu_i}{\partial\theta_i} = n_i\pi_i(1-\pi_i)\ .$$

Thus

$$\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}} = \text{diag}[n_i\pi_i(1-\pi_i)] = \mathbf{V}_\mu\ . \qquad (5.26)$$

The logistic regression model in (5.23) can be written as

$$\boldsymbol{l} = \mathbf{X}\boldsymbol{\beta} \qquad (5.27)$$

where

$$l_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)\ ,\quad \boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_m)$$

and

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{m1} \\ 1 & x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1k} & x_{2k} & \cdots & x_{mk} \end{bmatrix}\ .$$

Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\mathbf{P}\boldsymbol{l} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. Hence $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\boldsymbol{l} = \mathbf{0}$, since $l_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \ln\left(\frac{\mu_i}{n_i-\mu_i}\right)$ is a function of μ_i.

Furthermore

$$\frac{\partial l_i}{\partial\mu_i} = \frac{1}{n_i\pi_i(1-\pi_i)}\ .$$

Thus

$$\mathbf{G}_\mu = \frac{\partial \boldsymbol{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \mathbf{P}[\text{diag}(n_i\pi_i(1-\pi_i))]^{-1} = \mathbf{P}\mathbf{V}_\mu^{-1} .$$

From Proposition 1, the ML estimate for $\boldsymbol{\mu}$ subject to the constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ is

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}_\mu \mathbf{V}_\mu)'(\mathbf{G}_y \mathbf{V}_\mu \mathbf{G}_\mu')^{-1}\boldsymbol{g}(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||) \\
&= \mathbf{y} - (\mathbf{P}\mathbf{V}_\mu^{-1}\mathbf{V}_\mu)'(\mathbf{P}\mathbf{V}_y^{-1}\mathbf{V}_\mu\mathbf{V}_\mu^{-1}\mathbf{P}')^{-1}\mathbf{P}\boldsymbol{l}_y + o(||\mathbf{y} - \boldsymbol{\mu}||) \\
&= \mathbf{y} - \mathbf{P}(\mathbf{P}\mathbf{V}_y^{-1}\mathbf{P})^{-1}\mathbf{P}\boldsymbol{l}_y + o(||\mathbf{y} - \boldsymbol{\mu}||) \qquad (5.28)
\end{aligned}$$

where $\boldsymbol{l}_y' = (l_{1,y}, \ldots, l_{k,y})$ and $l_{i,y} = \ln\left(\frac{y_i}{n_i - y_i}\right)$, $i = 1, \ldots, k$ and iteration takes place over $\mathbf{y}$.

The estimated parameters $\widehat{\boldsymbol{\beta}}$ are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{l}^{(c)}, \qquad (5.29)$$

where $\boldsymbol{l}^{(c)}$ is the value of the vector of logits on convergence.

The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\boldsymbol{l}^{(c)}, \boldsymbol{l}^{(c)'})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} . \qquad (5.30)$$

Comment:

When $y_i = 0$, we record a frequency 1×10^{-8} or some small value. If any $y_i = n_i$, then enter the appropriate n_i as $n_i + (1 \times 10^{-8})$. This will eliminate the problem of a zero in the sample logit

$$\ln\left(\frac{y_i/n_i}{1 - y_i/n_i}\right) .$$

Example 5.3

To illustrate the ML estimation procedure in logistic regression in the case where some frequencies are zero, consider the data from Cox and Snell (1989). The data summarize a two-factor 5×4 industrial investigation in which the number, y_i, of ingots not ready for rolling out of n_i tested is shown for combinations of heating time, x_1, and soaking time, x_2.

TABLE 5.3: NUMBER, y_i, OF INGOTS NOT READY ROLLING OUT OF n_i TESTED

(Pairs $(y_i; n_i)$ are recorded in the cells)

Soaking time (x_2)	Heating time (x_1)			
	7	14	27	51
1.0	(0;10)	(0;31)	(1;56)	(3;13)
1.7	(0;17)	(0;43)	(4;44)	(0;1)
2.2	(0;7)	(2;33)	(0;21)	(0;1)
2.8	(0;12)	(0;31)	(1;22)	(0;0)
4.0	(0;9)	(0;19)	(1;16)	(0;1)

The logistic regression model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ , \quad i = 1, \ldots, 19 \ .$$

Using the estimation procedure of (5.28) gives the ML estimates

$$\widehat{\beta}_0 = -5.559165 \ , \ \widehat{\beta}_1 = 0.082031 \ , \ \text{and} \widehat{\beta}_2 = 0.056772 \ .$$

The deviance $D = 13.75$. These estimates are identical to the values obtained by PROC LOGISTIC of SAS/STAT (1990).

The IML program for this example can be found in the Appendix.

5.5.2 The Extreme Value Model

Another model which is suitable for modelling binomial data is the model

$$\pi_i = 1 - \exp[-\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})] \qquad (5.31)$$

or

$$\ln[-\ln(1 - \pi_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} \ . \qquad (5.32)$$

The link function for this distribution is called the complementary log-log link.

Suppose that there is only one explanatory variable x, which denotes the dosage of a toxic chemical. Let the random variable $Y_i = 1$ if the ith subject dies. Let T be the tolerance of a subject for the dosage and let $Y_i = 1$ when $T \leq x$. Let $G(t) = P(T \leq t)$ denote the cdf for T. For fixed dosage x

$$P(Y_i = 1) = \pi_i(x) = P(T \leq x) = G(x) \ .$$

If the model

$$\pi(x) = \exp[-\exp(\beta_0 + \beta_1 x)]$$

is considered, then using the log-log link

$$\ln[-\ln(\pi(x))] = \beta_0 + \beta_1 x \ .$$

This model is also called the extreme value model because the cdf of a random variable X having the extreme value distribution has the cdf given by

$$G(x) = \exp[-\exp(x - a)/b] \ \ , b > 0, -\infty < a < \infty \ .$$

Note that when the complementary log-log model holds for the probability of a success, the log-log model holds for the probability of a failure.

The ML estimates for the parameters in the model given in (5.32) can be found using the estimation procedure of Proposition 1. Letting $l_i = \ln[-\ln(1 - \pi_i)]$, we can write the model in (5.32) as

$$\boldsymbol{l} = \mathbf{X}\boldsymbol{\beta} \ . \qquad (5.33)$$

Now $l_i = \ln\left[-\ln(1-\frac{\mu_i}{n_i})\right]$, where $\mu_i = n_i\pi_i$. Thus

$$\begin{aligned}\frac{\partial l_i}{\partial \mu_i} &= \left[\frac{1}{-\ln\left(1-\frac{\mu_i}{n_i}\right)}\right]\left[\frac{-1}{\left(1-\frac{\mu_i}{n_i}\right)n_i}\right] \\ &= \frac{1}{(\mu_i - n_i)\ln\left(1-\frac{\mu_i}{n_i}\right)} \\ &= d_i \ . \end{aligned} \tag{5.34}$$

Furthermore, as for the logistic regression model, the implied constraints are

$$\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\boldsymbol{l} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \tag{5.35}$$

and

$$\mathbf{G}_\mu = \frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\mathbf{D}_d$$

where $\mathbf{D}_d = \text{diag}(d_i)$. The estimation procedure then becomes

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - (\mathbf{G}_\mu\mathbf{V}_\mu)'(\mathbf{G}_y\mathbf{V}_\mu\mathbf{G}_\mu')^{-1}\boldsymbol{l}_y + o(||\mathbf{y}-\boldsymbol{\mu}||) \tag{5.36}$$

where $\boldsymbol{l}_y' = (l_{1,y}, \ldots, l_{k,y})$ and $l_{i,y} = \ln\left[-\ln\left(1-\frac{y_i}{n_i}\right)\right]$, $\mathbf{V}_\mu$ is defined in (5.26) and iteration takes place over $\mathbf{y}$ and $\boldsymbol{\mu}$.

Example 5.4

Consider the data in Table 5.4, taken from Bliss (1935) and used by Dobson (1983), where Y is the number of beetles killed after $5h$ exposure to gaseous carbon disulphide at various concentrations.

Firstly fitting the logistic regression model

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \ ,$$

using the estimation procedure discussed earlier, gives the MLE's $\widehat{\beta}_0 = -60.71745$ and $\widehat{\beta}_1 = 34.27033$ for β_0 and β_1, which agree with the values obtained by Dobson.

TABLE 5.4: BEETLE MORTALITY DATA

Dose x_i ($\log_{10} CS_2$ mg 1^{-1})	Number of insects n_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Fitting the complementary log-log model

$$\ln[-\ln(1-\pi_i)] = \beta_0 + \beta_1 x_i$$

to the data, using the estimation procedure given in (5.36) gives the MLE's $\widehat{\beta}_0 = -39.5723$ and $\widehat{\beta}_1 = 22.0412$. So if for example the log dosage $x = 1.7$ then the estimated probability of survival is

$$1 - \hat{\pi} = \exp[-\exp(-39.57 + 22.04(1.7))] = 0.885 \ .$$

Similarly the estimated probability of survival at log dosage 1.8 is 0.332. The probability of survival is $\exp(22.04 \times 0.1) = 9.06$ higher for each 0.1 increase in log dosage. The deviance $D = 2\sum_i y_i \ln(y_i/\widehat{y}_i) = 3.45$, with $df = 6$, indicating that the fit for the model is adequate. For the deviance the sum $\sum_i$ is taken over all $(2 \times k)$ cells with frequencies y_i and $n_i - y_i$.

The programs for the logistic regression model and the extreme value distribution can be found in the Appendix.

Chapter 6

MODELS FOR POISSON, NORMAL, GAMMA AND INVERSE GAUSSIAN OBSERVATIONS

This chapter deals with models where the observations have either a Poisson, Normal, gamma or inverse Gaussian distribution. Regression models are formulated in terms of constraints and the ML estimates of the parameters are found using the proposed estimation procedure.

6.1 THE POISSON REGRESSION MODEL

Suppose that Y_i , $i = 1, \ldots, n$ are independent Poisson random variables with $E(Y_i) = \mu_i$, and $\text{var}(Y_i) = \mu_i$, $i = 1, \ldots, n$ and that the model to be fitted is

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \ , \tag{6.1}$$

where $\mathbf{X}$ is the appropriate design matrix and $\boldsymbol{\beta}$ is the vector of parameters. Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\boldsymbol{\mu} = \mathbf{0}$, and

$$\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}.$$

Furthermore $\mathbf{V} = \text{Cov}(\mathbf{y}) = \mathbf{D}_\mu$. Hence the estimation procedure is

$$\hat{\boldsymbol{\mu}}_c = \mathbf{y} - (\mathbf{PD}_\mu)'(\mathbf{PD}_\mu\mathbf{P}')^{-1}\mathbf{Py} \tag{6.2}$$

where iteration is over $\boldsymbol{\mu}$ and the iterative process is the same as that described in Chapter 2 equation (2.6).

Example 6.1

The data from Dobson (1983) are counts y_i observed at various values of a covariate x.

TABLE 6.1: POISSON REGRESSION DATA

y_i	2	3	6	7	8	9	10	12	15
x_i	−1	−1	0	0	0	0	1	1	1

Observing that the variability increases with x, assume that the responses Y_i are Poisson random variables with

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i \ , \quad i = 1, \ldots, 9 \ .$$

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \quad , \quad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ \vdots \\ 15 \end{pmatrix}$$

Dobson uses the iterative reweighted least squares procedure to find the ML estimates of the parameters. Using (6.2) to find the ML estimate $\hat{\boldsymbol{\mu}}_c$, $\text{Cov}(\mathbf{Y}) = \mathbf{D}_\mu$, must be estimated. Take $\boldsymbol{\mu}^{(0)} = \mathbf{y}$ to initialize the diagonal matrix and iterate to get the ML estimates, $\hat{\beta}_0 = 7.4516$ and $\hat{\beta}_1 = 4.9353$, which agree with those obtained by Dobson.

The procedure may also be used when the logarithms of the mean values diplay a linear trend when plotted against the logarithms of a predictor variable and the observations are assumed to have a Poisson distribution.

Example 6.2

Consider the data from Dobson (1990)(Exercise 4.1 pp. 46-47). The number of deaths from AIDS in Australia recorded in 14 consecutive quarters from Jan-Mar 1983, to Apr-Jun 1986 are presented in Table 6.2.

TABLE 6.2: AIDS DATA IN AUSTRALIA

Quarter (t)	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Deaths (Y_t)	0	1	2	3	1	4	9	18	23	31	20	25	37	45

Consider the number of deaths at time t as an independent Poisson variable with mean μ_t. If the logs of both the time variable, t and Y_t are plotted, then a possible linear trend can be observed. Thus a feasible model is

$$\mu_t = \exp(\alpha + \delta \ln(t)) \quad , \quad t = 1, 2, \ldots, 14;$$

$$\text{or} \quad \ln(\mu_t) = \alpha + \delta \ln(t).$$

This model may be written as $\ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$, where

$$\mathbf{X} = \begin{pmatrix} 1 & \ln(1) \\ 1 & \ln(2) \\ \vdots & \vdots \\ 1 & \ln(14) \end{pmatrix} \quad \text{and } \boldsymbol{\beta}' = (\alpha, \delta).$$

The model can be formulated in terms of the implied constraints $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$, where

$$\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\ln(\boldsymbol{\mu}) = \mathbf{P}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' .$$

and

$$\mathbf{G}_\mu = \frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\mathbf{D}_\mu^{-1} .$$

Thus the estimation procedure is

$$\begin{aligned} \hat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}_\mu \mathbf{D}_\mu)'[\mathbf{G}_y \mathbf{D}_\mu \mathbf{G}_\mu']^{-1}\mathbf{P}\ln(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||) \\ &= \mathbf{y} - \mathbf{P}[\mathbf{P}\mathbf{D}_y^{-1}\mathbf{P}]^{-1}\mathbf{P}\ln(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||) \qquad (6.3) \end{aligned}$$

where iteration takes place over $\mathbf{y}$.

Applying this procedure to the data gives the ML estimates $\widehat{\alpha} = -1.9442$ and $\widehat{\delta} = 2.1748$ for the parameters (which agree with those reported by Dobson).

6.2 MODELS FOR NORMAL OBSERVATIONS

We now consider models where the observations are assumed to have a Normal distribution. As a simple application consider a frequency table with k classes for the classification variable as shown in the table.

A_1	A_2	$\cdots$	A_k
y_1	y_2	$\cdots$	y_k

Suppose that Y_i, $i = 1, \ldots, k$ are independent $N(\mu_i; 1)$ variables. We wish to find the ML estimates $\widehat{\mu}_i$ for the model of symmetry $\mu_1 = \mu_k$, $\mu_2 = \mu_{k-1}$, and so on. As an illustration consider the following example.

