NYIKOS, HELENA

MULTILEVEL MODELS FOR THE ANLYSIS OF ORDINAL DATA

MSc **UP** **1994**

MULTILEVEL MODELS FOR THE ANALYSIS OF ORDINAL DATA

by

HELENA NYIKOS

Submitted in part fulfilment of the requirements for the degree of Master of Science in the subject Mathematical Statistics

Faculty of Science
University of Pretoria
Pretoria

July 1994

ACKNOWLEDGEMENTS

The guidance of my supervisor, Professor S.H.C. du Toit is gratefully appreciated.

I wish to thank Mrs. M. van Staden for her patience in typing this manuscript.

I also wish to thank my family and friends for their support and encouragement.

Contents

Notation

The following notation shall be adopted:

π	:	constant, $\pi = 3.14159...$
e	:	Euler's constant, $e = 2.71828...$
$\exp(x)$	:	$e^x, -\infty < x < \infty$
$\ell n\ x$	:	natural logarithm of the real number x, $x \geq 0$
δ_{ij}	:	Kronecker's delta (1 if $i = j$ and 0 if $i \neq j$)
$\mathbf{A} : (p \times q)$	:	matrix of order $p \times q$
$\mathbf{a} : (p \times 1)$	:	column vector of order $p \times 1$
a	:	scâlar
$\mathbf{A}'$	:	transpose of $\mathbf{A}$
$\mathbf{a}'$	:	transpose of $\mathbf{a}$ (a row vector)
a_{ij} or $[A]_{ij}$	:	the element in the i-th row and j-th column of $\mathbf{A}$
a_i or $[\mathbf{a}]_{i1}$	:	the i-th element of $\mathbf{a}$
$\mathbf{A}^{-1}$	:	inverse of $\mathbf{A}$
a^{ij}	:	$[\mathbf{A}^{-1}]_{ij}$
$\|\mathbf{A}\|$	:	determinant of $\mathbf{A}$
$tr[\mathbf{A}]$	:	trace of $\mathbf{A}$
$\mathbf{D}_a$	:	diagonal matrix with diagonal elements $a_{11}, a_{22}, \cdots$
Diag[$\mathbf{A}$]	:	diagonal matrix formed from the diagonal elements of $\mathbf{A}$
diag [$\mathbf{A}$]	:	columnvector formed from the diagonal elements of $\mathbf{A}$
vec[$\mathbf{A}$]	:	$(pq \times 1)$ vector formed from the q columns of the $p \times q$ matrix $\mathbf{A}$
vecs[$\mathbf{A}$]	:	$(p(p+1)/2 \times 1)$ vector formed from the nonduplicated elements of the $(p \times p)$ symmetric matrix $\mathbf{A}$
$\mathbf{0}$	:	null matrix, $[\mathbf{0}]_{ij} = 0$
$\mathbf{0}$ or $\mathbf{0}_p$	:	$(p \times 1)$ null vector, $[\mathbf{0}]_{i1} = 0$
$\mathbf{j}$ or $\mathbf{j}_p$	:	$(p \times 1)$ vector with unit elements, $[\mathbf{j}]_{i1} = 1$
$\mathbf{I}$ or $\mathbf{I}_p$	:	$(p \times p)$ identity matrix
$\mathbf{J}_{ij}$	:	matrix with all elements equal to zero with the exception of the element in the i-th row and j-th column which is equal to unity

$\mathbf{J}_{i1}$	:	column vector with all elements equal to zero with the exception of the i-th element which is equal to unity
$\mathbf{A} \otimes \mathbf{B}$	:	The right direct product or "Kronecker product " of matrices $\mathbf{A}$ and $\mathbf{B}$ defined by: $\begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1q}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2q}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1}\mathbf{B} & a_{p2}\mathbf{B} & \cdots & a_{pq}\mathbf{B} \end{pmatrix}$
$\frac{\partial \mathbf{A}}{\partial x}$	:	matrix with typical element $\frac{\partial a_{ij}}{\partial x}$
$\frac{\partial \mathbf{a}}{\partial x}$	:	column vector with typical element $\frac{\partial [\mathbf{a}]_{i1}}{\partial x}$
$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$	:	column vector with typical element $\frac{\partial f(\mathbf{x})}{\partial [\mathbf{x}]_{i1}}$
$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'}$	:	symmetric matrix with typical element $\frac{\partial^2 f(\mathbf{x})}{\partial [\mathbf{x}]_{i1} \partial [\mathbf{x}]_{j1}}$
$E(\mathbf{y}) : (p \times 1)$	:	expected value of the random vector $\mathbf{y}$ with typical element $E(y_i)$
$\mathrm{Cov}(\mathbf{y}, \mathbf{y}') : (p \times p)$	:	covariance matrix of the random vector $\mathbf{y}$ with typical element $E[y_i - E(y_i)]\ [y_j - E(y_j)]$
$\mathrm{Cov}(\mathbf{Y}) : (Np \times Np)$	:	covariance matrix of $\mathrm{Vec}(\mathbf{Y})$ with $\mathbf{Y} : (N \times p)$

CHAPTER 1

1 Introduction

Multilevel modelling has a wide application in the social sciences. The term multilevel refers to a hierarchical relationship among units in a system. In an education system, for example, students are members of classes, and classes are grouped within schools. We regard students as level 1 units, classes as level 2 units and schools as level 3 units.

Multilevel analysis allows characteristics of each group (for example the students of a specific class of a specific school) to be incorporated into models of individual behaviour, while also producing correct estimates of standard errors so that valid tests and intervals can be constructed.

In Chapter 2 general multilevel theory is discussed (Du Toit, 1993). The fixed parameter linear regression model is extended to a random parameter linear regression model. A general expression for the two-level model is obtained and extended to the general three-level model. Estimation procedures for the unknown parameters are discussed, in particular the method of iterative generalized least squares. The multilevel logit model is illustrated with an example. In this example use was made of the ML3 package (Prosser, Rashbash and Goldstein, 1989 and 1990). A description of this package is given in Appendix A2 and the specific program used is given in Appendix A3. A SAS program, also implemented in this example, is given in Appendix A1.

Chapter 3 deals with models for analysing data with an ordinal response variable. The logit, cumulative logit and McCullagh's proportional odds model are discussed (Du Toit and Lampbrecht, 1984; McCullagh, 1980). These models are illustrated with a practical application.

The theory of Chapter 2 and 3 are combined in Chapter 4. The three models for analysing data with an ordinal dependent variable are described in the context of multilevel theory. The emphasis is on the iterative procedure used to obtain estimates of the unknown parameters. This procedure used in this new modelling approach is illustrated with an example.

In Chapter 5 conclusions are drawn and suggestions for further research are made.

The theory discussed in Chapter 3 and 4 has been implemented in SAS computer programs. These programs are given in Appendix B and C.

Throughout the study use was made of a dataset obtained by the Human Sciences Research Council. This dataset was obtained through a project with the the title "The prospects for a free, democratic election" (De Kock, 1993).

CHAPTER 2

2 Multilevel Theory

2.1 Fixed parameter linear regression models

Consider a sample of N scholars which were taken from various schools. Suppose the relationship between a scholar's number series ability (x) and midyear mathematics score (y) is to be investigated. This relationship for scholar i is usually described by the model

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, 2, \cdots, N \tag{2.1.1}$$

For the total of N scholars (2.1.1) can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

that is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

It is usually assumed that $E(\mathbf{e}) = \mathbf{0}$ and $\text{Cov}(\mathbf{e}, \mathbf{e}') = \sigma^2 \mathbf{I}_N$. Under this assumption the fixed parameters, β_0 and β_1, are estimated using ordinary least squares (OLS) estimation. The OLS estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

with

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \text{ and } \text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}') = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

2.2 Repeated measurement fixed parameter models

Suppose that each experimental unit i has n responses. The j-th response for experimental unit i can adequately be described by the following linear regression model

$$y_{ij} = \mathbf{x}_j'\boldsymbol{\beta} + e_{ij} \qquad \begin{aligned} i &= 1, 2, \cdots, N \\ j &= 1, 2, \cdots, n \end{aligned} \tag{2.2.1}$$

An example of the model (2.2.1) is

$$\begin{aligned} \mathbf{x}_j'\boldsymbol{\beta} &= \beta_0 + \beta_1 t_j + \beta_2 t_j^2 \ , \\ \mathbf{x}_j &= (1, t_j, t_j^2)' \end{aligned}$$

The set of regression equations (2.2.1) for experimental unit i can also be written in matrix notation as

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_i \ , \ \ i = 1, 2, \cdots, N \tag{2.2.2}$$

The j-th row of the $n \times m$ design matrix $\mathbf{X}$ is the $1 \times m$ row vector $\mathbf{x}_j'$. It is further assumed that the vectors of error variates $\mathbf{e}_i$, $i = 1, 2, \cdots, N$ are identically and independently distributed with

$$E(\mathbf{e}_i) = \mathbf{0}, \ \text{Cov}(\mathbf{e}_i, \ \mathbf{e}_i') = \boldsymbol{\Sigma}$$

Different assumptions about the structure of $\boldsymbol{\Sigma}$ lead to the use of different estimators of the unknown parameters $\boldsymbol{\beta}$. Different estimators are obtained under the following conditions

(i) $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$

The ordinary least squares estimate $\hat{\boldsymbol{\beta}}_{OLS}$ of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\bar{\mathbf{y}} \ , \tag{2.2.3}$$

with

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i$$

(ii) $\boldsymbol{\Sigma}$ known

Under the assumption of multivariate normality the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{ML}$ of $\boldsymbol{\beta}$ is obtained, that is

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{y}} \tag{2.2.4}$$

(iii) $\boldsymbol{\Sigma}$ unknown and no structure imposed on the elements of $\boldsymbol{\Sigma}$

In this case $\boldsymbol{\Sigma}$ is replaced in (2.2.4) by its unbiased estimator S, where

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' \qquad (2.2.5)$$

and the generalized least squares estimator $\hat{\beta}_{GLS}$ of β is obtained, that is

$$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}^{-1}\bar{\mathbf{y}} \qquad (2.2.6)$$

(iv) $\mathbf{\Sigma} = \mathbf{\Sigma}(\gamma)$, where $\mathbf{\Sigma}(\gamma)$ is a structured matrix, for example

$$\begin{aligned}[\mathbf{\Sigma}]_{k,k} &= \sigma^2 \qquad &, k = 1, 2, \cdots, n\\ [\mathbf{\Sigma}]_{k,\ell} &= \rho \qquad &, k \neq \ell\end{aligned}$$

Now $\gamma = (\sigma^2, \rho)'$. The maximum likelihood estimates of β and γ are obtained as the solution of the likelihood equations

$$\frac{\partial \ell n L}{\partial \beta} = \mathbf{0}$$

$$\frac{\partial \ell n L}{\partial \gamma} = \mathbf{0} \ ,$$

where L denotes the appropriate likelihood function of $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$. In general, closed form solutions to the likelihood equations cannot be obtained. In this case an iterative procedure is used.

2.3 Random parameter repeated measurements models

Suppose that for model (2.2.2) it is more realistic to assume that the regression coefficients β vary from one experimental unit to another. A way to accommodate for this assumption, is to regard the unknown regression parameters as random variables. Thus redefining (2.2.2) the model

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{e}_i \ \ , i = 1, 2, \cdots, N \ , \qquad (2.3.1)$$

is obtained for experimental unit i. $\mathbf{X}$ is $(n \times m)$, $E(\mathbf{b}_i) = \beta$ and $\text{Cov}(\mathbf{b}_i \ , \mathbf{b}_i') = \mathbf{\Phi}$. It is usually assumed that $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_N$ are a random sample from a multivariate normal distribution and that $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N$ are an independent random sample from a $\mathbf{N}(\mathbf{0}, \mathbf{\Lambda})$ distribution.

Under the above assumptions, $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$ is a row of independently and identically distributed $\mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\boldsymbol{\Phi}\mathbf{X}' + \boldsymbol{\Lambda})$ variables.

The likelihood function of $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$ can be written as (Browne, 1991):

$$L = (2\pi)^{-Nn/2}|\Sigma|^{-N/2} \ \exp \ \left\{-\frac{N}{2}\mathrm{tr}\Sigma^{-1}\mathbf{G}\right\}, \qquad (2.3.2)$$

where

$$\boldsymbol{\Sigma} = \mathbf{X}\boldsymbol{\Phi}\mathbf{X}' + \boldsymbol{\Lambda}$$

and

$$\mathbf{G} = \frac{(N-1)\mathbf{S}}{N} + (\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})(\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})',$$

with $\bar{\mathbf{y}} = \frac{1}{N}\Sigma\mathbf{y}_i$ and S defined by (2.2.5).