Example 6.3
Suppose Y_i, $i = 1, \ldots, 6$ are independent $N(\mu_i, 1)$ variables and that the following vector of observations is found

$$\mathbf{y}' = (20,\ 15,\ 9,\ 11,\ 14,\ 18)\ .$$

The model of symmetry $\mu_1 = \mu_6, \mu_2 = \mu_5, \mu_3 = \mu_4$ can be written as

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix} \boldsymbol{\mu} = \mathbf{0}$$

or

$$\mathbf{A}\boldsymbol{\mu} = \mathbf{0}.$$

Once again $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{G} = \dfrac{\partial \boldsymbol{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \mathbf{A}$. In this case $\mathrm{Cov}(\mathbf{y}) = \mathbf{I}_6$

and the estimation procedure is

$$\begin{aligned}\hat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}\mathbf{I}_6)'(\mathbf{G}'\mathbf{I}_6\mathbf{G}')^{-1}\mathbf{G}\mathbf{y} \\ &= \mathbf{y} - \mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}\mathbf{G}\mathbf{y}\ .\end{aligned}$$

Here no iteration is required and the estimate is found directly from the expression above.
The vector of ML estimates is given by

$$\hat{\mathbf{y}} = (19,\ 14.5,\ 10,\ 10,\ 14.5,\ 19)\ ,$$

Another application is a regression problem with certain constraints. Suppose that Y_i , $i = 1, \ldots, n$ are independent Normal random variables with $E(Y_i) = \mu_i$, and $\text{var}(Y_i) = \sigma^2$, $i = 1, \ldots, n$. The variance-covariance of matrix $\mathbf{y}$ is $\text{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}$.

Example 6.4:
Consider a data set where the first 5 points lie (apart from random error) on the first line and that the last 5 points lie on the second line and that the 5th point is thus common to both lines. Suppose that we fit the two line segments to the data with the constraint that the lines must pass through the 5th point. Use the dummy variables X_1 and X_2 as defined in Table 6.3 and fit the the model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, with the constraint that the lines must pass through the point $(5; 10.2)$.

Let $\mathbf{P} = \mathbf{I}_9 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{r}' = (0,0,0,0,1,0,0,0,0)$ and

$$\mathbf{G} = \begin{bmatrix} \mathbf{P} \\ \mathbf{r}' \end{bmatrix}\ .$$

The function required to force the lines through the point $(5; 10.2)$ is

$$\boldsymbol{g}(\mathbf{y}) = \mathbf{G}\mathbf{y} - (0,0,0,0,0,0,0,0,10.2)' = \mathbf{0}\ .$$

TABLE 6.3: REGRESSION FOR NORMAL OBSERVATIONS

obs no.	X_0	X_1	X_2	Y
1	1	-4	0	2.3
2	1	-3	0	3.8
3	1	-2	0	6.5
4	1	-1	0	7.4
5	1	0	0	10.2
6	1	0	1	10.5
7	1	0	2	12.1
8	1	0	3	13.2
9	1	0	4	13.6

The ML estimate for the mean subject to the constraints above, is

$$\widehat{\boldsymbol{\mu}}_c = \mathbf{y} - \mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{y}) \; ,$$

which is obtained without iteration.

The ML estimates are $\widehat{\beta}_0 = 10.20$, $\widehat{\beta}_1 = 2.03$ and $\widehat{\beta}_2 = 0.89$.

Another example is that of fitting a quadratic function subject to the constraint that the funtion must pass through a certain point.

Example 6.5

A study was conducted to examine the relationship between the number of years of experience (X) and the annual salary (Y) for individuals in a particular profession. The following information is obtained for a representative sample of 16 such professionals (salary in thousands of dollars):

X	1	2	4	5	5	9	11	14	16	20	22	24	25	27	29	30
Y	23	27	29	34	38	46	48	54	54	59	58	59	61	63	59	60

When these data are plotted, a curvilinear trend will be observed. The ordinary least squares equation for a quadratic function is

$$\hat{y} = 19.9801 + 3.2152x - 0.0634x^2 \ .$$

If we desire a further constraint that the equation must pass through the point $(16;54)$, then the ML estimates for the parameters can be found using the estimation procedure.
The design matrix for fitting a quadratic function, is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ \vdots & \vdots & \vdots \\ 1 & 30 & 900 \end{bmatrix} .$$

Now let $\mathbf{P} = \mathbf{I}_{16} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\mathbf{r}' = (0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0)$ then

$$\mathbf{G} = \begin{bmatrix} \mathbf{P} \\ \mathbf{r}' \end{bmatrix}$$

and $\boldsymbol{g}(\mathbf{y}) = \mathbf{G}\mathbf{y} - (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,54)'$.

The constraint $\boldsymbol{g}(\mathbf{y}) = \mathbf{0}$, will force the function through the point $(16;54)$. The ML estimate is

$$\hat{\boldsymbol{\mu}}_c = \mathbf{y} - \mathbf{G}'(\mathbf{G}\mathbf{G}')^{-1}\boldsymbol{g}(\mathbf{y}) \ ,$$

which is obtained without iteration.

The ML estimates are $\widehat{\beta}_0 = 20.7487$, $\widehat{\beta}_1 = 2.9633$ and $\widehat{\beta}_2 = -0.0553$.

The following example illustrates the use of the double iteration, when the observations are normal and the variable Y varies non-linearly with X.

Example 6.6

Consider the following data, where X is an explanatory variable, and Y is the response variable.

x	0	0	0	1	1	1	2	2	3	3	3	4	4	5	5	5
y	5	7	9	7	10	8	11	9	16	13	14	25	24	34	32	30

It appears that Y varies non-linearly with X, and that the variance is approximately constant. Suppose further that $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ and that the model to be considered is $\ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. Let $\mathbf{P} = \mathbf{I}_{16} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then the constraints for the model can be written as $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\ln(\boldsymbol{\mu}) = \mathbf{0}$ and

$$\mathbf{G}_{\mu} = \frac{\partial}{\partial \mu}\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\mathbf{D}_{\mu}^{-1} .$$

The estimation procedure is given by

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}_{\mu}\mathbf{V}_{\mu})'(\mathbf{G}_y\mathbf{V}_{\mu}\mathbf{G}_{\mu})^{-1}\boldsymbol{g}(y) + o(||\mathbf{y} - \boldsymbol{\mu}||) \\
&= \mathbf{y} - (\mathbf{P}\mathbf{D}_{\mu}^{-1}\sigma^2\mathbf{I})'(\mathbf{P}\mathbf{D}_{y}^{-1}\sigma^2\mathbf{I}\mathbf{D}_{\mu}^{-1}\mathbf{P}')^{-1}\mathbf{P}\ln(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||) \\
&= \mathbf{y} - (\mathbf{P}\mathbf{D}_{\mu}^{-1})'(\mathbf{P}\mathbf{D}_{y}^{-1}\mathbf{D}_{\mu}^{-1}\mathbf{P}')^{-1}\mathbf{P}\ln(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||) .
\end{aligned}$$

Since $\mathbf{D}_{\mu}$ is unknown, iteration takes place over both $\boldsymbol{\mu}$ and $\mathbf{y}$. Applying the procedure to the data yields the ML estimates $\widehat{\beta}_0 = 1.7214$ and $\widehat{\beta}_1 = 0.3496$ and deviance $D = 52.30$, which agree with the estimates from the procedure Genmod of SAS. The program for this example can be found in the Appendix.

6.3 MODELS WITH GAMMA-DISTRIBUTED OBSERVATIONS

Jorgensen (1992) defines the exponential dispersion model for the univariate continuous case as

$$p(y;\theta,\lambda) = a(\lambda, y)e^{\lambda(y\theta - \kappa(\theta))}, \quad y \in \mathrm{I\!R} \tag{6.5}$$

where $a(\cdot)$ and $\kappa(\cdot)$ are suitable functions and λ and θ are parameters with domain $(\theta,\lambda) \in \Theta \times \Lambda \subseteq \mathrm{I\!R} \times \mathrm{I\!R}_+$. For λ known (6.5) is an exponential family and $\sigma^2 = 1/\lambda$ is called the dispersion parameter.

The p.d.f of the gamma distribution can be written in the form

$$\begin{aligned} p(y;\mu,\nu) &= \frac{\nu^\nu}{\Gamma(\nu)} y^{\nu-1} \exp\left[-\frac{\nu}{\mu} y + \ln\left(\frac{1}{\mu}\right)\right] \\ &= \frac{\nu^\nu}{\Gamma(\nu)} y^{\nu-1} \exp\left[\nu\left(-\frac{y}{\mu}\right) + \nu \ln\left(\frac{1}{\mu}\right)\right]. \end{aligned} \tag{6.6}$$

Thus from (6.5) with $\theta = -\dfrac{1}{\mu}$ and $\lambda = \nu$

$$a(\nu, y) = \frac{\nu^\nu}{\Gamma(\nu)} y^{\nu-1} \quad \text{and} \quad \kappa(\theta) = -\ln(-\theta), \quad \theta < 0.$$

The expectation is given by

$$E(Y) = \frac{\partial}{\partial\theta}\kappa(\theta) = -\left\{\frac{-1}{-\theta}\right\} = -\frac{1}{\theta} = \mu$$

and

$$\operatorname{var}(Y) = \sigma^2 \frac{\partial^2}{\partial\theta^2}\kappa(\theta) = \frac{1}{\nu}\left(\frac{1}{\theta^2}\right) = \frac{\mu^2}{\nu}. \tag{6.7}$$

The coefficient of variation is

$$\frac{\sqrt{\operatorname{var}(Y)}}{E(Y)} = \frac{\mu}{\mu\sqrt{\nu}} = \frac{1}{\sqrt{\nu}}\,. \tag{6.8}$$

If all observations are assumed to come from the same distribution with parameter ν, then the coefficient of variation is constant. See McCullagh and Nelder (1989) Chapter 8 for data with constant coefficient of variation.

If the observations come from different distributions then take $\nu_i = cw_i$, where w_i are some appropriate weights assigned to the different distributions. Then

$$\operatorname{var}(Y_i) = \frac{\mu_i^2}{\nu_i} = \frac{\mu_i^2}{cw_i}\,. \tag{6.9}$$

Estimation of parameters in the model

Suppose $Y_i, i = 1, \ldots, n$ come from a gamma distribution as defined in (6.6). Then $\mu_i = E(Y_i)$.

(i) Let $l_i = \ln(\mu_i)$ and consider covariates $X_1, \ldots, X_m$.

Suppose the model to be considered is

$$\boldsymbol{l} = \mathbf{X}\boldsymbol{\beta} \tag{6.10}$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_m)$ and $\mathbf{X}$ is the design matrix.
Let $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\boldsymbol{l} = \mathbf{0}$ and

$$\mathbf{G}_\mu = \frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\,\mathrm{diag}\left(\frac{1}{\mu_i}\right) = \mathbf{P}\mathbf{D}_\mu^{-1} \tag{6.11}$$

From Proposition 1, the estimation procedure is given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}_\mu \mathbf{V}_\mu)'[\mathbf{G}_y \mathbf{V}_\mu \mathbf{G}_\mu']^{-1}\boldsymbol{g}(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||) \\ &= \mathbf{y} - (\mathbf{P}\mathbf{D}_\mu^{-1}\mathbf{V}_\mu)'[\mathbf{P}\mathbf{D}_y^{-1}\mathbf{V}_\mu(\mathbf{P}\mathbf{D}_\mu^{-1})']^{-1}\mathbf{P}\ln(\mathbf{y}) + o(||\mathbf{y} - \boldsymbol{\mu}||)\end{aligned}$$

where $\mathbf{V}_\mu = \mathrm{diag}\left(\frac{\mu_i^2}{cw_i}\right)$. Since $\mathbf{V}_\mu$ is a function of $\boldsymbol{\mu}$, we iterate over $\boldsymbol{\mu}$ and $\mathbf{y}$.

(ii) Suppose we consider fitting the model

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{6.12}$$

then $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}\boldsymbol{\mu}$ and $\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{P}$.

The estimation procedure becomes

$$\begin{aligned}\hat{\boldsymbol{\mu}}_c &= \mathbf{y} - (\mathbf{G}_\mu \mathbf{V}_\mu)'[\mathbf{G}_y \mathbf{V}_\mu \mathbf{G}_\mu']^{-1}\boldsymbol{g}(\mathbf{y}) \\ &= \mathbf{y} - (\mathbf{P}\mathbf{V}_\mu)'[\mathbf{P}\mathbf{V}_\mu\mathbf{P}']^{-1}\mathbf{P}\mathbf{y}\end{aligned}$$

with $\mathbf{V}_\mu = \mathrm{diag}\left(\frac{\mu_i^2}{w_i}\right)$ and iteration takes place over $\boldsymbol{\mu}$.

(iii) Suppose we consider fitting an inverse linear model, where $\mu_i^* = l_i = \frac{1}{\mu_i}$, $i = 1, \ldots, n$ and $\boldsymbol{\mu}^* = \mathbf{l} = \mathbf{X}\boldsymbol{\beta}$. The constraints are

$$\boldsymbol{g}(\boldsymbol{\mu}^*) = \mathbf{P}\boldsymbol{\mu}^* = \mathbf{0}$$

so that

$$\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\mu}^*}\mathbf{P}\boldsymbol{\mu}^* = \mathbf{P}.$$

The estimation procedure is

$$\hat{\boldsymbol{\mu}}_c = \mathbf{y}^* - (\mathbf{G}\mathbf{V}_*)'(\mathbf{G}\mathbf{V}_*\mathbf{G}')^{-1}\mathbf{P}\mathbf{y}^* \,,$$

where $y_i^* = \dfrac{1}{y_i}$, is the ith element of $\mathbf{y}^*$ and $\mathbf{V}_*$ is the asymptotic covariance matrix of $\mathbf{y}^*$. The ith diagonal element of $\mathbf{V}_*$ is given by

$$\begin{aligned} v_{i*} &= \left(\frac{\partial l_i}{\partial \mu_i}\right)^2 \left(\frac{\mu_i^2}{w_i}\right) \\ &= \left(\frac{-1}{\mu_i^2}\right)^2 \left(\frac{\mu_i^2}{w_i}\right) \\ &= \frac{1}{w_i \mu_i^2} \,. \end{aligned}$$

Iteration takes place over $\boldsymbol{\mu}$.

Example 6.7

Consider the example of McCullagh and Nelder (1989) pp.306-311. The data for the example appear in Table 6.4. The details of the experiment are discussed in the above-mentioned reference on pp 306-309.

The models to be considered are of the form

$$f(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2/(T_i - \delta)$$

where $f(\mu_i)$ can be chosen as $\dfrac{1}{\mu i}$, $\ln(\mu_i)$ or simply μ_i.

The weighting variable is batch size and the values for w_i, $i = 1, \ldots, 23$ are given in the table.

For the log, inverse and identity links δ is given the values $58.644°C$, $33.5°C$ and $0.6°C$, respectively.

TABLE 6.4: MEAN DURATION OF EMBRYONIC PERIOD IN THE DEVELOPMENT OF Drosophila melanogaster

Temp (T_i) °*C*	*Duration* (Y_i) *(hours)*	*Batch* (w_i) *size*
14.95	67.5	54
16.16	57.1	182
16.19	56.0	153
17.15	48.4	129
18.20	41.2	64
19.08	37.80	94
20.07	33.33	82
22.14	26.50	57
23.27	24.24	135
24.09	22.44	188
24.81	21.13	217
24.84	21.05	141
25.06	20.39	37
25.06	20.41	84
25.80	19.45	196
26.92	18.77	104
27.68	17.79	148
28.89	17.38	83
28.96	17.26	95
29.00	17.18	232
30.05	16.81	148
30.80	16.97	195
32.00	18.20	58

By using the procedures described in (i) (ii) and (iii) we fit the respective models and find the following ML estimates for the parameters.

Link	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	Deviance
log	3.2014	−0.2648	−216.998	0.318
inverse	−0.0384	0.0036	0.0333	1.406
identity	−163.2180	2.9892	2664.2152	0.4713

6.4 MODELS WITH INVERSE GAUSSIAN OBSERVATIONS

The p.d.f. of the inverse Gaussian distribution is given by

$$f(y;\mu;\sigma) = \frac{1}{\sqrt{2\pi y^3\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2(\mu\sigma)^2 y}\right), \quad 0 < y < \infty \ .$$

For this distribution $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2\mu^3$. The canonical parameter $\theta = -\dfrac{1}{2\mu^2}$ and the dispersion parameter is σ^2. The appropriate link function is $\dfrac{1}{\mu^2}$ and the model to be considered is

$$\frac{1}{\mu_i^2} = \mathbf{x}_i'\boldsymbol{\beta} \ ,$$

where $\mathbf{x}_i'$ is the ith row of the design matrix and $\boldsymbol{\beta}$ is the vector of unknown parameters.

Let $\gamma_i = \dfrac{1}{\mu_i^2}$. Then $\gamma_i = \mathbf{x}_i'\boldsymbol{\beta}$ and the model can be written in terms of constraints as $\boldsymbol{g}(\boldsymbol{\gamma}) = \mathbf{P}\boldsymbol{\gamma} = \mathbf{0}$ and $\mathbf{G} = \dfrac{\partial \boldsymbol{g}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \mathbf{P}$, where $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

If we let $z_i = \frac{1}{y_i^2}$, then the asymptotic covariance matrix of $\mathbf{Z}$ is $\mathbf{V}_*$ with ith diagonal element

$$\begin{aligned} \mathrm{var}(Z_i) &= \left(\frac{\partial \gamma_i}{\partial \mu_i}\right)^2 \mathrm{var}(Y_i) \\ &= (-2\mu_i^{-3})^2 \mu_i^3 \sigma_i^2 = \frac{4\sigma_i^2}{\mu_i^3} \, . \end{aligned}$$

The estimation procedure is then given by

$$\hat{\boldsymbol{\mu}}_c = \mathbf{z} - (\mathbf{G}\mathbf{V}_*)'(\mathbf{G}\mathbf{V}_*\mathbf{G}')^{-1}\mathbf{P}\mathbf{z} \, .$$

If we assume that $\sigma_i^2 = \sigma^2$ for all i, then σ^2 cancels out in the expression above and since the μ_i are unknown, iteration takes place over $\boldsymbol{\mu}$.

6.5 AN APPROXIMATE SOLUTION FOR MODELS WITH A NON-LINEAR PARAMETER IN THE COVARIATE

If a model has a covariate which includes a non-linear function of a parameter, then the following estimation procedure may be followed. Suppose the covariate is represented by $f(x;\theta)$ with θ unknown, then we use a linear Taylor expression about an initial value θ_0 to approximate $f(x;\theta)$ and write

$$f(x;\theta) \simeq f(x;\theta_0) + (\theta - \theta_0)\left[\frac{\partial f(x;\theta)}{\partial \theta}\right]_{\theta=\theta_0} \, .$$

Thus if a non-linear term in the predictor variable x is given by $\beta f(x;\theta)$, then we write

$$\begin{aligned} \beta f(x;\theta) &\simeq \beta f(x;\theta_0) + \beta(\theta - \theta_0)\left[\frac{\partial f(x;\theta)}{\partial \theta}\right]_{\theta=\theta_0} \\ &= \beta u + \gamma v \\ \text{where } u &= f(x;\theta_0), \ \gamma = \beta(\theta-\theta_0) \text{ and } v = \left[\frac{\partial f(x;\theta)}{\partial \theta}\right]_{\theta=\theta_0} \, . \end{aligned}$$

Fit the model in the usual way and find $\widehat{\gamma}$ and $\widehat{\beta}$. Now

$$\frac{\gamma}{\beta} = \theta - \theta_0 \ .$$

Let $\theta_u = \dfrac{\widehat{\gamma}}{\widehat{\beta}} + \theta_0$ and use the updated value θ_u in the next iteration. Repeat the procedure until convergence is attained. At each stage the design matrix $\mathbf{X}$ must also be appropriately changed depending on the function $f(x;\theta)$.

Example 6.8

Consider the example discussed by McCullagh and Nelder (1989, pp.384-386). The data are for estimation of lowest cost mixtures of insecticides and synergists. They relate to assays on a grasshopper *Melanopus sanguinipes* $(F\cdot)$ with the insecticide carbofuran and the synergist pipeonyl butoxide (PB), which enhances the toxicity of the insecticide.

McCullagh first suggests the model

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \frac{\beta_2 x_2}{\delta + x_2}$$

where x_1 is the log dose of insecticide and x_2 is the dose of the synergist PB. The term $\dfrac{\beta_2 x_2}{\delta + x_2}$ is non-linear in the parameter δ, so we follow the discussion of the previous paragraph. Let

$$f(x_2, \delta) = \frac{\beta_2 x_2}{\delta + x_2} \ .$$

Then

$$\frac{\partial}{\partial \delta} f(x_2, \delta) = \frac{-\beta_2 x_2}{(\delta + x_2)^2} \ .$$

Thus approximate $f(x_2, \delta)$ by

$$f(x_2, \delta_0) + (\delta - \delta_0)\left[\frac{\partial}{\partial \delta} f(x_2, \delta)\right]_{\delta = \delta_0} = \frac{\beta_2 x_2}{\delta_0 + x_2} + \beta_2(\delta - \delta_0)\frac{(-x_2)}{(\delta_0 + x_2)^2} \ ,$$

and fit the model

$$\begin{aligned} \ln\left(\frac{\pi}{1-\pi}\right) &= \beta_0 + \beta_1 x_1 + \beta_2\left(\frac{x_2}{\delta_0 + x_2}\right) + \beta_2(\delta - \delta_0)\frac{(-x_2)}{(\delta_0 + x_2)^2} \\ &= \beta_0 + \beta_1 x_1 + \beta_2\left(\frac{x_2}{\delta_0 + x_2}\right) + \gamma\frac{(-x_2)}{(\delta_0 + x_2)^2} \end{aligned}$$

TABLE 6.5: DATA FROM ASSAY ON INSECTICIDE AND SYNERGIST

Number killed, y	*Sample size,* m	*Dose of insecticide*	*Dose of synergist*
7	100	4	0
59	200	5	0
115	300	8	0
149	300	10	0
178	300	15	0
229	300	20	0
5	100	2	3.9
43	100	5	3.9
76	100	10	3.9
4	100	2	19.5
57	100	5	19.5
83	100	10	19.5
6	100	2	39.0
57	100	5	39.0
84	100	10	39.0

where $\gamma = \beta_2(\delta - \delta_0)$.

This is done by initializing δ_0 and finding the design matrix at the first iteration using the covariates

$$x_1, \frac{x_2}{\delta_0 + x_2}, \frac{-x_2}{(\delta_0 + x_2)^2}.$$

Estimate the parameters in the model in the usual way and use

$$\delta_u = \frac{\widehat{\gamma}}{\widehat{\beta}_2} + \delta_0$$

as the updated value for δ in the next iteration. Re-evaluate the design matrix with the value of δ replaced in the expressions for the covariate. The estimation procedure of Proposition 1 is implemented here, with the model above written in terms of the implied constraints, $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$. In this case the design matrix is changed appropriately at each iteration.

Since the model above does not provide a good fit, McCullagh suggests the following model

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \ln(z-\theta) + \beta_2 \frac{x_2}{(\delta + x_2)}$$

where z is the dose of insecticide. The second and third terms are non-linear in the parameters and are linearized to give the form

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \ln(z-\theta_0) + \gamma_1\left(\frac{-1}{z-\theta_0}\right) + \beta_2 \frac{x_2}{\delta_0 + x_2} + \gamma_2 \frac{(-x_2)}{(\delta_0 + x_2)^2}$$

where θ_0 and δ_0 are the initial values. The covariates for the design matrix are

$$\ln(z-\theta_0), \quad \frac{-1}{z-\theta_0}, \quad \frac{x_2}{\delta_0 + x_2}, \quad \frac{-x_2}{(\delta_0 + x_2)^2} .$$

These are updated at each iteration where

$$\theta_u = \frac{\widehat{\gamma}_1}{\widehat{\beta}_1} + \theta_0$$

$$\text{and } \delta_u = \frac{\widehat{\gamma}_2}{\widehat{\beta}_2} + \delta_0$$

are the expressions for the updated values. Applying the estimation procedure with starting values $\delta_0 = 1.76$ and $\theta_0 = 1.5$ as recommended by McCullagh gives $\widehat{\theta} = 1.67$, $\widehat{\delta} = 2.06$ with a deviance $D = 18.70$ with $df = 15 - 5 = 10$.

The program for this example is given in the Appendix.

6.6 CONCLUSION

This chapter has illustrated the use of the ML estimation procedure for a number of models which belong to the class of models described as generalized linear models. Other examples presented emphasize the usefulness of modelling in terms of constraints. The ML estimation procedure of Proposition 1 can accommodate any choice of model assuming any predetermined distribution of the observations, provided that the covariance matrix of the variables can be computed as in the case of the generalized linear model. Another distinct advantage of the estimation procedure is that the constraints can be non-linear.

The applications of this ML estimation procedure, as presented in this work, are only some of the possible areas where the procedure may be implemented. There are, no doubt, many other applications and these areas will provide opportunities for further research.

REFERENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York: Wiley.

——— (1988). A model for agreement between ratings on an ordinal scale. *Biometrics 44, 539-548.*

——— (1990). *Categorical Data Analysis.* New York: Wiley.

Aitchison, J. and Silvey, S.D. (1958). Maximum-likelihood estimation of parameters subject to constraints. *Ann. Math. Stat.,* ***29****, 813-828.*

Akaike, H. (1973). Information Theory and Extention of the Maximum Likelihood Principle, *2nd Inter. Symp. on Information Theory. (Petrov, B.N. and Csaki, F. eds). Akademia Kiado, Budapest, 267-281.*

Anderson, E.B. (1991). *The Statistical Analysis of Categorical Data.* Springer-Verlag.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Chichester: Wiley.

Bhapkar, V.P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Amer. Statist. Assoc.* ***61****, 228-235.*

Bliss, C.I. (1935). The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* ***22****, 134-167.*

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

Breslow, N. (1982). Covariance adjustment of relative-risk estimates in matched studies. *Biometrics* ***38***, *661-672.*

Brown, L.D. (1986). *Fundamentals of Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Lecture Notes-Monograph Series, Volume 9.

Clogg, C.C. (1982). Some models for the analysis of association in multi-way cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* ***77***, *803-815.*

Cox, D.R. and Snell, E.J. (1989). (2nd ed.) *Analysis of Binary Data.* London: Chapman and Hall.

Crowther, N.A.S. and Shaw, T.M. (1989). An estimation procedure in categorical data analysis. *S.A. Statistical Journal,* ***23(1)***, *63-83.*

Deming, W.E. and Stefan W.W. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* ***11***, *427-444.* London: Chapman and Hall.

Dobson, A.J. (1983). *An Introduction to Statistical Modelling.* London: Chapman and Hall.

——— (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Everitt, B. S. (1977). *The Analysis of Contingency Tables.* London: Chapman and Hall.

Grizzle, J.E. Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics* ***25***, *489-504.*

Grizzle, J.E. and Williams, O.D. (1972). Loglinear models and tests of independence for contingency tables. *Biometrics* ***28***, *137-156.*

Goodman, L.A. (1979a). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* ***74***, *537-552.*

——— (1979b). Multiplicative models for square contingency tables with ordered categories. *Biometrica* ***66***, *413-418.*

——— (1983). The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* ***39***, *149-160.*

——— (1994). On quasi-independence and quasi-dependence in contingency tables, with special reference to ordinal triangular contingency tables. *J. Amer. Statist. Assoc.* ***89***, *1059-1063.*

Haber, M. and Brown, M.B. (1986). Maximum likelihood estimation for log-linear models when expected frequencies are subject to linear constraints. *J. Amer. Statist. Assoc.* ***81***, *477-482.*

Haberman, S.J. (1973). Loglinear models for frequency-data: sufficient statistics and likelihood equations. *Ann. Statist.* ***1***, *617-632.*

——— (1974a). *The Analysis of Frequency Data.* Chicago: University of Chicago Press.

——— (1974b). Log-linear models for frequency tables with ordered classifications. *Biometrica* ***36**, 589-600.*

——— (1978, 1979). *The Analysis of Qualitative Data*, Vols 1 and 2. New York: Academic Press.

Hedlund, R.D. (1978). Cross-over voting in a 1976 presidential primary. *Public Opinion Quart.* ***41**, 498-514.*

Jorgensen, B. (1986). Some properties of exponential dispersion models. *Scand. J. Statist.* ***13**, 187-198.*

——— (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc.* ***B49**, 127-162.*

——— (1992). Exponential dispersion models and extensions: a Review. *International Statistical Review* ***60**, 5-20.*

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* ***33**, 133-158.*

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* ***33**, 159-175.*

Little, J.A. and Wu, M.M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *J. Amer. Statist. Assoc.* ***86**, 87-95.*

Magnus, J.R. and Neudecker, H. (1991). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester: Wiley.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc.* ***B42****, 109-142.*

McCullagh, P. and Nelder, J.A. (1983), 2nd edn. (1989). *Generalized Linear Models.* London: Chapman and Hall.

Nelder , J. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc.* ***A135****, 370-384.*

Neyman, J. (1949). Contributions to the theory of the χ^2 test. *Proc. Berkeley Symp. Math. Statist. and Prob., 239-272.*

SAS Institute Inc. *SAS/IML^{TM} User's Guide for Personal Computers, Version 6 Edition.* Cary NC: SAS Institute Inc., 1985.

SAS Instsitute Inc. *SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2.* Cary NC: SAS Institute Inc., 1990.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* New York: Wiley.

Silvey, S.D. (1959). The Lagrangian multiplier test. *Ann. Math. Stat.,* ***30****, 389-407.*

Steyn, H.S., Crowther, N.A.S., Raath E.L. and Shaw, T.M. (1989). A general model for negative multinomial frequency counts. *S.A. Statistical Journal,* ***23(1)****, 85-117.*

Tanner, M.A. and Young, M.A. (1985). Modeling agreement among raters. *J. Amer. Statist. Assoc.* ***80****, 175-180.*

Tweedie, M.C.K. (1947). Functions of a statistical variate with given means, with special reference to Laplacian distributions. *Proc. Camb. Phil. Soc.* ***49****, 41-49.*

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). Generalized linear models specified in terms of linear constraints. *J. Roy. Statist. Soc.* ***B36****, 449-454.*

APPENDIX

The IML programs for certain examples are given in the Appendix and appear under the appropriate chapter heading and example number.