Suppose that

$$\boldsymbol{\Lambda} = \sigma^2\mathbf{I}$$

and therefore that

$$\boldsymbol{\Sigma} = \mathbf{X}\boldsymbol{\Phi}\mathbf{X}' + \sigma^2\mathbf{I}$$

Let

$$\gamma = \begin{pmatrix} \boldsymbol{\beta} \\ \text{vecs } \boldsymbol{\Phi} \\ \sigma^2 \end{pmatrix}$$

It was shown (see e.g. Browne and du Toit, 1992) that

$$\frac{\partial \ell n L}{\partial \gamma_i} = \frac{N}{2} \ \mathrm{tr} \ \mathbf{P}\frac{\partial \Sigma}{\partial \gamma_i} + N \ \mathrm{tr} \ \mathbf{R}\frac{\partial \mathbf{X}\boldsymbol{\beta}}{\partial \gamma_i} \qquad (2.3.3)$$

where $\mathbf{P} = \Sigma^{-1}(\mathbf{G} - \Sigma)\Sigma^{-1}$
and $\mathbf{R} = (\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'\Sigma^{-1}$. From expression (2.3.3) maximum likelihood (ML) estimators of

$\boldsymbol{\beta}, \boldsymbol{\Phi}$ and σ^2 can be obtained.

Let

$$\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and

$$\mathbf{W} = \frac{(N-1)}{N}\mathbf{S}$$

It then follows that these maximum likelihood estimators are

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\bar{\mathbf{y}} \\ \hat{\mathbf{\Phi}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{W}-\hat{\sigma}^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ \hat{\sigma}^2 &= (n-m)^{-1}\, tr[(\mathbf{I}-\mathbf{Q})(\mathbf{W}+\bar{\mathbf{y}}\bar{\mathbf{y}}')]\end{aligned}$$

2.4 Marginal maximum likelihood and the E-M algorithm

Suppose that the experimental units in (2.3.1) do not have an equal number of responses. That is, experimental unit i has n_i responses.

The assumption of an unequal number of measurements for each experimental unit is incorporated in the random parameter regressions model by allowing for different design matrices for different individuals. Let

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{e}_i\ , \quad i = 1, 2, \cdots, N \ , \tag{2.4.1}$$

where $\mathbf{X}_i$ is a $n_i \times m$ design matrix and where it is assumed that

$$E(\mathbf{b}_i) = \beta, \quad \mathrm{Cov}(\mathbf{b}_i, \mathbf{b}_i') = \mathbf{\Phi} \tag{2.4.2}$$

$$E(\mathbf{e}_i) = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{e}_i, \mathbf{e}_i') = \mathbf{\Lambda}_i \tag{2.4.3}$$

$$\mathrm{Cov}(\mathbf{b}_i, \mathbf{e}_i') = \mathbf{0} \tag{2.4.4}$$

Denote the $n_i \times n_i$ covariance matrix of $\mathbf{y}_i$ by $\mathbf{\Sigma}_i$, then it follows from (2.4.2) to (2.4.4) that

$$\mathbf{\Sigma}_i = \mathbf{X}_i\mathbf{\Phi}\mathbf{X}_i' + \mathbf{\Lambda}_i$$

If $E(\mathbf{y}_i)$ is denoted by $\boldsymbol{\xi}_i$, then

$$\boldsymbol{\xi}_i = \mathbf{X}_i\beta$$

If, as in the previous section, we assume that $\mathbf{b}_i$ is multivariate normal and that $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_N$ are a row of independently distributed normal random deviates, then

$$\mathbf{y}_i \sim N(\boldsymbol{\xi}_i, \mathbf{\Sigma}_i),$$

$$L(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N) = \prod_{i=1}^{N} (2\pi)^{-n_i/2} |\mathbf{\Sigma}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\xi}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\xi}_i)\right\} \tag{2.4.5}$$

From (2.4.5) it further follows (with the constant terms $\ell n 2\pi$ omitted) that:

$$\ell n L = -\frac{1}{2}\sum_{i=1}^{N} \ell n|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{N}(tr\Sigma_i^{-1}\mathbf{G}_i) ,$$

where

$$\mathbf{G}_i = (\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)'$$

Extending the results of Browne and Du Toit (1992) it follows that

$$\frac{\partial \ell n L}{\partial \gamma_\ell} = \frac{1}{2}\sum_{i=1}^{N} tr\ \mathbf{P}_i\frac{\partial \Sigma_i}{\partial \gamma_\ell} + \sum_{i=1}^{N} tr\ \mathbf{R}_i\frac{\partial \boldsymbol{\xi}_i}{\partial \gamma_\ell}, \qquad (2.4.6)$$

where

$$\begin{aligned} \mathbf{P}_i &= \Sigma_i^{-1}(\mathbf{G}_i - \Sigma_i)\Sigma_i^{-1} \ , \\ \mathbf{R}_i &= (\mathbf{y}_i - \boldsymbol{\xi}_i)'\Sigma_i^{-1} \end{aligned}$$

and where γ_ℓ denotes a typical element of the $(k \times 1)$ vector of unknown elements.

Usually it assumed that there is a parameter vector, say $(\boldsymbol{\tau})$, common to $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \cdots, \boldsymbol{\Lambda}_N$. That is

$$\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_i(\boldsymbol{\tau})$$

To obtain ML estimates of $\boldsymbol{\beta}, \boldsymbol{\Phi}$ and $\boldsymbol{\tau}$ an iterative procedure is required. A modified version of the computer program AUFIT (see Du Toit & Browne, 1982)) was written by du Toit (1991) to obtain these estimates. This estimation procedure involves the inversion of $n_i \times n_i$ matrices. For large values of n_i, the iterative procedure may become very time consuming. An alternative method for obtaining ML estimates was suggested by Bock (1990). This method is called the marginal maximum likelihood (MML) method. This method is briefly as follows

Denote the conditional density of $\mathbf{b}_i$ given $\mathbf{y}_i$ by $p(\mathbf{b}|\mathbf{y}_i)$ The following expressions are obtained

$$E(\mathbf{b}_i|\mathbf{y}_i) = (\mathbf{X}_i'\boldsymbol{\Lambda}_i^{-1}\mathbf{X}_i + \boldsymbol{\Phi}^{-1})^{-1}\mathbf{X}_i'\boldsymbol{\Lambda}_i^{-1}[\mathbf{y}_i - \mathbf{X}_i\beta] + \beta \qquad (2.4.7)$$

and

$$\text{Cov}(\mathbf{b}_i|\mathbf{y}_i) = (\mathbf{X}_i'\boldsymbol{\Lambda}_i^{-1}\mathbf{X}_i + \boldsymbol{\Phi}^{-1})^{-1} \qquad (2.4.8)$$

It is assumed that $\boldsymbol{\Lambda}_i = \sigma^2\mathbf{I}_{n_i}$. Expressions for the MML estimations are obtained as

$$\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N} E(\mathbf{b}|\mathbf{y}_i) . \qquad (2.4.9)$$

$$\hat{\mathbf{\Phi}} = \frac{1}{N}\sum_{i=1}^{N}\{\text{Cov }(\mathbf{b}|\mathbf{y}_i) + (E\mathbf{b}|\mathbf{y}_i - \hat{\boldsymbol{\beta}})(E\mathbf{b}|\mathbf{y}_i - \hat{\boldsymbol{\beta}})'\} \tag{2.4.10}$$

and

$$\begin{aligned}\hat{\sigma}^2 &= \left(\sum_{i=1}^{N} n_i\right)^{-1} \sum_{i=1}^{N} [(\mathbf{y}_i - \mathbf{X}_i E\mathbf{b}|\mathbf{y}_i)'(\mathbf{y}_i - \mathbf{X}_i E\mathbf{b}|\mathbf{y}_i) \\ &\quad + tr\mathbf{X}_i'\mathbf{X}_i \text{ Cov }(\mathbf{b}|\mathbf{y}_i)]\end{aligned} \tag{2.4.11}$$

All these expressions depend on $E(\mathbf{b}|y_i)$ and Cov$(\mathbf{b}|\mathbf{y}_i)$ By initially setting, for example, $\sigma^2 = 1, \boldsymbol{\beta} = \mathbf{0}$ and $\mathbf{\Phi} = \mathbf{0}$, initial estimates of $E(\mathbf{b}|y_i)$ and Cov$(\mathbf{b}|\mathbf{y}_i)$ may be obtained from (2.4.7) and (2.4.8). These estimates may then be used in (2.4.9), (2.4.10) and (2.4.11) to solve for $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{\Phi}}$ and $\hat{\sigma}^2$ respectively. These new values are then substituted into (2.4.7) and (2.4.8) and the procedure repeated until convergence is attained. This optimization algorithm is called expected maximization (EM-algorithm). Close approximations to the marginal maximum likelihood estimates of the parameters are obtained.

2.5 Two-level Models

Suppose that from N schools each data have been collected.That is, from each school n_i $(i = 1, 2, \cdots, N)$ scholars were included in the sample. The relationship between a scholar's number series ability (x) and midyear mathematics score (y) is to be investigated. For school i, a linear relationship between these variables can be written as follows

$$y_{ij} = b_{0i} + b_{1i}x_{ij} + e_{ij} \qquad i = 1, \cdots, N \;\; j = 1, \cdots, n_i. \tag{2.5.1}$$

In (2.5.1) y_{ij} and x_{ij} are the midyear mathematics and number series ability scores, respectively, for student j in school i. The intercept (b_{0i}) and the gradient (b_{1i}) varies between schools. Scholars are regarded as level 1 units, and schools as level 2 units. Using the notation of Goldstein and McDonald (1988), b_{0i} and b_{1i} can be expressed as

$$\begin{aligned} b_{0i} &= \beta_0 + u_{0i} \\ b_{1i} &= \beta_1 + u_{1i} \end{aligned} \tag{2.5.2}$$

and are therefore random variables at level 2 (school level). In matrix notation (2.5.1) for school i becomes

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{e}_i \tag{2.5.3}$$

that is

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix} \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}$$

The following assumptions are made:

(a) $\mathbf{X}_i$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ are non-random

(b) $\mathbf{e}_i$ has expected value $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}_{n_i}$

(c) $\mathbf{u}_i = \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix}$ has expected value $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}$

(d) $\text{Cov}(\mathbf{e}_i, \mathbf{u}_i') = \mathbf{0}$

Multivariate normality is usually assumed for $\mathbf{e}_i$ and $\mathbf{u}_i$.

To include certain characteristics of the schools, (2.5.2) can be extended. Suppose z_i is the percentage of scholars in school i whose parents are classified as having a high socio-economic status according to some definition. Now

$$b_{0i} = \beta_0 + \gamma_{01} z_i + u_{0i}$$

$$b_{1i} = \beta_1 + \gamma_{11} z_i + u_{1i}$$

and in matrix notation

$$\mathbf{b}_i = \boldsymbol{\beta} + \boldsymbol{\gamma} z_i + \mathbf{u}_i \quad \text{with} \quad \boldsymbol{\gamma} = \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \end{bmatrix}$$

In general if measurements were made on $(m-1)$ explanatory variables $x_1, \cdots, x_{m-1}$ (number series ability, pattern completion ability etc.) and on q level 2 related variables $z_{i1}, \cdots, z_{iq}$

$$\mathbf{b}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma} \mathbf{z}_i + \mathbf{u}_i \qquad (2.5.4)$$

where

$\boldsymbol{\beta}$ is an $(m \times 1)$ vector
$\boldsymbol{\Gamma}$ is an $(m \times q)$ matrix of fixed unknown coefficients
$\mathbf{z}_i$ is a $(q \times 1)$ vector of level 2 variables
$\mathbf{u}_i$ is an $(m \times 1)$ random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Phi}$

If (2.5.3) and (2.5.4) are combined then

$$\mathbf{y}_i = \mathbf{X}_i(\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{z}_i + \mathbf{u}_i) + \mathbf{e}_i \tag{2.5.5}$$

All the unknown fixed parameters of (2.5.5) can be combined into a single vector

$$\gamma^\star = \begin{bmatrix} \beta_0 \\ \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0q} \\ \vdots \\ \beta_{m-1} \\ \gamma_{m-1,1} \\ \gamma_{m-1,2} \\ \vdots \\ \gamma_{m-1,q} \end{bmatrix}$$