CHAPTER 2

Example 2.6

(i) constraint: $\sigma_{11} = 4, \ \ n = 100.$

```
proc iml worksize=50;
reset nolog;
x={ 3,3,2};
n=100;
V=J(3,3,0);
GG={1 0 0};
diff1=1;
i=0;
x1=x;
do while (diff1>0.000001);
i=i+1;
m=x;
x=x1;
V[1,1]=2#m[1]##2;   V[1,2]=2#m[3]##2;   V[1,3]=2#m[1]#m[3];
V[2,1]=2#m[3]##2;   V[2,2]=2#m[2]##2;   V[2,3]=2#m[2]#m[3];
V[3,1]=2#m[1]#m[3]; V[3,2]=2#m[2]#m[3]; V[3,3]=m[1]#m[2]+m[3]##2;
V=V/n;
Gm=GG;
diff=1;
j=0;
do while (diff>0.000001);
xv=x;
j=j+1;
Gx=GG;
g=Gx*x-4;
x=x-(Gm*V)`*ginv(Gx*V*Gm`)*g;
nt=x[+];
diff=(x-xv)`*(x-xv);
print i j nt x m;
end;
diff1=(m-x)`*(m-x);
end;
```

(iii) constraint: $\rho = \frac{1}{2}$.

```
proc iml worksize=50;
reset nolog;
x={ 3,3,2};
n=100;
r=0.5;
V=J(3,3,0);
Gm={0 0 0};
Gx={0 0 0};
diff1=1;
i=0;
x1=x;
do while (diff1>0.00000001);
i=i+1;
m=x;
x=x1;
V[1,1]=2#m[1]##2;   V[1,2]=2#m[3]##2;   V[1,3]=2#m[1]#m[3];
V[2,1]=2#m[3]##2;   V[2,2]=2#m[2]##2;   V[2,3]=2#m[2]#m[3];
V[3,1]=2#m[1]#m[3]; V[3,2]=2#m[2]#m[3]; V[3,3]=m[1]#m[2]+m[3]##2;
V=V/n;
c=sqrt(m[2]/m[1]);c1=-r#c/2; c2=-r/c/2; Gm[1]=c1; Gm[2]=c2; Gm[3]=1;
diff=1;
j=0;
do while (diff>0.00000001);
xv=x;
j=j+1;
c=sqrt(x[2]/x[1]);c1=-r#c/2; c2=-r/c/2; Gx[1]=c1; Gx[2]=c2; Gx[3]=1;
g=x[3]-r#sqrt(x[1]#x[2]);
x=x-(Gm*V)`*ginv(Gx*V*Gm`)*g;
nt=x[+];
diff=(x-xv)`*(x-xv);
print i j nt x m;
end;
diff1=(m-x)`*(m-x);
end;
```

CHAPTER 3

Example 3.1

```
proc iml worksize=80;
reset nolog;
n=4;
*------------>; y={ 72,714,655,41};
SIG=diag(y);
G={1 0 0  -1 ,
   0 1 -1 0 };
    itr=0;
    diff=1;
    m=y;
    do while (diff>0.00000001);
      gm=G*m;
      m1=m;
      m=y-(G*SIG)`*ginv(G*SIG*G`)*G*y;
      SIG=diag(m);
      diff=sqrt((m-m1)`*(m-m1));
      itr=itr+1;
    end;
   mi=1/m;
   X2=(y-m)`*(mi#(y-m));
   prob=1-probchi(x2,2);
   print  y[format=5.2] m[format=5.2];
   print X2 prob;
```

Example 3.2

```
proc iml;
* Drug comparisons Grizzle et al. (1969);
reset nolog;
* Frequency vector;
*---->;  f={6,16,2,4,2,4,6,6};
 n=sum(f);
* Matrix of constraints;
      A={0  0  1   1  -1 -1  0  0,
         0  1 -1   0   0  1  -1  0};
 m=f;
 diff=1;
 itr=0;
```

```
  do while (diff>0.00000001);
     S=diag(m);
    m1=m;
     m=f-(A*S)'*ginv(A*S*A')*A*f;
     diff=sqrt((m-m1)'*(m-m1));
     itr=itr+1;
   end;
mi=1/m;
G2=2*f'*log(f/m);
X2=(f-m)'*(mi#(f-m));
vec=x2||g2;
df=nrow(A);
prob=J(1,2,1)-probchi(vec,df);
vec1={"Pearson" "LR"  };
R={"Chi^2" "Df" "Prob"};
TEST=vec//J(1,2,df)//prob;
print "Expected frequencies, no. of iterations";
print m itr;
print" ";
print "Chi-squared statistics with attained significance levels";
print TEST[rowname=R colname=vec1 format=15.6];
```

Example 3.4

This program can also be used for loglinear models for higher dimensional tables. The information to be supplied is :

(a) the frequencies, entered as a vector "**x**".

(b) the number of variables, "nf".

(c) the names of the variables, "name".

(d) the number of levels of each variable, "**k**", a vector.

(e) the index vector, "nh" which references the λ's which will be set to zero, e.g. nh=4 will set all $\lambda_{ij}^{AB} = 0$.

(f) the matrix "**G1**" is used to impose constraints on the λ's, for example in the symmetry and quasi-symmetry models certain λ's are equated.
The output from the program is the following :

(a) the expected frequencies

(b) the parameter estimates for the model fitted

(c) the statistics, degrees of freedom and exact p-value.

```
proc iml worksize= 80;
options pagesize=200;
*-------------  FREQUENCY VECTOR(factor that changes slowest first );
*------------>; x={ 143, 156, 100,
                   119, 210, 141,
                    15,  72, 127};
*-------------  NUMBER OF VARIABLES  ----;
*------------>; nf=2;
*-------------  NAME OF VARIABLES  -----;
*--------->; name={"aff.","ideol."};
*---------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR  (MAX 6 FACTORS) --;
*--------->; k[1,]=3  ;k[2,]=3  ;k[3,]=0  ; k[4,]=0  ;k[5,]=0 ;k[6,]=0 ;
*---------------------------------------------------------------;
*---------- SPECIFICATION OF HYPOTHESIS MATRIX AH --------------;
*index vector nh of the hypothesis in the order (lambdas set to zero)
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  ens.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={4};
*----->;G1={0  0};
*----------------------------------------------------------------;
*-------------- CONSTRUCTION OF DESIGN MATRIX A ----------------;
reset nolog;
reset fw=10;
c=k[1,];
een=J(c,1,1);
d=c-1;
A=(i(d)//J(1,d,-1));
e=k[1];
do i=2 to nf;
c=k[i,];
one=J(c,1,1);
d=c-1;
Y=(I(d)//J(1,d,-1));
one1=j(e,1,1);
A1=A@one;
Y1=one1@Y;
```

```
A=A1||Y1;
A=A||hdir(A1,Y1);
e=k[i,]*e;
end;
A=j(e,1,1)||A;
*-----------------------------------------------------------------;
vg=1;
do i=1 to nf;
vg=vg//((k[i,]-1)*vg);
end;
kol=cusum(vg);
nrh=nrow(nh);
ii=nh[1,]-1;iii=nh[1,];
a1=kol[ii,]+1;a2=kol[iii,];
AH=A[,a1:a2];
do i=2 to nrh;
ii=nh[i,]-1;iii=nh[i,];
a1=kol[ii,]+1;a2=kol[iii,];
AH=AH||A[,a1:a2];
end;
*-------------- CONSTRUCTION OF THE INDEX VECTOR----------------;
tyd=name[1,1];
name1={""}//tyd;
do i=2 to nf;
name1=name1//concat(name1,name[i,1]);
end;
name1=rowcatc(name1);
nn=nrow(name1);
index={"mu"};
do i=2 to nn;
tyd=name1[i,1];
index=index//repeat(tyd,vg[i,1]);
end;
*-----------------------------------------------------------------;
*---------------- HYPOTHESIS MATRIX WITH STRUCTURE -------------;
sg=sum(g1*g1`);
if sg^=0 then AH=AH*G1`;
*-----------------------------------------------------------------;
free Y A1 Y1 tyd name1 ;
A=inv(A`*A)*A`;
```

```
lambda=A*log(x);
    x1=1/x;
varl=A*(x1#A‘);
stdl=sqrt(vecdiag(varl));
standl=lambda/stdl;
free varl stdl ;
gx=AH‘*log(x);
   m=x;
   gm=gx;
   itr=0;
   diff=1;
   do while (diff>0.000001);
   m1=m;
   mi=1/m;
   m=m-AH*inv(AH‘*(mi#AH))*gm;
   gm=AH‘*log(m);
   diff=sqrt((m-m1)‘*(m-m1));
   itr=itr+1;
   end;
lambdah=A*log(m);
varlh=vecdiag(A*(mi#A‘))-vecdiag(A*(mi#AH)*inv(AH‘*(mi#AH))*AH‘*(mi#A‘));
vecvar=varlh<>J(e,1,1E-10);
stdlh=sqrt(vecvar);
standlh=lambdah/stdlh;
df=ncol(ah);
X2=(x-m)‘*(mi#(x-m));
G2=2*x‘*log(x/m);
K2ft=4*(sqrt(x)-sqrt(m))‘*(sqrt(x)-sqrt(m));
Wald=gx‘*inv(AH‘*(x1#AH))*gx;
vec=x2||G2||K2ft||Wald;
prob=J(1,4,1)-probchi(vec,df);
*output;
vec1={"Pearson" "LR" "F-T" "Wald" };
R={"Chi^2" "Df" "Prob"};
Test=vec//J(1,4,df)//prob;
xr=nrow(x);
nrt=xr/k[nf];
x=shape(x,nrt);
m=shape(m,nrt);
print "------------ LOG.IML ----------------";
```

```
print"number of iterations=" itr; print" " ;
print x[format=7.1]  m[format=12.6]; print" ";
print  index lambda[format=12.6] standl[format=12.6]
       lambdah [format=12.6] standlh[format=12.6];
print" ";
print "Chi squared statistics with exceedance probabilities";
print Test[rowname=R colname=vec1 format=15.6];
```

Example 3.5

The program is the same as that of Example 3.4, hence only the segment for the information to be supplied is given. The matrix **G1** is the matrix of constraints for the symmetry model. For the quasi-symmetry model the matrix **G1** is the last row of the **G1** in the segment.

```
proc iml worksize= 50;
options pagesize=200;
*-------------  FREQUENCY VECTOR(factor that changes slowest first );
*------------>; x={ 198,65,59,
                   63,79,66,
                   89,76,272};
*------------- NUMBER OF VARIABLES ----;
*------------>; nf=2;
*------------- NAME OF VARIABLES ------;
*--------->; name={"e.","p."};
*---------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR  (MAX 6 FACTORS) --;
*--------->; k[1,]=3 ;k[2,]=3 ;k[3,]=0 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*---------------------------------------------------------------;
*---------- SPECIFICATION OF HYPOTHESIS MATRIX AH --------------;
*index vector nh of the hypothesis in the order (lambdas set to zero)
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  ens.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3,4};
*----->;G1={1 0 -1  0  0  0  0  0,
           0 1  0 -1  0  0  0  0,
           0 0  0  0  0  1 -1  0};
*---------------------------------------------------------------;
```

Example 3.6

The program is that of Example 3.4, hence only the segment for the information to be supplied is given.

```
proc iml worksize= 50;
options pagesize=200;
*-------------  FREQUENCY VECTOR(factor that changes slowest first);
*------------>; x={  72 ,47 ,
                   110 ,196,
                    44 ,179,
                    86 , 38,
                   173 , 283,
                    28 , 187};
*------------- NUMBER OF VARIABLES ----;
*------------>; nf=3;
*------------- NAME OF VARIABLES -----;
*--------->; name={"s.","e.","r."};
*---------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR  (MAX 6 FACTORS) --;
*--------->;k[1,]=2 ;k[2,]=3 ;k[3,]=2 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*---------------------------------------------------------------;
*---------- SPECIFICATION OF HYPOTHESIS MATRIX AH -------------;
*index vector nh of the hypothesis in the order (lambdas set to zero)
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  ens.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={6,8};
*----->;G1={0  0};
*---------------------------------------------------------------;
```

Example 3.7

This program allows the user to introduce covariates into the design matrix along with the other main effects to be included. The difference in this program is that the quantity "nh" now specifies the effects to be *added* to the design matrix and **not** *set to zero as in the previous cases.*

```
proc iml worksize=80;
options pagesize=200;
*-----> FREQUENCY VECTOR;
*----->;  x={ 143,156 ,100,
              119,210 ,141,
               15, 72 ,127};
```

```
*-----> COVARIATES TO BE INCLUDED IN DESIGN MATRIX;
*----->;cov={  -1  0,
                0  0,
                1  0,
                0 -1,
                0  0,
                0  1,
                1  1,
                0  0,
               -1 -1};
*------------- NUMBER OF FACTORS -------;
*------------>; nf=2;
*-------------  NAMES OF FACTORS -------;
*--------->; name={"p.","i."};
*-------------  NAMES OF COVARIATES;
*--------->;namecov={"t1","t2"};
*---------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR (MAX 6 FACTORS) ---;
*--------->; k[1,]=3 ;k[2,]=3 ;k[3,]=0 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*---------------------------------------------------------------;
*---------- SPECIFICATION OF DESIGN MATRIX ---------------------;
*index vector nh of main effects and interactions to be included
  1  A (I+A)B (I+(A+(I+A)B))C (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3};
*---------------------------------------------------------------;
*----------- CONSTRUCTION OF THE DESIGN MATRIX -----------------;
reset nolog;
reset fw=10;
c=k[1,];
one=J(c,1,1);
d=c-1;
A=(i(d)//J(1,d,-1));
e=k[1];
do i=2 to nf;
c=k[i,];
one=J(c,1,1);
d=c-1;
Y=(I(d)//J(1,d,-1));
```

```
one1=j(e,1,1);
A1=A@one;
Y1=one1@Y;
A=A1||Y1;
A=A||hdir(A1,Y1);
e=k[i,]*e;
end;
A=j(e,1,1)||A;
*---------- CONSTRUCTION OF THE INDEX VECTOR --------------------;
vg=1;
do i=1 to nf;
vg=vg//((k[i,]-1)*vg);
end;
tyd=name[1,1];
name1={""}//tyd;
do i=2 to nf;
name1=name1//concat(name1,name[i,1]);
end;
name1=rowcatc(name1);
nn=nrow(name1);
index={"mu"};
do i=2 to nn;
tyd=name1[i,1];
index=index//repeat(tyd,vg[i,1]);
end;
*---------------------------------------------------------------;
kol=cusum(vg);
nrh=nrow(nh);
AA=J(e,1,1);
index1=index[1,];
do i=1 to nrh;
ii=nh[i,]-1;iii=nh[i,];
a1=kol[ii,]+1;a2=kol[iii,];
AA=AA||A[,a1:a2];
index1=index1//index[a1:a2,];
end;
*---------------------------------------------------------------;
A=AA||cov;
index=index1//namecov;
nc=ncol(cov);
```

```
vg=vg//J(nc,1,1);
AH=I(e)-A*ginv(A`*A)*A`;
vgh=e-ncol(A);
call svd(ah,q,v,ah);
AH=AH[,1:vgh];
*-------------------------------------------------------------;
x=x<>J(e,1,1e-6);
x1=1/x;
*----->;G1={0  0};
*-------------------------------------------------------------;
*------------- HYPOTHESIS MATRIX WITH STRUCTURE ----------------;
sg=sum(g1*g1`);
if sg^=0 then AH=AH*G1`;
*-------------------------------------------------------------;
free name1 ;
A=inv(A`*A)*A`;
gx=AH`*log(x);
   m=x;
   gm=gx;
   itr=0;
   diff=1;
   do while (diff>0.000001);
   m1=m;
   mi=1/m;
   m=m-AH*ginv(AH`*(mi#AH))*gm;
*m=m<>J(e,1,1e-12);
   gm=AH`*log(m);
   diff=sqrt((m-m1)`*(m-m1));
   itr=itr+1;
   end;
xr=nrow(x);
lambda=A*log(m);
    mi=1/m;
varl=vecdiag(A*(mi#A`))-vecdiag(A*(mi#AH)*ginv(AH`*(mi#AH))*AH`*(mi#A`));
vecvar=varl;
stdl=sqrt(vecvar);
standl=lambda/stdl;
free varl stdl ;
X2=(x-m)`*(mi#(x-m));
G2=2*x`*log(x/m);
```

```
K2ft=4*(sqrt(x)-sqrt(m))'*(sqrt(x)-sqrt(m));
Wald=gx'*ginv(AH'*(x1#AH))*gx;
vec=x2||g2||K2ft||Wald;
prob=J(1,4,1)-probchi(vec,vgh);
*PRINTING OF THE OUTPUT;
vec1={"Pearson" "LR" "F-T" "Wald" };
R={"Chi^2" "Df" "Prob"};
TEST=vec//J(1,4,vgh)//prob;
xr=nrow(x);
nrt=xr/k[nf];
x=shape(x,nrt);
m=shape(m,nrt);
print "------------LOG.IML----------------";
print"number of iterations =" itr;
print" " ;
print x[format=7.1]  m[format=12.6] ;
print" ";
print  index lambda[format=12.6] standl[format=12.6];
print" ";
print "Chi-squared statistics with exact p-values";
print TEST[rowname=R colname=vec1 format=15.6];
print" ";
* calculation of AIC;
 aic=G2+2*(xr-vgh);
 print "Akaike Information Criterion ";
 print "AIC=" aic;
```

Example 3.8

This program is the same as that of Example 3.7, hence only the relevant information to be supplied, is given.

```
proc iml worksize=80;
options pagesize=500;
*-----> FREQUENCY VECTOR;
*----->;  x={  5,  6, 53, 16,
              7, 15,115, 25,
             13, 20,197, 96,
              7,  3, 33,  7,
              8,  9, 65, 23,
             13, 13,149, 56 };
*-----> COVARIATES TO BE INCLUDED IN DESIGN MATRIX;
```

```
*----->;cov={ 1.5    -1,
              0.5    -1,
             -0.5    -1,
             -1.5    -1,
              0       0,
              0       0,
              0       0,
              0       0,
             -1.5     1,
             -0.5     1,
              0.5     1,
              1.5     1,
              1.5    -1,
              0.5    -1,
             -0.5    -1,
             -1.5    -1,
              0       0,
              0       0,
              0       0,
              0       0,
             -1.5     1,
             -0.5     1,
              0.5     1,
              1.5     1 };
*------------- NUMBER OF FACTORS -------;
*------------>; nf=3;
*------------- NAMES OF FACTORS --------;
*--------->; name={"l.","a.","o."};
*--------->;namecov={"a.o","age"};
*--------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR (MAX 6 FACTORS) ---;
*--------->; k[1,]=2 ;k[2,]=3 ;k[3,]=4 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*--------------------------------------------------------------;
*---------- SPECIFICATION OF DESIGN MATRIX ---------------------;
*index vector nh of main effects and interactions to be INCLUDED
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,5};
*--------------------------------------------------------------;
```

Example 3.10

```
proc iml worksize=50;
reset nolog;
*---->; f={64,57,57,72,36,21,
          94,94,105,141,97,71,
          58,54,65,77,54,54,
          46,40,60,94,78,71};
n=15;
A={1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0  1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0  0  1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0  0  0  1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 0  1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 0  0  1 -1 0 0 0 0 -1 1 0 0 0 0 0 0 0,
   0 0 0 0 0 0 0 0  0  0  1 -1 0 0 0 0 -1 1 0 0 0 0 0 0,
   0 0 0 0 0 0 0 0 0 0 0 0 1 -1 0 0 0 0 -1 1 0 0 0 0,
   0 0 0 0 0 0 0 0 0 0 0 0 0 1 -1 0 0 0 0 -1 1 0 0 0,
   0 0 0 0 0 0 0 0 0 0 0 0 0 0  1 -1 0 0 0 0 -1 1 0 0,
   0 0 0 0 0 0 0 0 0 0 0 0 0 0  0  1 -1 0 0 0 0 -1 1 0,
   0 0 0 0 0 0 0 0 0 0 0 0 0 0  0  0  1 -1 0 0 0 0 -1 1};
 c=j(5,1,1);
 D=I(3)@c;  ;
 H=I(5)//I(5)//I(5);
 X=H||D;
 m=f;
 odds=A*log(m);
 P=I(n)-x*ginv(x`*x)*x`;
 P=P*A;
   itr=0;
   diff=1;
   do while (diff>0.00000001);
 SIG=diag(m);
 siginv=inv(sig);
  m1=m;
  m=m-(P)`*ginv(P*siginv*P`)*P*log(m);
   diff=sqrt((m-m1)`*(m-m1));
```

```
   itr=itr+1;
   end;
 b=ginv(x'*x)*x'*A*log(m);
 print b ;
 theta=exp(b);
 print theta;
   mi=1/m;
 G2=2*f'*log(f/m);
 X2=(f-m)'*(mi#(f-m));
 print G2 X2 ;
 print itr;
 print "Expected Frequencies";
 print m;
```

CHAPTER 4

Example 4.1 : Marginal Homogeneity

```
proc iml;
* Marginal Homogeneity For Vision Data ;
reset nolog;
f={1520,266,124,66,234,1512,432,78,
   117,362,1772,205,36,82,179,492};
A={0  1  1  1 -1  0  0  0 -1 0  0  0 -1  0  0  0,
   0 -1  0  0  1  0  1  1  0 -1 0  0  0 -1  0  0,
   0  0 -1  0  0  0 -1  0  1  1 0  1  0  0 -1  0};
     m=f;
     diff=1;
     itr=0;
   do while (diff>0.00000001);
     S=diag(m);
     m1=m;
     m=f-(A*S)'*ginv(A*S*A')*A*f;
     diff=sqrt((m-m1)'*(m-m1));
     itr=itr+1;
   end;
  mi=1/m;
 G2=2*f'*log(f/m);
 X2=(f-m)'*(mi#(f-m));
 print G2 X2 ;
 print itr;
```

```
print "Expected Frequencies";
print m;
```

Example 4.2: Conditional Symmetry

```
proc iml;
* Example taken from Agresti(1990) p364;
reset nolog;
X={1,1,1,1,1,1};
f={31, 5,5, 1e-6,12,1,1e-6,1e-6,14,1,2,1,6,1,1,1e-6};
C={0  1  0  0 -1  0  0  0  0  0  0  0  0  0  0  0,
   0  0  1  0  0  0  0  0 -1  0  0  0  0  0  0  0,
   0  0  0  1  0  0  0  0  0  0  0  0 -1  0  0  0,
   0  0  0  0  0  0  1  0  0 -1  0  0  0  0  0  0,
   0  0  0  0  0  0  0  1  0  0  0  0  0 -1  0  0,
   0  0  0  0  0  0  0  0  0  0  0  1  0  0 -1  0};
      P=I(6)-x*ginv(x`*x)*x`;
      K=P*C;
      m=f;
      diff=1;
      itr=0;
   do while (diff>0.00000001);
     Diagf=diag(m);
     Di=inv(diagf);
     m1=m;
     m=m-K`*ginv(K*Di*K`)*K*log(m);
     diff=sqrt((m-m1)`*(m-m1));
     itr=itr+1;
   end;
     G2=2*f`*log(f/m);
     mi=1/m;
     X2=(f-m)`*(mi#(f-m));
     print G2 X2 ;
     print itr;
     print " ";
     print "Expected Frequencies";
     print m[format=12.5];
     b=inv(x`*x)*x`*C*log(m);
  print "Estimated Regression Parameters";
  print b;
```

Example 4.3: Diagonals-parameter model

```
proc iml;
* Diagonals parameter on vision data;
reset nolog;
X={1 0 0,
   0 1 0,
   0 0 1,
   1 0 0,
   0 1 0,
   1 0 0};
xrow=nrow(x);
xcol=ncol(x);
df=xrow-xcol;
f={1520,266,124,66,
   234,1512,432,78,
   117,362,1772,205,
   36,82,179,492};
C={0  1  0  0 -1  0  0  0  0  0  0  0  0  0  0  0,
   0  0  1  0  0  0  0  0 -1  0  0  0  0  0  0  0,
   0  0  0  1  0  0  0  0  0  0  0  0 -1  0  0  0,
   0  0  0  0  0  0  1  0  0 -1  0  0  0  0  0  0,
   0  0  0  0  0  0  0  1  0  0  0  0  0 -1  0  0,
   0  0  0  0  0  0  0  0  0  0  0  1  0  0 -1  0};
     P=I(6)-x*ginv(x'*x)*x';
     K=P*C;
     m=f;
     diff=1;
     itr=0;
   do while (diff>0.00000001);
     Diagf=diag(m);
     Di=inv(diagf);
     m1=m;
     m=m-K'*ginv(K*Di*K')*K*log(m);
     diff=sqrt((m-m1)'*(m-m1));
     itr=itr+1;
   end;
 mi=1/m;
 G2=2*f'*log(f/m);
 X2=(f-m)'*(mi#(f-m));
```

```
 b=inv(x'*x)*x'*C*log(m);
*PRINTING OF THE OUTPUT;
 print itr;
 print " ";
 print "Expected Frequencies";
 print m[format=12.5];
 print "Estimated Regression Parameters";
 print b;
 vec=X2||G2;
 prob=J(1,2,1)-probchi(vec,df);
 vec1={"Pearson" "LR" };
 R={"Chi^2" "Df" "Prob"};
 TEST=vec//J(1,2,df)//prob;
 print" ";
 print "Chi-squared statistics with p-values";
 print TEST[rowname=R colname=vec1 format=15.6];
```

Example 4.4: Mobility Tables

This example utilizes the program of Example 3.6, introducing covariates into the design matrix. Only the relevant information to be supplied is given.

```
proc iml worksize=80;
options pagesize=200;
*----->Social Mobility : Haberman p 488;
*-----> FREQUENCY VECTOR;
*----->;  n={ 259,123,2,0,
              82,370,30,7,
              5,59,34,4,
              2,41,29,8};
 cov={0 1 0 0 0 ,
      1 0 0 0 0 ,
      2 0 0 0 0 ,
      3 0 0 0 0 ,
      1 0 0 0 0 ,
      0 0 1 0 0 ,
      1 0 0 0 0 ,
      2 0 0 0 0 ,
      2 0 0 0 0 ,
      1 0 0 0 0 ,
      0 0 0 1 0 ,
      1 0 0 0 0 ,
```

```
      3 0 0 0 0 ,
      2 0 0 0 0 ,
      1 0 0 0 0 ,
      0 0 0 0 1 };
*------------- NUMBER OF FACTORS -------;
*------------>; nf=2;
*------------- NAMES OF FACTORS -------;
*--------->; name={"husband","wife"};
*--------->;namecov={"d","a1","a2","a3","a4"};
*-------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR (MAX 6 FACTORS) ---;
*--------->; k[1,]=4 ;k[2,]=4 ;k[3,]=0 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*-------------------------------------------------------------;
*-------------------------------------------------------------;
*---------- SPECIFICATION OF DESIGN MATRIX ---------------------;
*index vector nh of main effects and interactions to be included
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3};
*-------------------------------------------------------------;
```

Example 4.5: Rater agreement.
This example also utilizes the "covariate program" of Example 3.6, hence only information to be supplied is given.

```
proc iml worksize=100;
options pagesize=500;
*-----> Loglinear models with covariates;
*-----> FREQUENCY VECTOR;
*----->;  n={ 38,5,0,1,
              33,11,3,0,
              10,14,5,6,
               3,7,3,10,
               5,3,0,0,
               3,11,4,0,
               2,13,3,4,
               1,2,4,14};
*-----> COVARIATES TO BE INCLUDED IN DESIGN MATRIX;
 cov={1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1,
 1 2 3 4 2 4 6 8 3 6 9 12 4 8 12 16 1 2 3 4 2 4 6 8 3 6 9 12 4 8 12 16};
```

```
 cov=cov‘;
*------------- NUMBER OF FACTORS -------;
*------------>; nf=3;
*------------- NAMES OF FACTORS --------;
*--------->; name={"s","r1","r2"};
*--------->; namecov={"d","b"};
*-------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS FOR EACH FACTOR (MAX 6 FACTORS) ---;
*--------->; k[1,]=2 ;k[2,]=4 ;k[3,]=4 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*-------------------------------------------------------------;
*-------------------------------------------------------------;
*-------SPECIFICATION OF DESIGN MATRIX ;
*index vector nh of main effects and interactions to be INCLUDED
  1  A  (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->; nh={2,3,4,5,6};
*-------------------------------------------------------------;
```

Example 4.6

```
proc iml worksize=200;
reset nolog;
x={ 11607,100,366,124,
     87,13677,515,302,
     172,225,17819,270,
     63,176,286,10192};
n=x[+];
Gx=J(3,16,0);
Gm=J(3,16,0);
x1=x;
do i=1 to 5;
*m=x<>0.0001;
m=x;
x=x1;
V=diag(m)-m*m‘/n;
Gm[1,2]= m[8]#m[13]; Gm[1,4]=-m[5]#m[14]; Gm[1,5]=-m[4]#m[14];
Gm[1,8]= m[2]#m[13]; Gm[1,13]=m[2]#m[8]; Gm[1,14]=-m[5]#m[4];
Gm[2,3]= m[12]#m[13]; Gm[2,4]=-m[9]#m[15]; Gm[2,9]=-m[4]#m[15];
Gm[2,12]= m[3]#m[13]; Gm[2,13]=m[3]#m[12]; Gm[2,15]=-m[9]#m[4];
Gm[3,7]=m[12]#m[14]; Gm[3,8]=-m[10]#m[15]; Gm[3,10]=-m[8]#m[15];
```

```
Gm[3,12]= m[7]#m[14]; Gm[3,14]=m[7]#m[12]; Gm[3,15]=-m[10]#m[8];
do j=1 to 5;
g1=x[2]#x[8]#x[13]-x[5]#x[4]#x[14];
g2=x[3]#x[12]#x[13]-x[9]#x[4]#x[15];
g3=x[7]#x[12]#x[14]-x[10]#x[8]#x[15];
g=g1//g2//g3;
Gx[1,2]= x[8]#x[13]; Gx[1,4]=-x[5]#x[14]; Gx[1,5]=-x[4]#x[14];
Gx[1,8]= x[2]#x[13]; Gx[1,13]=x[2]#x[8]; Gx[1,14]=-x[5]#x[4];
Gx[2,3]= x[12]#x[13]; Gx[2,4]=-x[9]#x[15]; Gx[2,9]=-x[4]#x[15];
Gx[2,12]= x[3]#x[13]; Gx[2,13]=x[3]#x[12]; Gx[2,15]=-x[9]#x[4];
Gx[3,7]=x[12]#x[14]; Gx[3,8]=-x[10]#x[15]; Gx[3,10]=-x[8]#x[15];
Gx[3,12]= x[7]#x[14]; Gx[3,14]=x[7]#x[12]; Gx[3,15]=-x[10]#x[8];
x=x-(Gm*V)`*ginv(Gx*V*Gm`)*g;
end;
end;
 xi=1/x;
x2=(x1-x)`*(xi#(x1-x));
lr=2*x1`*log(x1/x);
print lr x2;
print x1 m;
```