Now (2.5.5) can be written as

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{Z}_i\gamma^\star + \mathbf{X}_i\mathbf{u}_i + \mathbf{e}_i$$

where $\mathbf{Z}_i$ is an $m \times (m + mq)$ between-unit design matrix. The above model can again be rewritten as

$$\mathbf{y}_i = \mathbf{W}_i\gamma^\star + \mathbf{X}_{(2)i}\mathbf{u}_i + \mathbf{X}_{(1)i}\mathbf{e}_i \tag{2.5.6}$$

where $\mathbf{X}_{(2)}$ denotes the matrix of explanatory variables whose coefficients are random at level 2, $\mathbf{X}_{(1)}$ denotes the matrix of explanatory variables whose coefficients are random at level 1 and $\mathbf{W}_i = \mathbf{X}_i\mathbf{Z}_i$. Expression (2.5.6) is a general expression for the two-level model which allows for a complex level 1 covariance structure. Under the general assumptions given it follows that

$$E(\mathbf{y}_i) = \mathbf{W}_i\gamma^\star$$

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \mathbf{X}_{(2)_i}\boldsymbol{\Phi}_{(2)}\mathbf{X}_{(2)_i}' + \mathbf{X}_{(1)_i}\boldsymbol{\Phi}_{(1)}\mathbf{X}_{(1)_i}'$$

where

$$\boldsymbol{\Phi}_{(1)} = \text{Cov}(\mathbf{e}_i, \mathbf{e}_i') \qquad \boldsymbol{\Phi}_{(2)} = \text{Cov}(\mathbf{u}_i, \mathbf{u}_i')$$

2.6 Maximum likelihood estimation

The general two-level model is given as

$$\mathbf{y}_i = \mathbf{W}_i\gamma^\star + \mathbf{X}_{(2)i}\mathbf{u}_i + \mathbf{X}_{(1)i}\mathbf{e}_i$$

Denote the expected value and covariance matrix of $\mathbf{y}_i$ by $\boldsymbol{\xi}_i$ and $\boldsymbol{\Sigma}_i$, respectively. The log-likelihood function of $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$ may then be expressed as

$$\ell nL = -\frac{1}{2}\sum_{i=1}^{N}\left\{n_i\ell n2\pi + \ell n|\boldsymbol{\Sigma}_i| + tr\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)'\right\}$$

Maximum normal likelihood estimates of the unknown parameters are obtained by minimizing $-2\ell nL$ with the constant term omitted yielding the discrepancy function

$$F(\gamma) = \sum_{i=1}^{N}\left\{\ell n|\boldsymbol{\Sigma}_i| + tr\boldsymbol{\Sigma}_i^{-1}\mathbf{G}_{y_i}\right\},$$

where

$$\mathbf{G}_{y_i} = (\mathbf{y}_i - \boldsymbol{\xi}_i)(\mathbf{y}_i - \boldsymbol{\xi}_i)'$$

Its minimum $\frac{\partial F(\gamma)}{\partial \gamma} = \mathbf{0}$ yields the normal maximum likelihood estimator $\hat{\gamma}$ of the unknown vector of parameters γ.

If the model does not yield maximum likelihood estimates in closed form, it will be necessary to make use of an iterative procedure to minimize the discrepancy function. An optimization method based on the so-called Fischer scoring algorithm was developed by Browne and du Toit (1992). Fisher scoring algorithms require the gradient vector of the discrepancy function and use of the information matrix as an approximation to the Hessian matrix. Elements of the gradient vector, $\mathbf{g}(\gamma)$, and approximate Hessian matrix $\mathbf{H}(\gamma)$ of $\frac{1}{2}F(\gamma)$ are given by

$$\frac{1}{2}\frac{\partial F}{\partial \gamma_r} = [\mathbf{g}(\gamma)]_r = \sum_{i=1}^{N}\left\{tr\mathbf{Q}_i\frac{\partial \boldsymbol{\xi}_i}{\partial \gamma_r} + \frac{1}{2}tr\mathbf{P}_i\frac{\partial \boldsymbol{\Sigma}_i}{\partial \gamma_r}\right\},$$

where

$$\mathbf{Q}_i = (\mathbf{y}_i - \boldsymbol{\xi}_i)'\boldsymbol{\Sigma}_i^{-1}$$

$$\mathbf{P}_i = \boldsymbol{\Sigma}_i^{-1}(\mathbf{G}_{y_i} - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}$$

Let

$$[\mathbf{H}(\gamma)]_{r,s} = -E\left(\frac{\partial^2 \ell nL}{\partial \gamma_r \partial \gamma_s}\right)$$

then

$$\frac{1}{2}\frac{\partial^2 F}{\partial\gamma_r\partial\gamma_s} \approx [\mathbf{H}(\gamma)]_{r,s}$$
$$= \sum_{i=1}^{N}\left\{tr\left(\frac{\partial\boldsymbol{\xi}_i'}{\partial\gamma_r}\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\xi}_i}{\partial\gamma_s}\right) + \frac{1}{2}tr\left(\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\gamma_r}\boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\gamma_s}\right)\right\}$$

Suppose that γ_k is the k-th approximation to the $\hat{\gamma}$ which minimizes $F(\gamma)$. Let

$$\mathbf{g}_k = \mathbf{g}(\gamma_k), \mathbf{H}_k = \mathbf{H}(\gamma_k), \text{ and } F_k = F(\gamma_k).$$

The next approximation is obtained from

$$\hat{\gamma}_{k+1} = \hat{\gamma}_k + \alpha_k\delta_k,$$

where

$$\delta_k = -\mathbf{H}_k^{-1}\mathbf{g}_k$$

and α_k is a step size parameter chosen initially as 1 and then successively halved until $F_{k+1} \leq F_k$.

It was pointed out by Agresti (1990) that the Fisher scoring method resembles the Newton-Raphson method, the distinction being that Fisher scoring uses the expected value of the second derivative matrix.

2.7 Iterative generalized least squares

For convenience the general two-level model is written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{e}_i^\star \tag{2.7.1}$$

where $\boldsymbol{\beta}$ is a vector of fixed coefficients and $\mathbf{e}_i^\star$ is a vector of variables random at level one or two of the hierarchy. The matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the design matrices for the fixed and random variables in the model respectively.

Model (2.7.1) can be written in the form of model (2.5.6) if $\mathbf{W}_i = \mathbf{X}_i; \boldsymbol{\gamma}^\star = \boldsymbol{\beta}; [\mathbf{X}_{(2)i}\ \ \mathbf{X}_{(1)i}] = \mathbf{Z}_i$ and $\begin{pmatrix} \mathbf{u}_i \\ \mathbf{e}_i \end{pmatrix} = \mathbf{e}_i^\star$.

It is assumed that $\mathbf{u}_1, \cdots \mathbf{u}_N$ are a random sample of a $\mathbf{N}(\mathbf{0}, \boldsymbol{\Phi}_{(2)})$ random variable, and $\mathbf{e}_1, \cdots, \mathbf{e}_N$ are an independent random sample of a $\mathbf{N}(\mathbf{0}, \boldsymbol{\Phi}_{(1)})$ random variable. Under these assumptions the distribution of $\mathbf{y}_i$ is described, that is

$$\mathbf{V}_i = \text{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \mathbf{Z}_i\begin{pmatrix} \boldsymbol{\Phi}_{(2)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{(1)} \end{pmatrix}\mathbf{Z}_i'$$

and

$$E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$$

If $\mathbf{V}_i$ is assumed to be known, the generalized least squares (GLS) estimate $\hat{\beta}$ of β is obtained as the minimum of the quadratic function

$$\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)$$

and at the minimum

$$\hat{\beta} = \left[\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right]^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i\right) \tag{2.7.2}$$

The unknown parameters in $\mathbf{V}_i$ is the non-duplicated elements of the symmetric matrices $\mathbf{\Phi}_{(1)}$ and $\mathbf{\Phi}_{(2)}$. These elements are combined in the vector τ that is

$$\tau = \begin{pmatrix} \text{vecs}\mathbf{\Phi}_{(2)} \\ \text{vecs}\mathbf{\Phi}_{(1)} \end{pmatrix}$$

It is thus necessary to find the GLS estimator $\hat{\tau}$ of τ .

Now let

$$\mathbf{V} = \text{Diag}[\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_N], \tag{2.7.3}$$

$$\mathbf{X} : \left(\sum_{i=1}^{N} n_i \times m\right) = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \tag{2.7.4}$$

and

$$\mathbf{y} : \left(\sum_{i=1}^{N} n_i \times 1\right) = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} \tag{2.7.5}$$

Under the assumption that the $\mathbf{y}_i$'s are independently multivariate normal, it follows that

$$\mathbf{y} \sim \text{N}(\mathbf{X}\beta, \mathbf{V})$$

Suppose further that β is known and let

$$\mathbf{Y}^\star = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})', \tag{2.7.6}$$

then

$$E\mathbf{Y}^\star = \mathbf{V}$$

Note that $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{V}$ are respectively the mean vector and covariance matrix of $\mathbf{y}$ if the general model (2.7.1) is true. Denote the true mean vector and covariance matrix of $\mathbf{y}$ by $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}$ respectively. An estimate of $E(\mathbf{Y}^\star)$ is the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ and hence $W = \mathrm{Cov}(\hat{\boldsymbol{\Sigma}})$ is obtained from

$$\left[-E\frac{\partial^2 \ell nL}{\partial\sigma_{ij}\partial\sigma_{rs}}\right]^{-1} = [[H]_{ij,rs}]^{-1}$$

where

$$[H]_{ij,rs} = \frac{N}{2}tr\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_{ij}}\Sigma^{-1}\frac{\partial\Sigma}{\partial\sigma_{rs}}\right) \qquad \text{(Brown, 1991)} \tag{2.7.7}$$

Let $\mathbf{y}^\star = \mathrm{vecs}\mathbf{Y}^\star$ then an estimate of $E(\mathbf{y}^\star)$ is the ML estimator vecs $\hat{\boldsymbol{\Sigma}}$ of vecs $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ denotes the true population covariance matrix.

Using (cf Browne, 1991) the results $\frac{\partial\Sigma}{\partial\sigma_{ij}} = \mathbf{J}_{ij} + (1-\delta_{ij})\mathbf{J}_{ji}$ and $\mathrm{tr}[\mathbf{A}\mathbf{J}_{ij}\mathbf{B}\mathbf{J}_{rs}] = [\mathbf{A}]_{si}[\mathbf{B}]_{jr}$ as well as (2.7.7), it follows that

$$[\mathbf{W}^{\star^{-1}}]_{ij,rs} = \frac{(2-\delta_{ij})(2-\delta_{rs})N}{4}\left(\sigma^{ir}\sigma^{js} + \sigma^{is}\sigma^{jr}\right),$$

where $\mathbf{W}^\star$ denotes the covariance matrix of vecs $\hat{\boldsymbol{\Sigma}}$.

Using this result, it can be shown (Du Toit, 1992) that

$$\mathbf{W}^{\star^{-1}} = \frac{N}{2}\mathbf{G}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{G} \tag{2.7.8}$$

where $\mathbf{G}$ is defined by result 2 from Browne (1974).

If $\mathbf{Z}_i = [\mathbf{X}_{(2)i} \;\; \mathbf{X}_{(1)i}]$ then

$$\mathbf{V}_i = \mathbf{X}_{(2)i}\boldsymbol{\Phi}_{(2)}\mathbf{X}'_{(2)i} + \mathbf{X}_{(1)i}\boldsymbol{\Phi}_{(1)}\mathbf{X}'_{(1)i}$$

The following three results from Browne (1974) are applied:

1. $\mathrm{vec}\mathbf{C}\mathbf{A}\mathbf{C}' = \mathbf{C} \otimes \mathbf{C}\mathrm{vec}\mathbf{A}$

2. There is a unique matrix $\mathbf{G} : p^2 \times \frac{1}{2}p(p+1)$ such that

   $$\mathrm{vec}\mathbf{A} = \mathbf{G} \;\; \mathrm{vecs}\mathbf{A},$$

   with $\mathbf{A}$ a symmetric $p \times p$ matrix.

3. There is a non-unique matrix $\mathbf{H} : \frac{1}{2}p(p+1) \times p^2$ such that

$$\text{vecs}\mathbf{A} = \mathbf{H} \ \text{vec}\mathbf{A}$$

From result (1)

$$\text{vec}\mathbf{V}_i = \mathbf{X}_{(2)i} \otimes \mathbf{X}_{(2)i} \ \text{vec}\mathbf{\Phi}_{(2)} + \mathbf{X}_{(1)i} \otimes \mathbf{X}_{(1)i} \ \text{vec}\mathbf{\Phi}_{(1)}$$

Applying result (3)

$$\begin{aligned} \text{vecs}\mathbf{V}_i &= \mathbf{H}\text{vec}\mathbf{V}_i \\ &= \mathbf{H}(\mathbf{X}_{(2)i} \otimes \mathbf{X}_{(2)i})\text{vec}\mathbf{\Phi}_{(2)} + \mathbf{H}(\mathbf{X}_{(1)i} \otimes \mathbf{X}_{(1)i}) \ \text{vec}\mathbf{\Phi}_{(1)} \end{aligned}$$

Applying result (2) to $\text{vec}\mathbf{\Phi}_{(2)}$ and $\text{vec}\mathbf{\Phi}_{(1)}$

$$\begin{aligned} \text{vecs}\mathbf{V}_i &= \mathbf{H}(\mathbf{X}_{(2)i} \otimes \mathbf{X}_{(2)i})\mathbf{G}_m \text{vecs}\mathbf{\Phi}_{(2)} \\ &+ \mathbf{H}(\mathbf{X}_{(1)i} \otimes \mathbf{X}_{(1)i})\mathbf{G}_r \text{vecs}\mathbf{\Phi}_{(1)} \end{aligned} \quad (2.7.9)$$

If

$$\mathbf{X}_i^\star = \mathbf{H}((\mathbf{X}_{(2)i} \otimes \mathbf{X}_{(2)i})\mathbf{G}_m \ \ (\mathbf{X}_{(1)i} \otimes \mathbf{X}_{(1)i})\mathbf{G}_r)$$

then

$$\text{vecs}\mathbf{V}_i = \mathbf{X}_i^\star \tau$$

Let

$$\mathbf{X}^\star = \begin{bmatrix} \mathbf{X}_1^\star \\ \mathbf{X}_2^\star \\ \vdots \\ \mathbf{X}_N^\star \end{bmatrix}$$

The GLS estimator $\hat{\tau}$ of τ may be obtained as the minimum of the quadratic function

$$(\mathbf{y}^\star - \mathbf{X}^\star\tau)'\mathbf{W}^{\star^{-1}}(\mathbf{y}^\star - \mathbf{X}^\star\tau),$$

and hence

$$\hat{\tau} = (\mathbf{X}^{\star\prime}\mathbf{W}^{\star^{-1}}\mathbf{X}^\star)^{-1}\mathbf{X}^{\star\prime}\mathbf{W}^{\star^{-1}}\mathbf{y}^\star, \quad (2.7.10)$$

In practice $\mathbf{\Sigma}$ is unknown and is replaced in (2.7.8) by $\hat{\mathbf{V}} = \mathbf{V}(\hat{\tau})$.

Note that $\hat{\tau}$ can be evaluated from the alternative expression

$$\hat{\tau} = \left[\sum_{i=1}^{N} \mathbf{X}_i^{\star\prime} \mathbf{W}_i^{\star^{-1}} \mathbf{X}_i^{\star}\right]^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^{\star\prime} \mathbf{W}_i^{\star^{-1}} \mathbf{y}_i^{\star}\right),$$

where

$$\mathbf{W}_i^{\star^{-1}} = \frac{1}{2}\mathbf{G}'(\mathbf{V}_i^{-1} \otimes \mathbf{V}_i^{-1})\mathbf{G}$$

and

$$\mathbf{y}_i^{\star} = \text{vecs}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'.$$

When estimating $\hat{\boldsymbol{\beta}}$ according to (2.7.2) it is assumed that $\mathbf{V}$ is known. When estimating $\hat{\tau}$ according to (2.7.10) it is assumed that $\mathbf{W}^{\star}$ is a consistent estimator of $\text{Cov}(\mathbf{Y}^{\star}, \mathbf{Y}^{\star\prime})$ and that $\boldsymbol{\beta}$ is known, since $\boldsymbol{\beta}$ is required to evaluate $\mathbf{Y}^{\star}$ according to (2.7.6).

The following iterative procedure is used to obtain values of $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}$.

1. Set $\hat{\mathbf{V}}_i = \mathbf{I}$ in (2.7.2)
2. Obtain an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$
3. Calculate $\mathbf{W}^{\star}$ according to (2.7.8), with $\boldsymbol{\Sigma}$ replaced by $\hat{\mathbf{V}}$
4. Estimate $\hat{\tau}$ using (2.7.10) and obtain a revised estimate $\hat{\mathbf{V}} = \mathbf{V}(\hat{\tau})$ using (2.7.9)
5. Repeat steps (2) to (4) until convergence is obtained, for example $|\hat{\tau}_{k+1} - \hat{\tau}_k| < \varepsilon$; $|\hat{\boldsymbol{\beta}}_{k+1} - \hat{\boldsymbol{\beta}}_k| < \varepsilon$, with $\varepsilon_k = 10^{-6}$, where ε_k denotes a typical element of ε.

The algorithm described above (see e.g. Goldstein 1986) is known as iterative generalized least squares.

It can be shown (see e.g. Brown and du Toit, 1992 or Goldstein 1986) that under the assumption of multivariate normality, the iterative generalized least squares estimates are equivalent to the maximum likelihood estimates of the corresponding unknown parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}$.

2.8 Restricted maximum likelihood estimation

In Section 2.6 and 2.7 methods were discussed for the estimation of $\boldsymbol{\beta}$ and τ in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

$$\mathbf{e} = \begin{pmatrix} \mathbf{Z}_1\mathbf{e}_1^{\star} \\ \mathbf{Z}_2\mathbf{e}_2^{\star} \\ \vdots \\ \mathbf{Z}_N\mathbf{e}_N^{\star} \end{pmatrix}$$

and $\text{Cov}(\mathbf{e}, \mathbf{e}') = \mathbf{V}(\tau)$.
$\mathbf{V}, \mathbf{X}$ and $\mathbf{y}$ are defined by (2.7.3), (2.7.4) and (2.7.5) respectively, and $\boldsymbol{\beta}$ is a $(m \times 1)$ vector.

$\hat{\boldsymbol{\beta}}$ can also be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad (2.8.1)$$

It follows that

$$\text{Cov}(\mathbf{y}, \hat{\boldsymbol{\beta}}') = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1},$$

$$\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}') = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1},$$

and therefore from conditional multivariate normal results,

$$\text{Cov}(\mathbf{y}, \mathbf{y}'|\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}) = \mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}' \qquad (2.8.2)$$

One criticism of the ML approach to the estimation of τ (Harville, 1977) is that the ML estimator takes no account of the loss in degrees of freedom that results from estimating $\boldsymbol{\beta}$. In the restricted maximum likelihood approach (REML), inferences for τ are based on the likelihood function associated with $N^\star - m$ linearly independent error contrasts rather than on that associated with the full $(N^\star \times 1)$ data vector $\mathbf{y}$, $N^\star = \sum_{i=1}^{N} n_i$.

It can be shown that the log-likelihood function in this case with constant terms omitted is

$$-\frac{1}{2}\ell n|\mathbf{V}| - \frac{1}{2}\ell n|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Since (cf (2.8.2))

$$E(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' = \mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$$

Goldstein (1989) suggested that an updated estimate of $\mathbf{V}$ may be obtained if $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' + \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ is used, based on the current value of $\hat{\mathbf{V}}$; the term $\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ can be regarded as a bias correction term.

Since maximizing the log likelihood is equivalent to minimizing -2(log likelihood), we use

$$F = \ell n|\mathbf{V}| + tr\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' + \ell n|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|$$

The function to be minimized in REML is (cf Section 2.6) the ML function to be minimized plus $\ell n|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|$.

Goldstein (1989) showed that REML estimation is equivalent to iterative generalized least squares if the term $\mathbf{Y}^\star$ (cf (2.7.6)) is replaced by

$$\mathbf{S}^\star = (\mathbf{y} - \mathbf{X}\hat{\beta})(\mathbf{y} - \mathbf{X}\hat{\beta})' + \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$$

Let $\mathbf{S} = (\mathbf{y} - \mathbf{X}\hat{\beta})(\mathbf{y} - \mathbf{X}\hat{\beta})'$ then $\mathbf{S}^\star = \mathbf{S} + \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ and (cf (2.8.2)) $E(\mathbf{S}^\star) = \mathbf{V}$.

2.9 Extension to three-level models

The theory for two-level models can be extended to include a third level. If, for example area in which the schools are situated is also recorded, areas are level 3 units. Now (2.5.3) becomes

$$\mathbf{y}_{ik} = \mathbf{X}_{ik}\mathbf{b}_{ik} + \mathbf{e}_{ik} \qquad i = 1, 2, \cdots, N; k = 1, 2, \cdots, K; j = 1, 2, \cdots, n_{ik} \tag{2.9.1}$$

for area k, school i

where

$$\mathbf{y}_{ik} = \begin{bmatrix} y_{ik1} \\ y_{ik2} \\ \vdots \\ y_{ikn_{ik}} \end{bmatrix} \quad \text{and} \quad \mathbf{X}_{ik}$$

is an $n_{ik} \times m$ design matrix for the (ik)-th unit. The coefficients of the $m \times 1$ vector $\mathbf{b}_{ik}$ are considered to be random varying accross the level 2 and level 3 units. Again (2.9.1) can be written in an alternative form after expressing $\mathbf{b}_{ik}$ in terms of fixed parameters as well as coefficients that are random, varying across the level 3 units.

2.10 Hypothesis testing

A complex hypothesis about several elements of $\gamma^\star$ in (2.5.6) can be formulated. Make use of a $p \times \ell$ contrast matrix $\mathbf{C}$ with p the number of contrasts and $\ell = m + mq$. The hypothesis is written in the form

$$\mathbf{C}\gamma^\star = \mathbf{k}$$

where $\mathbf{k}$ is a $p \times 1$ vector (usually $\mathbf{k} = \mathbf{0}$). Now

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\mathbf{Z}_i\gamma^\star + \mathbf{X}_{(2)i}\mathbf{u}_i + \mathbf{X}_{(1)i}\mathbf{e}_i \\ &= \mathbf{W}_i\gamma^\star + \mathbf{R}_i\mathbf{r}_i \end{aligned}$$

where

$$\mathbf{R}_i = [\mathbf{X}_{(2)i}, \mathbf{X}_{(1)i}] \quad \text{and} \quad \mathbf{r}_i = \begin{bmatrix} \mathbf{u}_i \\ \mathbf{e}_i \end{bmatrix}$$

Using column stacking the models for N level 2 units (schools) can be combined.

$$\mathbf{y} = \mathbf{W}\gamma^\star + \mathbf{R}\mathbf{r}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N, \end{bmatrix}, \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_N, \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \vdots \\ \mathbf{R}_N \end{bmatrix} \text{ and } \mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_N \end{bmatrix}$$

Suppose $\hat{\gamma}^\star$ is the estimator of $\gamma^\star$ and $\hat{\mathbf{V}}$ the estimator of $\mathbf{V}$. It can be shown that $Cov(\hat{\gamma}^\star, \hat{\gamma}^{\star\prime}) = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}$

where

$$\mathbf{V}_i = \text{Cov}(\mathbf{y}_i, \mathbf{y}_i')$$

and

$$\mathbf{V} = \text{Diag}[\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_N]$$

For large samples $\mathbf{C}\hat{\gamma}^\star$ will have an approximate normal distribution with

$$E(\mathbf{C}\hat{\gamma}^\star) = \mathbf{C}\gamma^\star$$

and

$$\text{Cov}(\mathbf{C}\hat{\gamma}^\star, \mathbf{C}\hat{\gamma}^{\star\prime}) = \mathbf{C}(\mathbf{W}'\hat{\mathbf{V}}^{-1}\mathbf{W})^{-1}\mathbf{C}'$$

If the hypothesis $\mathbf{C}\gamma^\star = \mathbf{k}$ is true

$$U = (\mathbf{C}\hat{\gamma}^\star - \mathbf{k})'\{\mathbf{C}(\mathbf{W}'\hat{\mathbf{V}}^{-1}\mathbf{W})^{-1}\mathbf{C}'\}^{-1}(\mathbf{C}\hat{\gamma}^\star - \mathbf{k})$$

has an approximate χ^2-distribution with p degrees of freedom. A set of 100(1-α)% simultaneous confidence intervals for the p elements of $\mathbf{C}\gamma^\star$ is given by the p intervals

$$\mathbf{C}_i'\hat{\gamma}^\star \pm \{\mathbf{C}_i'(\mathbf{W}'\hat{\mathbf{V}}\mathbf{W})^{-1}\mathbf{C}_i\chi^2_{\ell,\alpha}\}^{0,5}$$

where $p \leq \ell$ and $\mathbf{C}_i'$ denotes the i-th row of $\mathbf{C}$, and $\chi^2_{\ell,\alpha}$ is the critical value of the χ^2 distribution with ℓ degrees of freedom.

2.11 Multilevel logit models

Serious inferential errors are often made if sample aspects such as clustering and stratification are ignored and data treated as if obtained from a simple random sampling scheme. Multilevel modelling can be used to take account of these sampling aspects if they are present.