Example 4.7

The first program uses the cross-product ratios for the set of constraints.

```
proc iml worksize=200;
reset nolog;
x={ 11607,100,366,124,
     87,13677,515,302,
     172,225,17818,270,
     63,176,286,10192};
n=x[+];
Gx=J(5,16,0);
Gm=J(5,16,0);
x1=x;
do i=1 to 5;
m=x<>0.0001;
x=x1;
V=diag(m)-m*m`/n;
Gm[1,5]= m[12]; Gm[1,8]=-m[9]; Gm[1,9]=-m[8]; Gm[1,12]= m[5];
Gm[2,3]= m[8]; Gm[2,4]=-m[7]; Gm[2,7]=-m[4]; Gm[2,8]= m[3];
Gm[3,2]= m[15]; Gm[3,3]=-m[14]; Gm[3,14]=-m[3]; Gm[3,15]= m[2];
```

```
Gm[4,5]= m[15]; Gm[4,7]=-m[13]; Gm[4,13]=-m[7]; Gm[4,15]= m[5];
Gm[5,9]= m[14]; Gm[5,10]=-m[13]; Gm[5,13]=-m[10]; Gm[5,14]= m[9];
do j=1 to 5;
g1=x[5]#x[12]-x[8]#x[9];
g2=x[3]#x[8]-x[4]#x[7];
g3=x[2]#x[15]-x[3]#x[14];
g4=x[5]#x[15]-x[7]#x[13];
g5=x[9]#x[14]-x[10]#x[13];
g=g1//g2//g3//g4//g5;
Gx[1,5]= x[12]; Gx[1,8]=-x[9]; Gx[1,9]=-x[8]; Gx[1,12]= x[5];
Gx[2,3]= x[8]; Gx[2,4]=-x[7]; Gx[2,7]=-x[4]; Gx[2,8]= x[3];
Gx[3,2]= x[15]; Gx[3,3]=-x[14]; Gx[3,14]=-x[3]; Gx[3,15]= x[2];
Gx[4,5]= x[15]; Gx[4,7]=-x[13]; Gx[4,13]=-x[7]; Gx[4,15]= x[5];
Gx[5,9]= x[14]; Gx[5,10]=-x[13]; Gx[5,13]=-x[10]; Gx[5,14]= x[9];
x=x-(Gm*V)'*ginv(Gx*V*Gm')*g;
nt=x[+];
end;
end;
print x m;
```

The next program places constraints on the frequencies and is the same as that in Example 3.6 hence only the information to to be entered is given.

```
proc iml worksize= 80;
*-------------  FREQUENCY VECTOR(slowest changing vector first);
*------------>; x={ 11607, 100,  366,  124,
                      87, 13677,  515,  302,
                     172,   225, 17819, 270,
                      63,   176,  286 , 10192};
xr =nrow(x);
x=x<>J(xr,1,1e-12);
*-------------  NUMBER OF VARIABLES ----;
*------------>; nf=2;
*-------------  NAMES OF VARIABLES -----;
*--------->; Name={"1980.","1985."};
*--------------------------------------------------------------;
k=j(6,1,0);
*----- NUMBER OF VARIABLES FOR EACH VARIABLE (MAX 6 VARIABLES)--;
*------>; k[1,]=4 ;k[2,]=4 ;k[3,]=0 ; k[4,]=0 ;k[5,]=0 ;k[6,]=0 ;
*--------------------------------------------------------------;
*-------- SPECIFICATION OF HYPOTHESIS MATRIX AH ----------------;
```

```
*index vector nh:effects to be included in the design matrix  :
  1  A  (I+A)B (I+(A+(I+A)B))C (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3};
*---------------------------------------------------------------;
*---------- SPECIFICATION OF THE CONSTRAINTS -------------------;
*--->; C1={ 1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0,
           0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0,
           0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0,
           0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1};
cov=C1`;
*---------------------------------------------------------------;
```

Example 4.8

This program uses the cross-product ratios to set up the constraints as described in the chapter.

```
proc iml worksize=200;
reset nolog;
x={ 11, 23, 12, 15,  8,
     9, 10,  4,  1,  0,
     6,  4,  4,  0,  0,
     4,  5,  0,  0,  0,
     5,  0,  0,  0,  0};
n=x[+];
Gx=J(6,25,0);
Gm=J(6,25,0);
x1=x;
do i=1 to 5;
m=x;
x=x1;
V=diag(m)-m*m`/n;
Gm[1,1]= m[7]; Gm[1,2]=-m[6]; Gm[1,6]=-m[2]; Gm[1,7]= m[1];
Gm[2,1]= m[8]; Gm[2,3]=-m[6]; Gm[2,6]=-m[3]; Gm[2,8]= m[1];
Gm[3,1]= m[9]; Gm[3,4]=-m[6]; Gm[3,6]=-m[4]; Gm[3,9]= m[1];
Gm[4,1]= m[12]; Gm[4,2]=-m[11]; Gm[4,11]=-m[2]; Gm[4,12]= m[1];
Gm[5,1]= m[13]; Gm[5,3]=-m[11]; Gm[5,11]=-m[3]; Gm[5,13]= m[1];
Gm[6,1]= m[17]; Gm[6,2]=-m[16]; Gm[6,16]=-m[2]; Gm[6,17]= m[1];
do j=1 to 5;
g1=x[1]#x[7]-x[2]#x[6];
g2=x[1]#x[8]-x[3]#x[6];
```

```
g3=x[1]#x[9]-x[4]#x[6];
g4=x[1]#x[12]-x[2]#x[11];
g5=x[1]#x[13]-x[3]#x[11];
g6=x[1]#x[17]-x[2]#x[16];
g=g1//g2//g3//g4//g5//g6;
Gx[1,1]= x[7]; Gx[1,2]=-x[6]; Gx[1,6]=-x[2]; Gx[1,7]= x[1];
Gx[2,1]= x[8]; Gx[2,3]=-x[6]; Gx[2,6]=-x[3]; Gx[2,8]= x[1];
Gx[3,1]= x[9]; Gx[3,4]=-x[6]; Gx[3,6]=-x[4]; Gx[3,9]= x[1];
Gx[4,1]= x[12]; Gx[4,2]=-x[11]; Gx[4,11]=-x[2]; Gx[4,12]= x[1];
Gx[5,1]= x[13]; Gx[5,3]=-x[11]; Gx[5,11]=-x[3]; Gx[5,13]= x[1];
Gx[6,1]= x[17]; Gx[6,2]=-x[16]; Gx[6,16]=-x[2]; Gx[6,17]= x[1];
x=x-(Gm*V)`*ginv(Gx*V*Gm`)*g;
nt=x[+];
end;
end;
print x m;
```

Example 4.9

This program allows the user to place constraints on the *frequencies*. The quantity "nh" still specifies the effects to be *included* in the design matrix.

```
proc iml worksize= 80;
*-------------  FREQUENCY VECTOR(slowest changing vector first);
*--->Structural zeros example;
*------------>; x={  64, 70,11,
                    83, 95, 0,
                     0,  0, 32};
xr =nrow(x);
x=x<>J(xr,1,1e-12);
*-------------  NUMBER OF VARIABLES ----;
*------------>; nf=2;
*-------------  NAMES OF VARIABLES -----;
*---------->; Name={"a.","b."};
*---------------------------------------------------------------;
k=j(6,1,0);
*----- NUMBER OF VARIABLES FOR EACH VARIABLE (MAX 6 VARIABLES)--;
*------->; k[1,]=3 ;k[2,]=3 ;k[3,]=0 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0 ;
*---------------------------------------------------------------;
*---------- SPECIFICATION OF HYPOTHESIS MATRIX AH --------------;
*index vector nh:effects to be INCLUDED in the design matrix  :
  1 A (I+A)B (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
```

```
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3};
*---------------------------------------------------------------;
*--------- SPECIFICATION OF THE CONSTRAINTS --------------------;
*--->; C1={ 0  0  0  0  0  1  0  0  0  ,
           0  0  0  0  0  0  1  0  0  ,
           0  0  0  0  0  0  0  1  0 };

cov=C1`;
*---------------------------------------------------------------;
*---------- CONSTRUCTION OF THE DESIGN MATRIX ------------------;
reset nolog;
reset fw=10;
c=k[1,];
one=J(c,1,1);
d=c-1;
A=(i(d)//J(1,d,-1));
e=k[1];
do i=2 to nf;
c=k[i,];
one=J(c,1,1);
d=c-1;
Y=(I(d)//J(1,d,-1));
one1=j(e,1,1);
A1=A@one;
Y1=one1@Y;
A=A1||Y1;
A=A||hdir(A1,Y1);
e=k[i,]*e;
end;
A=j(e,1,1)||A;
*---------- CONSTRUCTION OF THE INDEX VECTOR -------------------;
vg=1;
do i=1 to nf;
vg=vg//((k[i,]-1)*vg);
end;
kol=cusum(vg);
nrh=nrow(nh);
AA=J(e,1,1);
do i=1 to nrh;
```

```
ii=nh[i,]-1;iii=nh[i,];
a1=kol[ii,]+1;a2=kol[iii,];
AA=AA||A[,a1:a2];
end;  print AA[format=3.0];
*---------------------------------------------------------------;
A=AA||cov;
nc=ncol(A);
call svd(A,q,v,A);
Q1=J(nc,1,.001)><Q;
dfh=1000*sum(Q1); print dfh;
vgh=e-dfh;
A=A[,1:dfh];
AH=I(e)-A*ginv(A`*A)*A`;
*---------------------------------------------------------------;
*---------------------------------------------------------------;
x=x<>J(e,1,1e-6);
x1=1/x;
*----->;G1={0  0};
*---------------------------------------------------------------;
*-------------- HYPOTHESIS MATRIX WITH STRUCTURE ---------------;
sg=sum(g1*g1`);
if sg^=0 then AH=AH*G1`;
*---------------------------------------------------------------;
gx=AH`*log(x);
   m=x;
   gm=gx;
   itr=0;
   diff=1;
   do while (diff>0.000001);
   m1=m;
   mi=1/m;
   m=m-AH*ginv(AH`*(mi#AH))*gm;
m=m<>J(e,1,1e-12);
   gm=AH`*log(m);
   diff=sqrt((m-m1)`*(m-m1));
   itr=itr+1;
   end;
    mi=1/m;
vgh=e-dfh;
X2=(x-m)`*(mi#(x-m));
```

```
G2=2*x'*log(x/m);
K2ft=4*(sqrt(x)-sqrt(m))'*(sqrt(x)-sqrt(m));
Wald=gx'*ginv(AH'*(x1#AH))*gx;
vec=x2||g2||K2ft||Wald;
prob=J(1,4,1)-probchi(vec,vgh);
*PRINTING OF THE OUTPUT;
vec1={"Pearson" "LR" "F-T" "Wald" };
R={"Chi^2" "Df" "Prob"};
TEST=vec//J(1,4,vgh)//prob;
nrt=xr/k[nf];
x=shape(x,nrt);
m=shape(m,nrt);
print "------------LOG.IML-----------------";
print"number of iterations =" itr;
print" " ;
print x[format=7.1]  m[format=12.6] ;
print" ";
print "Chi-squared statistics with exact p-values";
print TEST[rowname=R colname=vec1 format=15.6];
```

Example 4.10

The program for this example is the same as Example 4.9, hence the relevant segment with input is given.

```
proc iml worksize= 50;
*-------------  FREQUENCY VECTOR(slowest changing vector first);
*---> Structural zeros example ;
*------------>; x={ 0 , 1029, 2240, 1413,
                   346, 548, 1287, 0 };
xr =nrow(x);
x=x<>J(xr,1,1e-12);
*-------------  NUMBER OF VARIABLES ----;
*------------>; nf=2;
*-------------  NAMES OF VARIABLES -----;
*--------->; Name={"a.","b."};
*-------------------------------------------------------------;
k=j(6,1,0);
*----- NUMBER OF VARIABLES FOR EACH VARIABLE (MAX 6 VARIABLES)--;
*------->; k[1,]=2 ;k[2,]=4 ;k[3,]=0 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0;
*-------------------------------------------------------------;
*---------- SPECIFICATION OF HYPOTHESIS MATRIX AH --------------;
```

```
*index vector nh:effects to be included in the design matrix  :
  1  A (I+A)B (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3};
*----------------------------------------------------------------;
*------------ SPECIFICATION OF THE CONSTRAINTS -----------------;
*--->; C1={ 1  0  0  0  0  0  0  0  ,
           0  0  0  0  0  0  0  1};
cov=C1`;
*----------------------------------------------------------------;
```

Example 4.11
This program is also the same as that of Example 4.9, hence only the relevant information to be supplied is listed.

```
proc iml worksize= 80;
*-------------  FREQUENCY VECTOR(slowest changing vector first);
*------------>; x={  4 , 2 , 9, 7,
                     0 , 0 , 4, 8,
                    42 , 7 ,19,10,
                    57 ,20 ,71,31};
xr =nrow(x);
x=x<>J(xr,1,1e-12);
*-------------  NUMBER OF VARIABLES ----;
*------------>; nf=3;
*-------------  NAMES OF VARIABLES -----;
*--------->; Name={"h.","a.","s."};
*----------------------------------------------------------------;
k=j(6,1,0);
*----- NUMBER OF VARIABLES FOR EACH VARIABLE (MAX 6 VARIABLES)--;
*------>; k[1,]=4 ;k[2,]=2 ;k[3,]=2 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0 ;
*----------------------------------------------------------------;
*--------- SPECIFICATION OF HYPOTHESIS MATRIX AH ---------------;
*index vector nh:effects to be included in the design matrix  :
  1 A (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2,3,4,5,6,7};
*----------------------------------------------------------------;
*--------- SPECIFICATION OF THE CONSTRAINTS --------------------;
*--->; C1={ 0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0,
           0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0};
```

```
cov=C1`;
*-------------------------------------------------------------;
```