Suppose that in cluster i $(i = 1, 2, \cdots, N)$ there are n_i individuals (e.g. children in a school). Each of these individuals belong to a certain category j $(j = 1, \cdots, C)$ (e.g. children in a certain class). Let n_{ij} be the number of individuals in cluster i, category j (e.g. the number of children in school i, class j). Clusters are level 2 units and categories are level 1 units.

The proportions $\frac{n_{ij}}{n}$ can be modelled to study variation across the clusters. A different variable is modelled instead. Let p_{ij} be the proportion of individuals in cluster i, category j which answers "a" to a certain "a/b" (dichotomous) question. Let

$$E(p_{ij}) = \pi_{ij}$$

Note that the C proportions, $p_{i1}, p_{i2}, \cdots, p_{ic}$ need not sum to 1 and that cluster i need not contain all C categories.

The following are goals in the analysis:

(a) Estimate the overall proportions of "a-answers" in each category.

(b) Estimate the variation in the π_{ij}'s across clusters. If this variation is large determine whether it could be accounted for by using cluster level covariates.

The distribution of raw proportions can be problematic, particularly when the extremes of 0 and 1 occur. The proportions are therefore transformed by using the logit (log(odds)):

$$\begin{aligned} y_{ij} &= log\left\{ \frac{p_{ij}}{1-p_{ij}} \right\} \qquad i = 1, \cdots, N \;\; j = 1, \cdots, C \\ &= log\left\{ \frac{m_{ij}}{n_{ij}-m_{ij}} \right\} \end{aligned}$$

where m_{ij} = number of "a-answers" in cluster i, category j.

To ensure the correct application of the logit let

$$m_{ij} = 0,25 \text{ if } m_{ij} = 0$$

and

$$m_{ij} = n_{ij} - 0,25 \text{ if } m_{ij} = n_{ij}$$

The simplest model to fit is the variance components model. For cluster i:

$$\begin{aligned} y_{i1} &= b_{i0} + \gamma_1 + e_{i1}/\sqrt{n_{i1}} \\ y_{i2} &= b_{i0} + \gamma_2 + e_{i2}/\sqrt{n_{i2}} \\ &\vdots \\ y_{ic} &= b_{i0} + \gamma_c + e_{ic}/\sqrt{n_{ic}} \end{aligned}$$

In matrix notation

$$\mathbf{y}_i = b_{i0}\mathbf{j} + \mathbf{X}_i\gamma + \mathbf{D}_i\mathbf{e}_i \tag{2.11.1}$$

where

$$\mathbf{j} : C \times 1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_c \end{bmatrix} \qquad \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ic} \end{bmatrix}$$

$$\mathbf{X}_i = \mathbf{I}_c \qquad \mathbf{D}_i : C \times C = \begin{bmatrix} \frac{1}{\sqrt{n_{i1}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{n_{i2}}} & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & & & \frac{1}{\sqrt{n_{ic}}} \end{bmatrix} \tag{2.11.2}$$

It is assumed that $b_{io} = u_{io}$
where $E(u_{io}) = u_o$ and $Var(u_{io}) = \sigma_o^2$. It is also assumed that $\mathbf{e}_1, \cdots, \mathbf{e}_N$ is a random sample of $\mathbf{e}$, which has mean zero and covariance matrix $\mathbf{D}_{\sigma_j^2}$. It follows that

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \sigma_o^2 \mathbf{j}\mathbf{j}' + \mathbf{D}_i \mathbf{D}_{\sigma_j^2} \mathbf{D}_i$$

and hence $var(y_{ij}) = \sigma_o^2 + \sigma_j^2/n_{ij}$

Goal (b) is achieved by estimating the level 2 variance σ_o^2. This estimate, $\hat{\sigma}_o^2$, gives an indication of the degree to which the population category proportions vary across clusters. Goal (a) is achieved by estimating the overall proportions of "a-answers" in each category. These estimates are $\hat{\gamma}_i, \hat{\gamma}_2, \cdots, \hat{\gamma}_c$. In the case of a large value for $\hat{\sigma}_o^2$, a more complex model can be fitted. If socio-economic status is to be incorporated in the model, this is done by extending (2.11.2) to

$$b_{i0} = \gamma_{00} + \gamma_{01} z_i + u_{i0}$$

where z_i is an indicator variable for socio-economic status (high,low etc.). Expression (2.11.1) can be written in the form of (2.5.6) with

$$\mathbf{W}_i = \mathbf{I}_c, \;\; \mathbf{X}_{(1)i} = \mathbf{D}_i \;\text{ and }\; \mathbf{X}_{(2)i} = \mathbf{j}$$

Example

Between March and May 1993 a countrywide survey was conducted by the HSRC among 8 366 adult black South Africans, including those in Venda and Ciskei. The title of this project was "The prospects for a free, democratic election" (De Kock, 1993). The specific categorical response used for a practical application is:

Are you going to take part in the forthcoming election for an interim government in South Africa?

The question (VOTE) has two possible outcomes namely,

1. yes, or
2. no

This is a typical situation where the multilevel analysis approach can be applied to fit a logit model. Persons are clustered according to the cross-classification of language group, gender and education level. There are four language groups (LANG) namely,

1. Xhosa
2. N.Sotho + Swazi
3. S.Sotho, Ndebele, Tsonga and others
4. Zulu

There are two gender groups (GENDER) namely,

1. Male
2. Female

There are four education level groups (EDUC) namely,

1. No education
2. Gr. 1-Gr. 7
3. Std. 6-Std. 9
4. Std. 10 +

Alltogether a maximum of $4 \times 2 \times 4 = 32$ clusters (CLUS) could be formed. Only 28 clusters are formed due to the fact that certain cross-classifications do not exist in this sample.

Within each cluster, five categories (CAT) are formed. These five categories are the five possible outcomes to the question (DAYS):

Would you prefer the coming election to take place on a single day or over two or more days?

The five possibles outcomes are:

1. One day
2. Two to three days

3. Four to five days
4. Six to seven days
5. More than seven days

The two goals in the analysis are:

(a) Estimate the overall proportions of "yes-answers" in each category.

(b) Estimate the variation across clusters.

The analysis was done according to the following steps:

Step 1

A dataset was created with the necessary variables in SAS. This program with a description is found in Appendix A1.

Step 2

The ML3 package was used for the analysis. A description of this package is found in Appendix A2. The specific ML3 program used for this example is found in Appendix A3.

The following output is obtained from the ML3 program

```
rand

                              LEVEL 3

  PARAMETER            ESTIMATE    S. ERROR     PREV. ESTIMATE     NCONV

                              LEVEL 2

  PARAMETER            ESTIMATE    S. ERROR     PREV. ESTIMATE     NCONV
CONS      /CONS          0.3969      0.1166             0.3972         2

                              LEVEL 1

  PARAMETER            ESTIMATE    S. ERROR     PREV. ESTIMATE     NCONV
RP1       /RP1            5.163       1.587              5.163         3
RP2       /RP2             4.59       1.718              4.574         1
RP3       /RP3            7.011       2.215              6.981         1
RP4       /RP4            16.75       4.805               16.8         3
RP5       /RP5            7.462       2.813              7.494         1
fixe


  PARAMETER            ESTIMATE    S. ERROR     PREV. ESTIMATE
```

FP1	0.7003	0.1468	0.7004
FP2	1.21	0.1337	1.21
FP3	1.676	0.1434	1.676
FP4	1.854	0.1705	1.854
FP5	1.696	0.1343	1.696

Goal (b) is achieved. From the output obtained after specifying the RAND command the estimate of the level 2 variance, σ_o^2, is $\hat{\sigma}_o^2 = 0.3969$. The estimated standard error is 0.1166. Convergence was achieved at the second iteration (NCONV = 2). The 95% confidence interval for this estimate is

$$(0.1637 \; ; \; 0.6301)$$

There is significant variation in category proportions of "yes-answers" (voters) among clusters. A more complex model may be fitted to the data to account for this variation.

The level 1 estimates are not interpreted.

Goal (a) is achieved. The estimates of the overall logarithms of the odds of voting in each category of DAYS are obtained from the output after specification of the FIXE command. Recall that in section 1.10 the odds are defined as

$$\frac{m_{ij}}{n_{ij} - m_{ij}}$$

where m_{ij} = number of "a-answers" (voters) in cluster i, category j and n_{ij} = total number of people in cluster i, category j.

The estimates and their respective confidence intervals are

Estimate	95% confidence interval
$\hat{\gamma}_1 = 0.7003$	(0.4067 ; 0.9939)
$\hat{\gamma}_2 = 1.21$	(0.9426 ; 1.4774)
$\hat{\gamma}_3 = 1.676$	(1.3892 ; 1.9628)
$\hat{\gamma}_4 = 1.854$	(1.513 ; 2.195)
$\hat{\gamma}_5 = 1.696$	(1.4274 ; 1.9646)

These estimates are all significant. For interpretation purposes the antilog of these estimates are calculated. That is, the odds of answering "yes" (to vote) if in category (1) (one day) of DAYS is

$$e^{\hat{\gamma}_1} = 2.01436$$

The odds for all categories of DAYS are given below.

	DAYS category	Odds
(1)	One day	2.0144
(2)	two-three days	3.3535
(3)	four-five days	5.3441
(4)	six-seven days	6.3853
(5)	more than seven days	5.4521

The odds of voting increases with the number of days wanted for election.

A specific odds ratio can now be calculated, for instance the odds ratio of DAYS category (4) to DAYS category (1). That is

$$\text{Odds ratio} = \frac{6.3853}{2.01436} = 3.1699$$

The odds of voting if in DAYS category (4) (election must take place over six to seven days) is thrice the odds of voting if in DAYS category (1) (election must take place in one day). Other relevant odds ratios can also be obtained.

As an example of hypothesis testing consider the following:

The FTES command is used to test the hypothesis

$$\gamma_1 - \gamma_4 = 0$$

The ML3 statements and output are as follows:

```
retr voting.ws
inpu c30
1 0 0 -1 0 0
ftes c30

CONTRASTS

FP1                    1.00
FP2                    0.00
FP3                    0.00
FP4                   -1.00
FP5                    0.00
result                -1.15
chi square ( 1 df)    62.00
+/-95% c.i.(sep.)      0.29
+/-95% c.i.(sim.)      0.49
```

```
 chi sq for simultaneous contrasts(5 df) =  62.00
stop
```

The above output is used to obtain a 95% simultaneous confidence interval for this contrast which is

$$\begin{aligned} & (-1.15 - 0.49 ; -1.15 + 0.49) \\ = & (-1.64 ; -0.66) \end{aligned}$$

The value -1.15 is obtained as $\hat{\tau}_1 - \hat{\tau}_4$. Since the above interval does not contain zero, the null hypothesis is rejected. Since the above interval does not contain zero, The overall proportion of voters in category 1 (one day election) is significantly lower than the overall proportion of voters in category 4 (six to seven days for the election).

2.12 Summary

This chapter deals with the theoretical aspects of multilevel models. The main concept is the extention of the familiar fixed parameter linear regression model to the random parameter linear regression model. Various approaches for the estimation of unknown parameters are discussed. The important estimation method is that of iterative generalized least squares. This method will be applied in modelling discussed further on. It is evident from this chapter that multilevel modelling is a powerful modelling approach. For further reading and examples see Bock (1989), Goldstein (1987 and 1991) and Wiggens (1990).

3 Models for analysing data with an ordinal dependent variable

3.1 The sampling distribution

An entire set of explanatory variables can be cross-classified to form a set of r subpopulations. Similarly the cross-classification of a set of c response variables gives a set of c response levels. Thus, regardless of the number of underlying variables, the multidimensional contingency table can be represented as a two-dimensional array representing the cross-classification of the response levels with the subpopulations. Table 3.1 illustrates the allocation of a sample of size n to the rc cells.

Table 3.1

		Response level				
		1	2	$\cdots$	c	
Subpopulation	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$

There are c response levels and r subpopulations. The cell frequency is denoted by n_{ij} for subpopulation i, and response level $j(i = 1, \cdots, r\, ; j = 1, \cdots, c)$. The row totals $n_{i.}$ represent the sample sizes for the r subpopulations with $\sum_{i=1}^{r} n_{i.} = n$. The underlying sampling distribution is assumed to be product multinomial. For each of the r subpopulations the sampling process is multinomial and the r samples are assumed to be mutually independent. The theoretical cell probabilities are denoted by π_{ij} and satisfy the condition

$$\sum_{j=1}^{c} \pi_{ij} = 1 \qquad i = 1, \cdots, r\, ; j = 1, \cdots, c$$

These densities are given in Table 3.2.