CHAPTER 5

Example 5.1: Cumulative Logit Model.
This program can be used to fit any cumulative logit model with covariates. Care should be taken on setting up the covariates for the populations and must be entered in the *correct order* according to the frequency order of the variables. The response variable is the slowest changing variable.

```
proc iml worksize=80;
options pagesize=500;
*-----> FREQUENCY VECTOR;
*----->; x={  5 ,  7 , 13,
              7 ,  8 , 13,
              6 , 15 , 20,
              3 ,  9 , 13,
             53 , 115, 197,
             33 ,  65, 149,
             16 , 25 ,  96,
              7 , 23 ,56};
cov={ -1 ,
       0 ,
       1 ,
      -1 ,
       0 ,
       1 };
*-------------  NUMBER OF FACTORS -------;
*------------>; nf=3;
*-------------  NAMES OF FACTORS -------;
*--------->; name={"o.","l.","a."};
*--------->;namecov={"age"};
*-------------------------------------------------------------;
k=j(6,1,0);
*---------- NUMBER OF LEVELS OF EACH FACTOR (MAX 6 VARIABLES) --;
*------->; k[1,]=4 ;k[2,]=2 ;k[3,]=3 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0 ;
*-------------------------------------------------------------;
*-------------------------------------------------------------;
*--------- SPECIFICATION OF DESIGN MATRIX ---------------------;
*index vector nh for main effects and interactions to be included
```

```
  1 A (I+A)B (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD etc.;
*----------->;  nh={2};
*---------------------------------------------------------------;
*-- G1 for equal Language and Age parameters;
*G1={0  1  0  0 -1  0  0  0  0,
     0  0  0  0  1  0  0 -1  0,
     0  0  1  0  0 -1  0  0  0,
     0  0  0  0  0  1  0  0 -1};

*-- G1 for equal  Age parameters;
  G1={ 0  0  1  0  0 -1  0  0  0,
     0  0  0  0  0  1  0  0 -1};
 *G1=J(1,9,0);
*----------- CONSTRUCTION OF THE DESIGN MATRIX ----------------;
reset nolog;
reset fw=10;
c=k[2,];
one=J(c,1,1);
d=c-1;
A=(i(d)//J(1,d,-1));
e=k[2];
do i=3 to nf;
c=k[i,];
one=J(c,1,1);
d=c-1;
Y=(I(d)//J(1,d,-1));
one1=j(e,1,1);
A1=A@one;
Y1=one1@Y;
A=A1||Y1;
A=A||hdir(A1,Y1);
e=k[i,]*e;
end;
A=j(e,1,1)||A;
free Y Y1 A1;
nsubpop=e;
*---------- CONSTRUCTION OF THE INDEX VECTOR -------------------;
vg=1;
do i=2 to nf;
```

```
vg=vg//((k[i,]-1)*vg);
end;
tyd=name[2,1];
name1={""}//tyd;
do i=3 to nf;
name1=name1//concat(name1,name[i,1]);
end;
name1=rowcatc(name1);
nn=nrow(name1);
index={"mu"};
do i=2 to nn;
tyd=name1[i,1];
index=index//repeat(tyd,vg[i,1]);
end;
*--------------------------------------------------------------;
kol=cusum(vg);
nrh=nrow(nh);
AA=J(e,1,1);
index1=index[1,];
do i=1 to nrh;
ii=nh[i,]-1;iii=nh[i,];
a1=kol[ii,]+1;a2=kol[iii,];
AA=AA||A[,a1:a2];
index1=index1//index[a1:a2,];
end;
A=AA;
free AA;
*--------------------------------------------------------------;
nc=ncol(cov);
if any ( cov^=J(e,nc,0)) then do; A=A||cov;
index1=index1//namecov;
vg=vg//J(nc,1,1); end;
*--------- FULL DESIGN,HYPOTHESIS MATRIX AND INDEX VECTOR ------;
c=k[1,1];
d=c-1;
A=I(d)@A;
*%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
*For other Logit types replace the following segment with relevant;
*segment as described ;
*%%%%%%%%%%%%%%%%%%%%% FROM HERE %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
```

```
C1=J(d,c,0);
DO i=1 to d;
ki=c-i;
c1[i,]=J(1,i,1)||J(1,ki,0);
end;   print c1;
C2=J(d,c,1)-C1; print c2;
C1=C1@I(e);
C2=C2@I(e);
*%%%%%%%%%%%%%%%%%%%% TO HERE %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%;
lgtx=log(C1*x)-log(C2*x);
index=repeat(index1,d,1);
free  tyd name1 ;
A=A*inv(A'*A);
    x1=1/x;
e=e*d;
AH=I(e)-A*inv(A'*A)*A';
call svd(AH,Q,V,AH);
free Q V;
nca =ncol(A);
nch=e - nca;
nrg=nrow(G1);
AH=AH[,1:nch];
If any ( G1^=J(nrg,nca,0))
 then do;
  AH=AH||A*G1';
 nch = nch + nrg;
 end;
   Gx1= AH'*(diag(1/(C1*x))*C1 - diag(1/(C2*x))*C2);
   g=AH'*lgtx;
   m=x;
   x1=x;
   lgtm=lgtx;
   itr=0;
   diff1=1;
   i=0;
   do while (diff1>0.000001);
   i=i+1;
   m=x;
   x=x1;
   Gm= AH'*(diag(1/(C1*m))*C1 - diag(1/(C2*m))*C2);
```

```
    j=0;
    diff=1;
    do while (diff>0.000001);
    xv=x;
    j=j+1;
    Gx= AH'*(diag(1/(C1*x))*C1 - diag(1/(C2*x))*C2);
    g=AH'*(log(C1*x)-log(C2*x));
    x=x-(m'#Gm)'*inv(Gx*(m#Gm'))*g;
    diff=sqrt((x-xv)'*(x-xv));
    itr=itr+1;
    end;
    diff1=sqrt((m-x)'*(m-x));
    lgtm=log(C1*m)-log(C2*m);
  end;
  xr=nrow(x);
  lambda=A'*(log(C1*m)-log(C2*m));
    GA = A'*(diag(1/(C1*m))*C1 - diag(1/(C2*m))*C2);
  free A C1 C2;
  varl=vecdiag((m'#GA)*GA')-vecdiag(GA*(m'#Gx1)'*inv(Gx1*(m#Gx1'))*(m'#Gx1)*GA');
  vecvar=varl<>J(nca,1,1E-10);
  stdl=sqrt(vecvar);
  standl=lambda/stdl;
X2=(x1-m)'*((x1-m)/m);
G2=2*x1'*log(x1/m);
K2ft=4*(sqrt(x1)-sqrt(m))'*(sqrt(x1)-sqrt(m));
vec=x2||g2||K2ft;
prob=J(1,3,1)-probchi(vec,nch);
*PRINTING OF THE OUTPUT;
vec1={"Pearson" "LR" "F-T"  };
R={"Chi^2" "Df" "Prob"};
TEST=vec//J(1,3,nch)//prob;
xr=nrow(x);
nrt=xr/k[nf,];
x1=shape(x1,nrt);
m=shape(m,nrt);
lgtm=shape(lgtm,d);
lgtm =lgtm';
print "------------CUMULATIVE LOGIT TYPES-----------------";
namedep=name[1,1];
print "analysis of" namedep;
```

```
print"number of iterations=" itr; print" " ;
print x1[format=7.1] m[format=12.6] ;  print" ";
print  index lambda[format=12.6] standl[format=12.6];
print" ";
print "Chi-squared statistics with exceedance probabilities";
print TEST[rowname=R colname=vec1 format=15.6]; print" ";
aic =G2+2*(e-nch);
 print "Akaike Information Criterion ";
 print "AIC=" aic;
```

In order to run any of the other logits described in the paragraph make the alterations at the indicated place in the program above by adding one of the segments, depending on which logit model is to be considered.

The other options are as follows:

1. Adjacent category logits.

```
C1=J(d,c,0);
C1=J(d,1,0)||I(d);
C2=J(d,c,0);
C2=I(d)||J(d,1,0);
C1=C1@I(e);
C2=C2@I(e);
```

2. Continuation ratio logits (Method 1).
   $\ln(f1/f2+f3+f4)$, $\ln(f2/f3+f4)$, $\ln(f3/f4)$

```
C1=J(d,c,0);
C1=J(d,1,0)||I(d);
C2=J(d,c,0);
DO i=1 to d;
ki=c-i;
C2[i,]=J(1,i,1)||J(1,ki,0);
end;
C1=C1@I(e);
C2=C2@I(e);
```

3. Continuation ratio logits (Method 2).
   $\ln(f1/(f2+f3)),\ln(f2/f3)$

```
      C1=J(d,c,0);
      C1=I(d)||J(d,1,0);
      C2=J(d,c,0);
      DO i=1 to d;
      ki=c-i;
      C2[i,]=J(1,i,0)||J(1,ki,1);
      end;
      C1=C1@I(e);
      C2=C2@I(e);
```

Example 5.2

```
proc iml worksize=120;
*---- Proportional Hazards Model;
options pagesize=500;
*-----> frequency vector;
*----->; f={2309,2050,5397,4125,3315,2737,
            2138,1360,2333,1626,2044,1263,
            4011,2528,3505,3313,3472,2530,
            2794,1967,1506,2032,1105,1848,
             784,2475,645,2264,477,2397};
*--------------- COVARIATES -----------;
   cov=J(6,1,0);
*------------- NUMBER OF VARIABLES ----;
*------------>; nf=3;
*------------- NAMES OF VARIABLES -----;
*--------->; Name={"i.","y.","s."};
*--------->; Namecov={" "};
*---------------------------------------------------------------;
k=j(6,1,0);
*----------NUMBER OF LEVELS OF EACH VARIABLE (MAX 6 VARIABLES)--;
*-------->;k[1,]=5 ;k[2,]=3 ;k[3,]=2 ;k[4,]=0 ;k[5,]=0 ;k[6,]=0 ;
*---------------------------------------------------------------;
*---------------------------------------------------------------;
*------------- SPECIFICATION OF DESIGN MATRIX ----------------- ;
*index vector nh in the correct order:
  1 A (I+A)B  (I+(A+(I+A)B))C  (I+A+(I+A)B+(I+(A+(I+A)B)C)D  etc.
  =1 A B AB C AC BC ABC D AD BD ABD CD ACD BCD ABCD ens.;
*----------->;  nh={2,3};
*---------------------------------------------------------------;
*----------- MATRIX OF CONSTRAINTS ON PARAMETERS ---------------;
```

```
   G1={0  1  0  0  0 -1  0  0  0  0  0  0  0  0  0  0,
       0  0  0  0  0  1  0  0  0 -1  0  0  0  0  0  0,
       0  0  0  0  0  0  0  0  0  1  0  0  0 -1  0  0,
       0  0  1  0  0  0 -1  0  0  0  0  0  0  0  0  0,
       0  0  0  0  0  0  1  0  0  0 -1  0  0  0  0  0,
       0  0  0  0  0  0  0  0  0  0  1  0  0  0 -1  0,
       0  0  0  1  0  0  0 -1  0  0  0  0  0  0  0  0,
       0  0  0  0  0  0  0  1  0  0  0 -1  0  0  0  0,
       0  0  0  0  0  0  0  0  0  0  0  1  0  0  0 -1};
 *G1=J(1,16,0);
* --------- CONSTRUCTION OF DESIGN MATRIX A --------------------;
reset nolog;
reset fw=10;
c=k(|2,|);
one=J(c,1,1);
d=c-1;
A=(i(d)//J(1,d,-1));
e=k(|2|);
do i=3 to nf;
c=k(|i,|);
one=J(c,1,1);
d=c-1;
Y=(I(d)//J(1,d,-1));
one1=j(e,1,1);
A1=A@one;
Y1=one1@Y;
A=A1||Y1;
A=A||hdir(A1,Y1);
e=k(|i,|)*e;
end;
A=j(e,1,1)||A;
*---------- CONSTRUCTION OF THE INDEX (PARAMETER) VECTOR -------;
vg=1;
do i=2 to nf;
vg=vg//((k(|i,|)-1)*vg);
end;
tyd=name(|2,1|);
name1={""}//tyd;
do i=3 to nf;
name1=name1//concat(name1,name(|i,1|));
```

```
end;
name1=rowcatc(name1);
nn=nrow(name1);
index={"mu"};
do i=2 to nn;
tyd=name1(|i,1|);
index=index//repeat(tyd,vg(|i,1|));
end;
*-------------------------------------------------------------;
col=cusum(vg);
nrh=nrow(nh);
AA=J(e,1,1);
index1=index(|1,|);
do i=1 to nrh;
ii=nh(|i,|)-1;iii=nh(|i,|);
a1=col(|ii,|)+1;a2=col(|iii,|);
AA=AA||A(|,a1:a2|);
index1=index1//index(|a1:a2,|);
end;
A=AA;
free AA;
*-------------------------------------------------------------;
nc=ncol(cov);
if  any (cov^=J(e,nc,0)) then do; A=A||cov;
index1=index1//namecov;
vg=vg//J(nc,1,1); end;
AH=I(e)-A*inv(A`*A)*A`;
call svd(AH,Q,V,AH);
free Q V;
nca =ncol(A);
nch=e - nca;
AH=AH(|,1:nch|);
*-------------------------------------------------------------;
*-------------------------------------------------------------;
*-------- FULL DESIGN MATRIX, HYPOTHESIS MATRIX AND INDEX VECTOR;
c=k(|1,1|);
d=c-1;
A=I(d)@A;
C1 =j(d,c,0);
Do i= 1 to d;
```

```
ki=c-i;
c1(|i,|)=J(1,i,0)||J(1,ki,1);
end;
C1=C1@I(e);
index=repeat(index1,d,1);
*-------------------------------------------------------------;
free Y A1 Y1 tyd name1 ;
A=A*inv(A'*A);
e1=e;
 GG=J(1,c,1)@I(e1);
rows=gg*f;
irows = 1/rows;
p=diag(J(c,1,1)@irows)*f;
    x1=1/p;
e=e*d;
pr=nrow(p);
AH=I(d)@AH;
nca =ncol(A);
nrg=nrow(G1);
nch=e - nca;
 test =0;
If nch=0 then do;
   If any ( G1^=J(nrg,nca,0)) then do;
   AH=A*G1';
  end;
 else do ;
 AH=A*G1';
 test=1;
  end;
 end;
  If nch>0 then do;
  AH=AH(|,1:nch|);
  If any (G1^=J(nrg,nca,0)) then AH=AH||A*G1';
  end;
   m=p;
  If any (G1^=J(nrg,nca,0) ) then nch = nch+nrg;
   gm=AH'*log(-(log(C1*m)));
   gp=gm;
 nt=J(c,1,1)@rows;
 sig=diag(m/nt)-(gg*diag(m/nt))'*inv(gg*diag(m/nt)*gg')*(gg*diag(m/nt));
```

```
 sig1=sig;
 diff=1;
If test =0 then do;
Gx1=AH'*((1/log(C1*p)/(C1*p))#C1);
 GW=Gx1;
  x=p;
  m=p;
 x1=p;
 itr=0;
 diff1=1;
 i=0;
 do while (diff1>0.000001);
 i=i+1;
 m=x;
 x=x1;
   Gm = AH'*((1/log(C1*m)/(C1*m))#C1);
sig=diag(m/nt)-(gg*diag(m/nt))'*inv(gg*diag(m/nt)*gg')*(gg*diag(m/nt));
 j=0;
 diff=1;
 do while (diff>0.000001);
 xv=x;
 j=j+1;
   Gx = AH'*((1/log(C1*x)/(C1*x))#C1);
 g=AH'*(log(-log(C1*x)));
 x=x-(Gm*sig)'*ginv(Gx*sig*Gm')*g;
 diff=sqrt((x-xv)'*(x-xv));
 itr=itr+1;
 end;
 diff1=sqrt((m-x)'*(m-x));
end;
end;
lambda=A'*log(-(log(C1*m)));
   GA = A'*((1/log(C1*m)/(C1*m))#C1);
if test=0 then
sig=sig-(Gm*sig)'*inv(Gm*sig*Gm')*(Gm*sig);
*varl=vecdiag((m'#GA)*GA')-vecdiag(GA*(m'#G)'*ginv(G*(m#G'))*(m'#G)*GA');
varl=vecdiag(GA*sig*GA');
vecvar=varl<>J(nca,1,1E-10);
stdl=sqrt(vecvar);
standl=lambda/stdl;
```

```
if test=0 then do;
m1=diag(J(c,1,1)@rows)*m;
X2=(f-m1)‘*((f-m1)/m1);
G2=2*f‘*log(f/m1);
K2ft=4*(sqrt(f)-sqrt(m1))‘*(sqrt(f)-sqrt(m1));
Wald=gp‘*inv(GW*sig1*GW‘)*gp;
vec=x2||g2||K2ft||Wald;
prob=J(1,4,1)-probchi(vec,nch);
end;
vec1={"Pearson" "LR" "F-T" "Wald" };
R={"Chi^2" "Df" "Prob"};
Test=vec//J(1,4,nch)//prob;
xr=nrow(f);
nrt=xr/k[nf];
print "------------PROPHAZ.IML----------------";
print p [format=12.6] m[format =12.6];
print"number of iterations=" itr;
```