Table 3.2

		Response level				
		1	2	$\cdots$	c	
	1	π_{11}	π_{12}	$\cdots$	π_{1c}	$\pi_{1.}$
	2	π_{21}	π_{22}	$\cdots$	π_{2c}	$\pi_{2.}$
Subpopulation	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	π_{r1}	π_{r2}	$\cdots$	π_{rc}	$\pi_{r.}$

The $(rc \times 1)$ vector of cell densities is denoted by

$$\pi = \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \vdots \\ \pi_{1c} \\ \pi_{21} \\ \pi_{22} \\ \vdots \\ \pi_{2c} \\ \vdots \\ \pi_{r1} \\ \pi_{r2} \\ \vdots \\ \pi_{rc} \end{bmatrix}$$

The sample proportions $\frac{n_{ij}}{n_{i.}} = p_{ij}$ provide estimators of the parameters π_{ij}. Under the product multinomial assumption

$$E\left[\frac{n_{ij}}{n_{i.}}\right] = \pi_{ij} \qquad i = 1, \cdots, r \; ; \; j = 1, \cdots, c$$

$$Var\left[\frac{n_{ij}}{n_{i.}}\right] = \pi_{ij}(1 - \pi_{ij})/n_{i.} \qquad i = 1, \cdots, r \; ; \; j = 1, \cdots, c$$

$$Cov\left[\frac{n_{ij}}{n_{i.}}, \frac{n_{ik}}{n_{i.}}\right] = -\pi_{ij}\pi_{ik}/n_{i.} \qquad i = 1, \cdots, r \; ; \; j = 1, \cdots, c$$

$$Cov\left[\frac{n_{ij}}{n_{i.}}, \frac{n_{\ell k}}{n_{\ell.}}\right] = 0 \qquad i \neq \ell \;\; i = 1, \cdots, r \; ; \; j = 1, \cdots, c$$

The $(rc \times 1)$ vector of estimators is denoted by

$$\mathbf{p} = \begin{bmatrix} \frac{n_{11}}{n_{1.}} \\ \frac{n_{12}}{n_{1.}} \\ \vdots \\ \frac{n_{1c}}{n_{1.}} \\ \frac{n_{21}}{n_{2.}} \\ \frac{n_{22}}{n_{2.}} \\ \vdots \\ \frac{n_{2c}}{n_{2.}} \\ \vdots \\ \frac{n_{r1}}{n_{r.}} \\ \frac{n_{r2}}{n_{r.}} \\ \vdots \\ \frac{n_{rc}}{n_{r.}} \end{bmatrix} = \begin{bmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1c} \\ p_{21} \\ p_{22} \\ \vdots \\ p_{2c} \\ \vdots \\ p_{r1} \\ p_{r2} \\ \vdots \\ p_{rc} \end{bmatrix}$$

3.2 Ordinal response variables

Ordinal variables are often treated as qualitative, being analysed using methods for nominal variables. But in many respects, ordinal variables more closely resemble interval variables than nominal variables. They possess important quantitative features:

Each level has a greater or smaller magnitude of the characteristic than another level and, though often not possible to measure, there is usually an underlying continuous variable present.

Often numerical scores are assigned to ordinal categories (Stoker, 1982). The purpose of the scoring process might be to approximate relative distances for an underlying continuous scale. This requires good judgement and guidance from researchers who use the scale. In this section various models are discussed which eliminates the need for assigning scores (Du Toit and Lamprecht, 1984). These models are discussed in the context of a specific example. Consider the following example with an ordinal response variable and two explanatory variables:

Variable	Type	Description	Categories
MENT	Response	Mental impairment	1. Well
			2. Mild
			3. Moderate
			4. Impaired
EVENT	Explanatory	One important life event such as child birth, new job, divorce, death in in family that occured to the subject within the past three years	1. Yes 2. No
SES	Explanatory	Socio-economic status	1. High
			2. Low

MENT is the ordinal response variable and EVENT and SES are the two explanatory variables which are cross-classified to form four subpopulations. Table 3.3 analogous to Table 3.1 is constructed.

Table 3.3

		Ordered response categories of MENT				
		Well	Mild	Moderate	Impaired	
Subpopulation	Yes, High	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
	Yes, Low	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
	No, High	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
	No, Low	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$

Three models are discussed to analyse Table 3.3. Using the notation of Section 3.1 the probability, π_{ij}, of the j-th response for subpopulation i, is estimated by $p_{ij} = \frac{n_{ij}}{n_{i.}}$ $(i = 1, \cdots, 4 : j = 1, \cdots, 4)$. Some function is constructed defined on the response probabilities, by using these estimates. The function of the true probabilities is assumed to follow a linear model in terms of the design structure of the subpopulations.

3.3 The logit model

Table 3.4 is constructed by applying the logit transformation to the estimated probabilities. The last category within each subpopulation is used as baseline-category.

Table 3.4

		F_1	F_2	F_3
Subpopulation	Yes, High	$\ell n\frac{p_{11}}{p_{14}}$	$\ell n\frac{p_{12}}{p_{14}}$	$\ell n\frac{p_{13}}{p_{14}}$
	Yes, Low	$\ell n\frac{p_{21}}{p_{24}}$	$\ell n\frac{p_{22}}{p_{24}}$	$\ell n\frac{p_{23}}{p_{24}}$
	No, High	$\ell n\frac{p_{31}}{p_{34}}$	$\ell n\frac{p_{32}}{p_{34}}$	$\ell n\frac{p_{33}}{p_{34}}$
	No, Low	$\ell n\frac{p_{41}}{p_{44}}$	$\ell n\frac{p_{42}}{p_{44}}$	$\ell n\frac{p_{43}}{p_{44}}$

Define the (4×3) matrix $\mathbf{F}$ as $\mathbf{F} = (\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)$

Let

$$\begin{aligned}
\gamma_{1m} = E(F_{1m}) &= \theta_1 + DUMMY1 \star \lambda_1^{EVENT} + DUMMY2 \star \lambda_1^{SES} \\
\gamma_{2m} = E(F_{2m}) &= \theta_2 + DUMMY1 \star \lambda_2^{EVENT} + DUMMY2 \star \lambda_2^{SES} \\
\gamma_{3m} = E(F_{3m}) &= \theta_3 + DUMMY1 \star \lambda_3^{EVENT} + DUMMY2 \star \lambda_3^{SES}
\end{aligned}$$

γ_{im} $(i = 1, \cdots, 3\, ; m = 1, \cdots 4)$ is a linear model in the unknown parameters to be estimated.

Now let

$$E(\mathbf{F}) = \mathbf{Z}\boldsymbol{\beta}$$

where $\mathbf{Z}$ is a (4×3) design matrix

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

and $\boldsymbol{\beta}$ is the (3×3) matrix of unknown parameters

$$\boldsymbol{\beta} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \lambda_1^{EVENT} & \lambda_2^{EVENT} & \lambda_3^{EVENT} \\ \lambda_1^{SES} & \lambda_2^{SES} & \lambda_3^{SES} \end{bmatrix}$$

If $\mathbf{F}$ is the (12×1) vector $\mathbf{F} = \begin{bmatrix} \ell n\frac{p11}{p14} \\ \ell n\frac{p12}{p14} \\ \ell n\frac{p13}{p14} \\ \ell n\frac{p21}{p24} \\ \vdots \\ \ell n\frac{p43}{p44} \end{bmatrix}$

it follows that

$$E(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}_r$$

where $\mathbf{X}$ is a (12×9) design matrix

that is

$$\mathbf{X} = \mathbf{Z} \otimes \mathbf{I}_3$$

and $\boldsymbol{\beta}_r$ is the (9×1) vector of unknown parameters

$$\boldsymbol{\beta}_r = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \lambda_1^{EVENT} \\ \lambda_2^{EVENT} \\ \lambda_3^{EVENT} \\ \lambda_1^{SES} \\ \lambda_2^{SES} \\ \lambda_3^{SES} \end{bmatrix}$$

Consider for example the "Yes, Low" subpopulation. In this subpopulation

$$\begin{bmatrix} \gamma_{12} \\ \gamma_{22} \\ \gamma_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \boldsymbol{\beta} r$$

that is

$$\begin{aligned} \gamma_{12} &= \theta_1 + \lambda_1^{EVENT} - \lambda_1^{SES} \\ \gamma_{22} &= \theta_2 + \lambda_2^{EVENT} - \lambda_2^{SES} \\ \gamma_{32} &= \theta_3 + \lambda_3^{EVENT} - \lambda_3^{SES} \end{aligned}$$

3.4 The cumulative logit model

Another way to use ordered response categories is by forming logits, within each subpopulation i, of the cumulative probabilities,

$$F_{ij} = \pi_{i1} + \cdots + \pi_{ij} \qquad i = 1, \cdots, 4\,; j = 1, \cdots, 4$$

where j denotes the j-th response category.

The cumulative logits are defined as

$$\begin{aligned} L_{ij} = &\ logit[F_{ij}] = log\left[\frac{F_{ij}}{1-F_{ij}}\right] \\ = &\ log\left[\frac{\pi_{i1}+\cdots+\pi_{ij}}{\pi_{ij+1}+\cdots+\pi_{i4}}\right] \end{aligned}$$

Note that each cumulative logit uses all four response categories. Using these cumulative logits, Table 3.5 is constructed.

Table 3.5

		F_1	F_2	F_3
Subpopulation	Yes, High	$\ell n \frac{p_{11}}{p_{12}+p_{13}+p_{14}}$	$\ell n \frac{p_{11}+p_{12}}{p_{13}+p_{14}}$	$\ell n \frac{p_{11}+p_{12}+p_{13}}{p_{14}}$
	Yes, Low	$\ell n \frac{p_{21}}{p_{22}+p_{23}+p_{24}}$	$\ell n \frac{p_{21}+p_{22}}{p_{23}+p_{24}}$	$\ell n \frac{p_{21}+p_{22}+p_{23}}{p_{24}}$
	No, High	$\ell n \frac{p_{31}}{p_{32}+p_{33}+p_{34}}$	$\ell n \frac{p_{31}+p_{32}}{p_{33}+p_{34}}$	$\ell n \frac{p_{31}+p_{32}+p_{33}}{p_{34}}$
	No, Low	$\ell n \frac{p_{41}}{p_{42}+p_{43}+p_{44}}$	$\ell n \frac{p_{41}+p_{42}}{p_{43}+p_{44}}$	$\ell n \frac{p_{41}+p_{42}+p_{43}}{p_{44}}$

The same model as in the case of the logit model in Section 3.3 is defined. With the (12×1) vector $\mathbf{F}$ formed by now stacking the cumulative logits rowwise. The difference between the two models is that in the cumulative logit model cumulative comparisons are made compared to the single comparisons in the logit model. The rationale behind the cumulative logits is the following:

If for subpopulation i

$$\ell n \frac{p_{i1}}{p_{i2}+p_{i3}+p_{i4}} > 0 \text{ i.e. } \frac{p_{i1}}{p_{i2}+p_{i3}+p_{i4}} > 1$$

a concentration of responses in category 1 is implied. (mental impairment is well for subpopulation i).

If for subpopulation i

$$\ell n \frac{p_{i1}+p_{i2}}{p_{i3}+p_{i4}} > 0 \text{ i.e. } \frac{p_{i1}+p_{i2}}{p_{i3}+p_{i4}} > 1$$

a majority of responses in categories 1 and 2 is implied. (mental impairment is well to mild for subpopulation i).

This reasoning is used to describe all the subpopulations.

3.5 McCullagh's Proportional odds model

In McCullagh's proportional odds model cumulative logits, as described in Section 3.4, are modelled as in the cumulative logit model (McCullagh, 1980). The main difference between the two models is that McCullagh's model is given in terms of less unknown parameters than the cumulative logit model.