Example 5.3

```
proc iml;
* Logistic regression Dobson Cox & Snell p11;
reset nolog;
X={1   7  1.0,
   1   7  1.7,
   1   7  2.2,
   1   7  2.8,
   1   7  4.0,
   1  14  1.0,
   1  14  1.7,
   1  14  2.2,
   1  14  2.8,
   1  14  4.0,
   1  27  1.0,
   1  27  1.7,
   1  27  2.2,
   1  27  2.8,
   1  27  4.0,
   1  51  1.0,
   1  51  1.7,
   1  51  2.2,
```

```
   1  51  4.0};

yi={1e-6,1e-6,1e-6,1e-6,1e-6,1e-6,1e-6,2,1e-6,1e-6,
    1,4,1e-6,1,1,3,1e-6,1e-6,1e-6};
ni={10,17,7,12,9,31,43,33,31,19,56,44,21,22,16,13,1,1,1};
xr=nrow(X);
 e=j(xr,1,1);
 pi=yi/ni;
 logit=log(pi/(e-pi));
 var=ni#pi#(e-pi);
 ivar=1/var;
 SIG=diag(var);
 P=I(xr)-x*ginv(x`*x)*x`;
 G=P*diag(ivar);
    m=yi;
    diff=1;
    itr=0;
   do while (diff>0.00000001);
     m1=m;
     m=m-(G*SIG)`*ginv(G*SIG*G`)*P*logit;
     pi=m/ni;
     ri=pi/(e-pi);
     ww=ni#pi#(e-pi);
     iww=1/ww;
     G=P*diag(iww);
     SIG=diag(ww);
     logit=log(ri);
     diff=sqrt((m-m1)`*(m-m1));
     itr=itr+1;
   end;
 b=inv(x`*x)*x`*logit;
 A=inv(x`*x)*x`;
 covb=A*diag(iww)*A`;
 se=sqrt(vecdiag(covb));
 ei=(yi-m)#(yi-m)/m;
 chi2=sum(ei);
 print "Estimated Regression Parameters";
 print b se[format=12.6] ;  print " ";
 print itr;   print " ";
 print "Expected Frequencies";
```

```
 print m;   print " ";
 print "Chi-squared Statistic";
 print chi2;
*---- Calculate the Deviance;
 G2yi=2*yi`*log(yi/m);
 nmyi=ni-yi;
 nmyh=ni-m;
 g2nmyi=2*nmyi`*log(nmyi/nmyh);
 Deviance=G2yi+g2nmyi;
 print " ";
 print Deviance;
```

Example 5.4

```
proc iml;
* The extreme value distribution Dobson p.77;
reset nolog;
X={1 1.6907,
   1 1.7242,
   1 1.7552,
   1 1.7842,
   1 1.8113,
   1 1.8369,
   1 1.8610,
   1 1.8839};
yi={6,13,18,28,52,53,61,60};
ni={59,60,62,56,63,59,62,60.00000001};
e=j(8,1,1);
pi=yi/ni;
lx=log(-log(e-pi));
var=ni#pi#(e-pi);
S=diag(var);
P=I(8)-x*ginv(x`*x)*x`;
Z=X;
x1=yi;
m=yi;
d=(yi-ni#e)#log(e-pi);
di=1/d;
D=diag(di);
G=P*D;
do i=1 to 10;
```

```
x=x1;
do j=1 to 10 ;
pi=x/ni;
  d=(x-ni#e)#log(e-pi);
  di=1/d;
  D=diag(di);
G_x=P*D;
lx=P*log(-log(e-pi));

x=x - (G*S)`*ginv(G_x*S*G`)*lx;
x=x`;
print  x;
x=x`;
end;
m=x;
 pi=m/ni;
 ww=ni#pi#(e-pi);
   d=(m-ni#e)#log(e-pi);
 di=1/d;
  D=diag(di);
G=P*D;
S=diag(ww);
end;
b=inv(Z`*Z)*Z`*log(-log(e-pi));
A=inv(Z`*Z)*Z`;
covb=A*D*S*D*A`;
se=sqrt(vecdiag(covb));
print "Estimated Regression Parameters";
print " ";
print b se[format=12.6] ; print " ";
print "Expected Frequencies"; print " ";
print m;
*---- Calculate the Deviance;
G2yi=2*yi`*log(yi/m);
 nmyi=ni-yi;
 nmyh=ni-m;
g2nmyi=2*nmyi`*log(nmyi/nmyh);
Deviance=G2yi+g2nmyi;
print " ";
print Deviance;
```

CHAPTER 6

Example 6.1

```
proc iml worksize=50;
options pagesize=250;
reset nolog;
n=9;  r=5;
x={-1,-1,0,0,0,0,1,1,1};
x1=j(n,1,1)||x;
*------- ----->; y={ 2,3,6,7,8,9,10,12,15};
sig=diag(y);
b=inv(x1`*sig*x1)*x1`*sig*y;
print b;
yh=x1*b;
sig=diag(yh);
G=I(n)-x1*ginv(x1`*x1)*x1`;
    itr=0;
    diff=1;
    m=y;
   do while (diff>0.00000001);
     gm=G*m;
     m1=m;
     m=y-(G*SIG)`*ginv(G*SIG*G`)*G*y;
     SIG=diag(m);
     diff=sqrt((m-m1)`*(m-m1));
     itr=itr+1;
   end;
 b=inv(x1`*x1)*x1`*m;
 print b ;
 *print  SIG[format=5.2] SIGM[format=5.2];
 print y[format=5.2]  m[format=5.2];
 print itr;
```

Example 6.2

```
proc iml worksize=50;
options pagesize=200;
reset nolog;
n=14;
t={1,2,3,4,5,6,7,8,9,10,11,12,13,14};
t=log(t);
```

```
x=J(n,1,1)||t;
*------------>; y={ 1e-6,1,2,3,1,4,9,18,23,31,20,25,37,45};
sig=diag(y);
isig=inv(sig);
P=I(n)-x*ginv(x'*x)*x';
    itr=0;
    diff=1;
    m=y;
   do while (diff>0.00000001);
     m1=m;
     m=m-(P)'*ginv(P*isig*P')*P*log(m);
     sig=diag(m);
     isig=inv(sig);
     diff=sqrt((m-m1)'*(m-m1));
     itr=itr+1;
   end;
 b=inv(x'*x)*x'*log(m);
 print b ;
 print y[format=5.2]  m[format=5.2];
 print itr;
```

Example 6.6

```
 proc iml worksize=120;
 reset nolog;
*-normal-errors:SAS Report P-243(genmod) page 53;
y={5,7,9,7,10,8,11,9,16,13,14,25,24,34,32,30};
x={0,0,0,1,1,1,2,2,2,3,3,3,4,4,5,5,5};
n=nrow(x);
one=j(n,1,1);
x=one||x;
yi=1/y;
Dyi=diag(yi) ;
P=I(n)-x*ginv(x'*x)*x';
G=P*Dyi;
y1=y;
 do i = 1 to 10;
 y=y1;
   do j= 1 to 10;
yi=1/y;
Dyi=diag(yi);
```

```
gy=P*log(y);
G_y=P*Dyi ;
    y=y-(G)`*ginv(G_y*G`)*gy;
 end;
yi=1/y;
Dyi=diag(yi);
G=P*Dyi;
end;
b=inv(x`*x)*x`*log(y);
dev=(y1-y)`*(y1-y);
print "Estimated Parameters.";
print b ;
print "Deviance.";
print dev;
print "Observed an Predicted Values";
print y[format=12.6]  y1[format=12.6];
```

Example 6.7

```
 proc iml worksize=120;
 reset nolog;
*--gamma-errors:McCullagh and Nelder p307 using log link;

y={67.5,57.1,56.0,48.4,41.2,37.80,33.33,26.50,24.24,22.44,21.13,
   21.05,20.39,20.41,19.45,18.77,17.79,17.38,17.26,17.18,16.81,
   16.97,18.20};

 x={14.95  54,
    16.16  182,
    16.19  153,
    17.15  129,
    18.20   64,
    19.08   94,
    20.07   82,
    22.14   57,
    23.27   135,
    24.09  188,
    24.81  217,
    24.84  141,
    25.06   37,
    25.06   84,
```

```
    25.80  196,
    26.92  104,
    27.68  148,
    28.89   83,
    28.96   95,
    29.00  232,
    30.05  148,
    30.80  195,
    32.00   58};

n=nrow(x);
one=j(n,1,1);
temp=x[,1];
itemp=1/(temp-one#58.644);
wi=x[,2];
iwi=1/wi;
yi2=y#y;
x=one||temp||itemp;
ss=yi2#iwi;
P=I(n)-x*ginv(x`*x)*x`;
    itr=0;
    diff=1;
    m=y;
   do while (diff>0.00000001);
     mi=1/m;
     gm=P*log(m);
     G=P*diag(mi);
     m1=m;
     m=m-(ss`#G)`*ginv(G*(ss#G`))*gm;
     mi2=m#m;
     ss=mi2#iwi;
     diff=sqrt((m-m1)`*(m-m1));
     itr=itr+1;
   end;
 b=inv(x`*x)*x`*log(m);
 dev=-2*wi`*(log(y/m)-(y-m)#mi);
 print "Estimated Parameters.";
 print b ;  print " ";
 print "Deviance.";
 print dev;  print " ";
```

```
print "Observed an Predicted Values";
print y[format=12.6]  m[format=12.6];
print " ";
print "Number of Iterations.";
print itr;
```

Identity link

```
 proc iml worksize=120;
 reset nolog;
*--gamma-errors:McCullagh and Nelder p307 using identity link;

 y={67.5, ... ,18.20};
 x={14.95  54,
       .     .
       .     .
    32.00   58};

n=nrow(x);
one=j(n,1,1);
temp=x[,1];
itemp=1/(temp-one#0.6);
wi=x[,2];
iwi=1/wi;
yi2=y#y;
x=one||temp||itemp;
ss=yi2#iwi;
P=I(n)-x*ginv(x`*x)*x`;
    itr=0;
    diff=1;
    m=y;
   do while (diff>0.00000001);
     mi=1/m;
     gm=P*m;
     G=P;
     m1=m;
     m=m-(ss`#G)`*ginv(G*(ss#G`))*gm;
     mi2=m#m;
     ss=mi2#iwi;
     diff=sqrt((m-m1)`*(m-m1));
     itr=itr+1;
```

```
   end;
 b=inv(x'*x)*x'*m;
 dev=-2*wi'*(log(y/m)-(y-m)#mi);
print "Estimated Parameters.";
   (see preceding program)
```

Inverse link

```
 proc iml worksize=120;
 reset nolog;
*--gamma-errors:McCullagh and Nelder p307 using inverse link;

 y={67.5, ... ,18.20};
 x={14.95  54,
      .     .
      .     .
    32.00   58};

n=nrow(x);
one=j(n,1,1);
temp=x[,1];
itemp=1/(temp-one#33.5);
wi=x[,2];
iwi=1/wi;
yi2=y#y;
x=one||temp||itemp;
ss=yi2#iwi;
P=I(n)-x*ginv(x'*x)*x';
    itr=0;
    diff=1;
    m=y;
   do while (diff>0.00000001);
     mi=1/m;
     mi2=mi#mi;
     imi2=1/mi2;
     ss=mi2#iwi;
     gm=P*mi;
     G=P;
     m1=m;
     mi=mi-(ss'#G)'*ginv(G*(ss#G'))*gm;
     m=1/mi;
```

```
    diff=sqrt((m-m1)`*(m-m1));
    itr=itr+1;
  end;

 b=inv(x`*x)*x`*mi;
 Dev=-2*wi`*(log(y/m)-(y-m)#mi);
 print "Estimated Parameters.";
      (see preceding program)
```

Example 6.8

```
proc iml;
* Non-linear parameter in the covariate: McCullagh p385;
reset nolog;

yi={7,59,115,149,178,229,5,43,76,4,57,83,6,57,84};
ni={100,200,300,300,300,300,100,100,100,100,100,100,100,100,100};
dosein={4,5,8,10,15,20,2,5,10,2,5,10,2,5,10};
dosesy={0,0,0,0,0,0,3.9,3.9,3.9,19.5,19.5,19.5,39,39,39};
e=j(15,1,1);
pi=yi/ni;
logit=log(pi/(e-pi));
var=ni#pi#(e-pi);
ivar=1/var;
SIG=diag(var);
 delta=1.7;
x1=log(dosein);
den=(delta#e+dosesy);
x2=dosesy/den;
den2=(delta#e+dosesy)#den;
x3=-dosesy/den2;
x=e||x1||x2||x3;
P=I(15)-x*ginv(x`*x)*x`;
G=P*diag(ivar);
  m=yi;
  diff=1;
  itr=0;
   do while (verskil>0.00000001);
  m1=m;
  m=m-(G*SIG)`*ginv(G*SIG*G`)*P*logit;
  pi=m/ni;
```

```
  ri=pi/(e-pi);
 ww=ni#pi#(e-pi);
 iww=1/ww;
SIG=diag(ww);
 logit=log(ri);
   diff=sqrt((m-m1)`*(m-m1));
   itr=itr+1;
b=inv(x`*x)*x`*logit;
 delta=b[4,]/b[3,]+delta;  print delta;
x1=log(dosein);
den=(delta#e+dosesy);
x2=dosesy/den;
den2=(delta#e+dosesy)#den;
x3=-dosesy/den2;
x=e||x1||x2||x3;
P=I(15)-x*ginv(x`*x)*x`;
G=P*diag(ivar);
   end;
print  delta ;
b=inv(x`*x)*x`*logit;
A=inv(x`*x)*x`;
covb=A*diag(iww)*A`;
se=sqrt(vecdiag(covb));
print "Estimated Regression Parameters";
print " ";
print b se[format=12.6] ;  print " ";
print itr; print " ";
print "Expected Frequencies"; print " ";
print m;
*---- Calculate the Deviance;
G2yi=2*yi`*log(yi/m);
 nmyi=ni-yi;
 nmyh=ni-m;
g2nmyi=2*nmyi`*log(nmyi/nmyh);
Deviance=G2yi+g2nmyi;
print " ";
print Deviance;
```