Now

$$
\begin{aligned}
\gamma_{1m} = E(F_{1m}) &= \theta_1 + DUMMY1 \star \lambda^{EVENT} + DUMMY2 \star \lambda^{SES} \\
\gamma_{2m} = E(F_{2m}) &= \theta_2 + DUMMY1 \star \lambda^{EVENT} + DUMMY2 \star \lambda^{SES} \\
\gamma_{3m} = E(F_{3m}) &= \theta_3 + DUMMY1 \star \lambda^{EVENT} + DUMMY2 \star \lambda^{SES} \quad m = 1,2,3,4
\end{aligned}
$$

If **F** is defined as in the case of the cumulative logit model and

$$E(\mathbf{F}) = \mathbf{X}\boldsymbol{\beta}_r$$

The design matrix is now defined as

$$
\underset{(12 \times 5)}{\mathbf{X}} = \begin{bmatrix}
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & -1 \\
0 & 1 & 0 & 1 & -1 \\
0 & 0 & 1 & 1 & -1 \\
1 & 0 & 0 & -1 & 1 \\
0 & 1 & 0 & -1 & 1 \\
0 & 0 & 1 & -1 & 1 \\
1 & 0 & 0 & -1 & -1 \\
0 & 1 & 0 & -1 & -1 \\
0 & 0 & 1 & -1 & -1
\end{bmatrix}
$$

and the (5×1) vector of unkown parameters is

$$
\boldsymbol{\beta}_r = \begin{bmatrix}
\theta_1 \\
\theta_2 \\
\theta_3 \\
\lambda^{EVENT} \\
\lambda^{SES}
\end{bmatrix}
$$

Consider for example the "Yes, Low" subpopulation. In this subpopulation

$$
\begin{bmatrix} \gamma_{12} \\ \gamma_{22} \\ \gamma_{32} \end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & 1 & -1 \\
0 & 1 & 0 & 1 & -1 \\
0 & 0 & 1 & 1 & -1
\end{bmatrix} \boldsymbol{\beta}_r
$$

that is

$$
\begin{aligned}
\gamma_{12} &= \theta_1 + \lambda^{EVENT} - \lambda^{SES} \\
\gamma_{22} &= \theta_2 + \lambda^{EVENT} - \lambda^{SES} \\
\gamma_{32} &= \theta_3 + \lambda^{EVENT} - \lambda^{SES}
\end{aligned}
$$

3.6 Estimation of the unknown parameters

3.6.1 The Delta method

The delta method (Agresti, 1990) is essentially making use of a first-order Taylor series expansion to obtain an approximate expression for the covariance matrix of a vector of stochastic variables, whose elements are again functions of stochastic variables.

Suppose x is a random variable with

$$E(x) = \mu$$

Suppose further a function F of x is differentiable at μ. The first-order Taylor series expansion for F, evaluated in the neighbourhood of μ is

$$F(x) = F(\mu) + (x-\mu)\frac{\partial F(x)}{\partial x}|_{x=\mu} + \text{error term}$$

If the error term is sufficiently small, it follows that

$$E[F(x)] = F(\mu)$$

and

$$Var[F(x)] = \left[\frac{\partial F(x)}{\partial x}|_{x=\mu}\right]^2 .Var(x)$$

Suppose now $\mathbf{F}$ is a $(u \times 1)$ vector of functions of the $(r \times 1)$ random vector $\mathbf{x}$ where

$$E(\mathbf{x}) = \boldsymbol{\mu}$$

The first-order Taylor series expansion for $\mathbf{F}(\mathbf{x})$ is

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\boldsymbol{\mu}) + \boldsymbol{\Delta}(\mathbf{x}-\boldsymbol{\mu}) + \text{vector of error terms}$$

where

$$\underset{(u \times r)}{\boldsymbol{\Delta}} = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}'}|_{\mathbf{x}=\boldsymbol{\mu}} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_r} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_r} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial F_u}{\partial x_1} & \frac{\partial F_u}{\partial x_2} & \cdots & \frac{\partial F_u}{\partial x_r} \end{bmatrix}_{|\mathbf{x}=\boldsymbol{\mu}} \qquad (3.6.1.1)$$

Now, for the elements of the vector of error terms sufficiently small, it follows that

$$E[\mathbf{F}(\mathbf{x})] = \mathbf{F}(\mu)$$

and

$$\text{Cov}[\mathbf{F}(\mathbf{x}), \mathbf{F}(\mathbf{x})'] = \Delta\text{Cov}(\mathbf{x}, \mathbf{x}')\Delta'. \qquad (3.6.1.2)$$

3.6.2 Generalized least squares estimation

Suppose the response variable Y has c categories and there are independent multinomial samples of sizes $n_{1.}, n_{2.}, \cdots, n_{r.}$ at r levels of an explanatory variable, or r levels of several explanatory variables.

Let

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_r \end{bmatrix} \quad \text{where} \quad \pi_i = \begin{bmatrix} \pi_{i1} \\ \pi_{i2} \\ \vdots \\ \pi_{ic} \end{bmatrix}$$

denote the cell probabilities. Let $\mathbf{p}$ denote the corresponding sample proportions. If $\mathbf{V}_i$ denotes the covariance matrix of $\mathbf{p}_i$, the sample proportions for subpopulation i, then

$$\underset{(rc \times rc)}{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 & & & \mathbf{0} \\ & \mathbf{V}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{V}_r \end{bmatrix} \qquad (3.6.2.1)$$

denotes the covariance matric of $\mathbf{p}$. If $\mathbf{F}(\pi)$ is a vector of response functions

$$\underset{(u \times 1)}{\mathbf{F}(\pi)} = \begin{bmatrix} \mathbf{F}_1(\pi) \\ \vdots \\ \mathbf{F}_u(\pi) \end{bmatrix}$$

the model has the form

$$\mathbf{F}(\pi) = \mathbf{X}\beta$$

where β is a $(t \times 1)$ vector of parameters, and $\mathbf{X}$ is a $(u \times t)$ design matrix. Let $\mathbf{F}(\mathbf{p})$ denote the sample response functions. It is assumed that $\mathbf{F}$ is differentiable at π.

Using the results from Section 3.6.1 it follows that

$$\mathbf{V}_F = \mathrm{Cov}[\mathbf{F}(\mathbf{p}), \mathbf{F}(\mathbf{p})'] \quad = \boldsymbol{\Delta}\mathrm{Cov}(\mathbf{p}, \mathbf{p}')\boldsymbol{\Delta}' \\ = \boldsymbol{\Delta}\mathbf{V}\boldsymbol{\Delta}'$$

where

$$\underset{(u \times rc)}{\boldsymbol{\Delta}} = \frac{\partial \mathbf{F}(\mathbf{p})}{\partial \mathbf{p}'}|_{\mathbf{p}=\boldsymbol{\pi}}$$

Let $\hat{\mathbf{V}}_F$ denote the sample version of $\mathbf{V}_F$, in which estimated proportions ($\hat{\boldsymbol{\pi}}$) are substituted in $\boldsymbol{\Delta}$ and $\mathbf{V}$. During the iteration process, $\hat{\boldsymbol{\pi}}$ is calculated at each iteration. The generalized least squares (GLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{F}(\mathbf{p})$$

which is obtained by minimizing the quadratic form

$$[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]$$

with respect to $\boldsymbol{\beta}$. The GLS estimator has an asymtotic multivariate normal distribution, with estimated covariance matrix

$$\hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}$$

Hypotheses about effects of explanatory variables have the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$$

where $\mathbf{C}$ is a known $(c \times t)$ matrix with $c \leq t$, having rank c. The estimator $\mathbf{C}\hat{\boldsymbol{\beta}}$ of $\mathbf{C}\boldsymbol{\beta}$ has an asymptotic normal distribution, with mean $\mathbf{0}$ under the null hypothesis and with covariance matrix estimated by

$$\mathbf{C}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{C}'.$$

Thus, the Wald statistic

$$W_c = \hat{\boldsymbol{\beta}}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$$

has an approximate chi-squared distribution with $df = c$.

3.7 A practical application

In the example to follow the logit model, cumulative logit model and McCullagh's proportional odds model are fitted.

The same HSRC dataset as described in the example in Section 2.11 is used. The specific ordinal response used in this example is:

Will there be more or less violence in the run-up to the coming election?

The question has three possible outcomes, namely,

1. Less violence
2. The same level of violence
3. More violence

It is clear that this response is ordinal.

The presence of children and gender are the two explanatory variables used. These two variables with their possible outcomes are:

Do you have Children?	and	Gender
1. Yes		1. Male
2. No		2. Female

These two variables are cross-classified to form four subpopulations. The following two-way table with frequencies is obtained.

Table 3.6
Ordered response categories

			Less	Same	More
	Yes,	Male	1142	610	329
Subpopulation	Yes,	Female	1501	962	493
	No,	Male	882	447	280
	No,	Female	739	368	264

When fitting the logit model as described in Section 3.3 the vector of unknown parameters to be estimated is

$$\beta_r = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \lambda_1^{CHILD} \\ \lambda_2^{CHILD} \\ \lambda_1^{GENDER} \\ \lambda_2^{GENDER} \end{bmatrix}$$

A SAS program was used to obtain the estimates of the unknown parameters. This program is given in Appendix B. The output obtained form this program when fitting the logit model is:

```
Iteration 1:              BHAT
                     1.1332163
                      0.527565
                      0.046051
                     0.1227896
                     0.0636332
                     0.0085279


Iteration 2:              BHAT
                     1.1331235
                      0.526377
                     0.0461133
                     0.1234798
                     0.0636577
                     0.0085053

                         .
                         .
                         .
                         .
                         .

Iteration 15:             BHAT
                     1.1331237
```

0.5263752
0.0461128
0.1234797
0.0636574
0.0084913

ANALYSIS OF RESULTS

SUMSTAT PARAMETER	DF	ESTIMATE	CHI-SQ	PROB	STD
1	1	1.1331237	1261.7998	0.0001	0.0318994
2	1	0.5263752	223.9714	0.0001	0.0351722
3	1	0.0461128	2.0670613	0.1505113	0.0320734
4	1	0.1234797	12.21372	0.0004744	0.0353323
5	1	0.0636574	4.0879555	0.0431899	0.0314844
6	1	0.0084913	0.0608492	0.8051587	0.0344227

RESPONSE FUNCTION

DESIGN	ACTUAL	PREDICTED	RESIDUAL
	1.2444786	1.2428939	0.0015848
	0.6174012	0.6583461	-0.040945
	1.1133777	1.1155791	-0.002201
	0.6685053	0.6413636	0.0271417
	1.1474025	1.1506683	-0.003266
	0.467769	0.4113868	0.0563822
	1.0293488	1.0233536	0.0059953
	0.3321338	0.3944042	-0.06227

After 15 iterations $\hat{\beta}_r$ is obtained as

$$\hat{\beta}_r = \begin{bmatrix} 1.133 \\ 0.526 \\ 0.046 \\ 0.123 \\ 0.064 \\ 0.008 \end{bmatrix}$$

For each estimate a Wald statistic is calculated (see Section 3.6.2). These values are given in the column "CHI-SQ". The probabilities for the test of significance are given in the column "PROB" and the standard errors of the estimates in the column "STD".

It follows that the estimated parameters $\hat{\lambda}_1^{CHILD}$ and $\hat{\lambda}_1^{GENDER}$ are not significant on the 5% level of significance.

The column named "PREDICTED" gives the values

$$\begin{bmatrix} \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \\ \hat{\gamma}_{21} \\ \hat{\gamma}_{22} \\ \hat{\gamma}_{31} \\ \hat{\gamma}_{32} \\ \hat{\gamma}_{41} \\ \hat{\gamma}_{42} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_1 + \hat{\lambda}_1^{CHILD} + \hat{\lambda}_1^{GENDER} \\ \hat{\theta}_2 + \hat{\lambda}_2^{CHILD} + \hat{\lambda}_2^{GENDER} \\ \hat{\theta}_1 + \hat{\lambda}_1^{CHILD} - \hat{\lambda}_1^{GENDER} \\ \hat{\theta}_2 + \hat{\lambda}_2^{CHILD} - \hat{\lambda}_2^{GENDER} \\ \hat{\theta}_1 - \hat{\lambda}_1^{CHILD} + \hat{\lambda}_1^{GENDER} \\ \hat{\theta}_2 - \hat{\lambda}_2^{CHILD} + \hat{\lambda}_2^{GENDER} \\ \hat{\theta}_1 - \hat{\lambda}_1^{CHILD} - \hat{\lambda}_1^{GENDER} \\ \hat{\theta}_2 - \hat{\lambda}_2^{CHILD} - \hat{\lambda}_2^{GENDER} \end{bmatrix} = \begin{bmatrix} 1.24 \\ 0.66 \\ 1.12 \\ 0.64 \\ 1.15 \\ 0.41 \\ 1.02 \\ 0.39 \end{bmatrix}$$

The column named "ACTUAL" gives the actual values analogous to Table 3.4. The column named "RESIDUAL" gives the differences between the actual and the predicted values. To interpret the results the predicted probabilities are calculated according to the algorithm in the SAS program described in Appendix B. These probabilities are

Table 3.7

Ordered response categories

			Less	Same	More
	Yes,	Male	0.54	0.30	0.16
Subpopulation	Yes	Female	0.51	0.32	0.17
	No,	Male	0.56	0.26	0.18
	No,	Female	0.53	0.28	0.19

The same pattern is observed for every subpopulation. The majority of people feel that there will be less violence. Less people feel that the violence level will be the same and few people feel that the violence will be more. Notice that males feel slightly more positive than females and that people without children feel slightly more positive than people with children.

When fitting the cumulative logit model to the data the output from the program is:

```
Iteration 1:                BHAT
                        0.140655
                       1.5686709
                       -0.031167
                       0.0742865
                       0.0578138
                       0.0463137
```

Iteration 2:

BHAT
0.1404128
1.5698197
-0.031064
0.0738719
0.0578293
0.0464383

.
.
.
.
.

Iteration 15:

BHAT
0.1404129
1.5698174
-0.031065
0.0738737
0.0578319
0.0464209

ANALYSIS OF RESULTS

SUMSTAT PARAMETER	DF	ESTIMATE	CHI-SQ	PROB	STD
1	1	0.1404129	36.581257	0.0001	0.0232155
2	1	1.5698174	2646.2782	0.0001	0.0305163
3	1	-0.031065	1.7683288	0.183589	0.0233605
4	1	0.0738737	5.8024429	0.0160039	0.0306679
5	1	0.0578319	6.5244464	0.0106402	0.022641
6	1	0.0464209	2.3800386	0.1228945	0.03009

RESPONSE FUNCTION

DESIGN	ACTUAL	PREDICTED	RESIDUAL
	0.1957209	0.1671803	0.0285406
	1.6724555	1.690112	-0.017656
	0.0311257	0.0515165	-0.020391
	1.6086262	1.5972701	0.0113561
	0.1932656	0.2293093	-0.036044
	1.5573925	1.5423646	0.0150278
	0.1564085	0.1136455	0.0427631

```
1.4334598    1.4495228    -0.016063
```

After 15 iterations $\hat{\beta}_r$ is obtained as

$$\hat{\beta}_r = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\lambda}_1^{CHILD} \\ \hat{\lambda}_2^{CHILD} \\ \hat{\lambda}_1^{GENDER} \\ \hat{\lambda}_2^{GENDER} \end{bmatrix} = \begin{bmatrix} 0.140 \\ 1.570 \\ -0.031 \\ 0.074 \\ 0.058 \\ 0.046 \end{bmatrix}$$

The estimated parameters $\hat{\lambda}_1^{CHILD}$ and $\hat{\lambda}_2^{GENDER}$ are not significant on the 5% level of significance.

The predicted probabilities are given in Table 3.8

Table 3.8

Ordered response categories

			Less	Same	More
	Yes,	Male	0.54	0.30	0.16
Subpopulation	Yes,	Female	0.51	0.32	0.17
	No,	Male	0.56	0.26	0.18
	No,	Female	0.53	0.28	0.19

Exactly the same pattern is obtained when fitting the cumulative logit model. The estimated expected cumulative logits can also be interpreted. Consider, for example, the subpopulation of males with children. Using the first two values from the "PREDICTED" column

$$e^{0.1671803} \simeq 1.18 \quad \text{and} \quad e^{1.690112} \simeq 5.42$$

The value of 1.18 indicates that the probability of being in the first category of the response (less violence) is more or less the same as the probability of being in one of the other categories (same level of violence or more). The value of 5.42 indicates that the probability of being in one of the first two categories of the response is 5.42 times the probability of being in the last category of the response. When applying this reasoning to all the subpopulations the same conclusions are made as with the interpretation of the predicted probabilities.

When fitting McCullagh's model to the data the output from the program is:

```
Iteration 1:                  BHAT
                         0.1351666
                         1.5831722
                         -0.005263
                         0.0549315


Iteration 2:                  BHAT
                         0.1338374
                         1.5920083
                         -0.005649
                         0.0546225

                              .
                              .
                              .
                              .
                              .

Iteration 15:                 BHAT
                         0.1338783
                         1.5920762
                         -0.005786
                         0.0546363
```

ANALYSIS OF RESULTS

SUMSTAT PARAMETER	DF	ESTIMATE	CHI-SQ	PROB	STD
1	1	0.1338783	33.48314	0.0001	0.0231365
2	1	1.5920762	2753.7311	0.0001	0.0303391
3	1	-0.005786	0.0672258	0.7954196	0.0223168
4	1	0.0546363	6.3700896	0.011606	0.0216475

RESPONSE FUNCTION

DESIGN	ACTUAL	PREDICTED	RESIDUAL
	0.1957209	0.1827283	0.0129926
	1.6724555	1.6409262	0.0315293
	0.0311257	0.0734558	-0.04233
	1.6086262	1.5316537	0.0769725
	0.1932656	0.1943009	-0.001035

```
1.5573925   1.6524988   -0.095106
0.1564085   0.0850284   0.0713801
1.4334598   1.5432263   -0.109766
```

After 15 iterations $\hat{\beta}_r$ is obtained as

$$\hat{\beta}_r = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\lambda}^{CHILD} \\ \hat{\lambda}^{GENDER} \end{bmatrix} = \begin{bmatrix} 0.134 \\ 1.592 \\ -0.006 \\ 0.055 \end{bmatrix}$$

The estimated parameter $\hat{\lambda}^{CHILD}$ is not significant on the 5% level of significance.

The predicted probabilities are given in Table 3.9

Table 3.9

Ordered response categories

			Less	Same	More
	Yes,	Male	0.55	0.29	0.16
Subpopulation	Yes,	Female	0.52	0.30	0.18
	No,	Male	0.55	0.29	0.16
	No,	Female	0.52	0.30	0.18

The same pattern is obtained, that is in every subpopulation the majority of people feel that there will be less violence. Males feel slightly more positive than females and there is no difference in perception between people with and without children. Again the estimated expected cumulative logits can also be interpreted which will result in the same conclusions.

3.8 Summary

Three models are discussed for analysing data with an ordinal dependent variable. These models, the logit, cumulative logit and McCullagh's proportional odds model, eliminates the need for assigning scores to ordinal categories. Generalized least squares estimation is applied to obtain estimates of the unknown parameters. Use is made of the delta method to obtain an approximate expression for the covariance matrix of a vector of stochastic variables whose elements are again functions of stochastic variables.

4 Multilevel models for analysing data with an ordinal dependent variable

4.1 Theoretical Aspects

Consider the example with an ordinal response and two explanatory variables discussed in Section 3.2. If area of residence is also recorded, areas can be regarded as level 2 units, whereas the people in the sample can be regarded as level 1 units. Recall that the two-level model discussed in Section 2.5 can be written as model (2.5.3) for area i, that is

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{e}_i \qquad i = 1, 2, \cdots, N \tag{4.1.1}$$

Where $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{u}_i$

If

$$\text{Cov}(\mathbf{e}_i, \mathbf{e}_i') = \mathbf{V}_i \;\text{ and }\; \text{Cov}(\mathbf{u}_i, \mathbf{u}_i') = \mathbf{\Phi}$$

then

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \mathbf{X}_i\mathbf{\Phi}\mathbf{X}_i' + \mathbf{V}_i = \mathbf{\Sigma}_i \tag{4.1.2}$$

In terms of the logit model of Section 3.3,

$\mathbf{y}_i$ is the (12×1) vector of response functions $\mathbf{F}$
$\mathbf{X}_i$ is the (12×9) design matrix $\mathbf{X} = \mathbf{Z} \otimes \mathbf{I}_3$
$\boldsymbol{\beta}$ is the (9×1) vector of unknown paramaters $\boldsymbol{\beta}_r$.

If expression (2.7.2) is used to obtain estimates of the unknown parameters it follows that

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i\right]^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{y}_i\right) \tag{4.1.3}$$

In expression (4.1.3) it is assumed that $\mathbf{\Sigma}_i$ is known, which implies that $\mathbf{\Phi}$ and $\mathbf{V}_i$ is known. $\mathbf{V}_i$ is calculated as the covariance matrix of the vector of response functions as in the logit model of Section 3.3. The task of estimating $\mathbf{\Phi}$ is more cumbersome.

In the context of Section 2.7, $\mathbf{\Phi}$ is analogous to $\mathbf{\Phi}_{(2)}$. In this case it is only necessary to estimate $\text{vecs}\mathbf{\Phi}_{(2)}$ or $\text{vecs}\mathbf{\Phi}$. Let

$$\boldsymbol{\tau} = (\text{vecs}\mathbf{\Phi})$$

The GLS estimator of $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ may be obtained as the minimum of the quadratic function

$$(\mathbf{y}^\star - \mathbf{X}^\star\boldsymbol{\tau})'\mathbf{W}^{\star-1}(\mathbf{y}^\star - \mathbf{X}^\star\boldsymbol{\tau}) \tag{4.1.4}$$

where

$$\begin{aligned} \mathbf{y}^\star &= \text{vecs}\mathbf{Y}^\star \\ &= \text{vecs}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \qquad \text{(Du Toit, 1993)} \end{aligned} \tag{4.1.5}$$

with $\mathbf{y}$ the vector of response functions of all N areas, $\mathbf{X}$ is the matrix formed by stacking the design matrices of each area, and $\boldsymbol{\beta}$ is assumed to be known.

The matrix $\mathbf{X}^\star$ is formed by stacking $\mathbf{X}_i^\star$, that is

$$\mathbf{X}^\star = \begin{bmatrix} \mathbf{X}_1^\star \\ \mathbf{X}_2^\star \\ \vdots \\ \mathbf{X}_N^\star \end{bmatrix}$$

where

$$\mathbf{X}_i^\star = \mathbf{H}(\mathbf{X} \otimes \mathbf{X})\mathbf{G} \tag{4.1.6}$$

with $\mathbf{H}$ and $\mathbf{G}$ as defined in Section 2.7. $\mathbf{W}^\star$ is defined by (2.7.8) that is,

$$\mathbf{W}^{\star^{-1}} = \frac{N}{2}\mathbf{G}'(\boldsymbol{\Sigma}^{\star-1} \otimes \boldsymbol{\Sigma}^{\star-1})\mathbf{G}, \tag{4.1.7}$$

where $\boldsymbol{\Sigma}^\star$ denotes the true population covariance matrix of $\mathbf{y}$.

Let $\boldsymbol{\Phi} = \mathbf{T}\mathbf{T}'$ be the Cholesky decomposition of $\boldsymbol{\Phi}$, where $\mathbf{T}$ is a lower triangular matrix. Expression (4.1.4) can be written as

$$\begin{aligned} Q &= (\mathbf{y}^\star - \mathbf{X}^\star\text{vecs}\mathbf{T}\mathbf{T}')'\mathbf{W}^{\star-1}(\mathbf{y}^\star - \mathbf{X}^\star\text{vecs}\mathbf{T}\mathbf{T}') \\ &= (\mathbf{y}^\star - \hat{\mathbf{y}}^\star)'\mathbf{W}^{\star-1}(\mathbf{y}^\star - \hat{\mathbf{y}}^\star), \end{aligned} \tag{4.1.8}$$

with

$$\hat{\mathbf{y}}^\star = \mathbf{X}^\star\text{vecs}\mathbf{T}\mathbf{T}'.$$

Let $\mathbf{W}^{\star-1} = \mathbf{L}\mathbf{L}'$ be the Cholesky decomposition of $\mathbf{W}^{\star-1}$ where $\mathbf{L}$ is a lower triangular matrix.

Now if $\mathbf{u} = \mathbf{L}'\mathbf{y}^\star$ and $\hat{\mathbf{u}} = \mathbf{L}'\hat{\mathbf{y}}^\star$

then

$$\begin{aligned} Q &= (\mathbf{u}-\hat{\mathbf{u}})'(\mathbf{u}-\hat{\mathbf{u}}) \\ &= \sum_{i=1}^{I}(u_i-\hat{u}_i)^2 \qquad i=1,2,\cdots,I \end{aligned} \tag{4.1.9}$$

The GLS estimator of $\hat{\tau}$ of $\boldsymbol{\tau}$ is obtained by minimizing (4.1.9) with respect to $\mathbf{T}$. This is done to ensure a positive definite matrix $\hat{\boldsymbol{\Phi}}$.

Let vecs$\mathbf{T} = \boldsymbol{\gamma}$. Elements of the gradient vector $\mathbf{g}(\boldsymbol{\gamma})$ are given by

$$\begin{aligned} [\mathbf{g}(\boldsymbol{\gamma})]_r = \frac{\partial Q}{\partial_{t_{ij}}} &= -2\sum_{i=1}^{I}(u_i-\hat{u}_i)\frac{\partial \hat{u}_i}{\partial_{t_{ij}}} \\ &= -2(\mathbf{u}_i-\hat{\mathbf{u}}_i)'\frac{\partial \hat{\mathbf{u}}_i}{\partial_{t_{ij}}} \\ &= -2(\mathbf{y}^\star-\hat{\mathbf{y}}^\star)\mathbf{W}^{\star-1}\frac{\partial \hat{\mathbf{y}}\star}{\partial_{t_{ij}}} \end{aligned} \tag{4.1.10}$$

The first element of g is the derivative of Q with respect to the first element of vecs$\mathbf{T}$, that is, $\frac{\partial Q}{\partial_{t_{11}}}$ etc. Elements of the approximate Hessian matrix $\mathbf{H}(\boldsymbol{\gamma})$ is given by

$$\begin{aligned} \frac{\partial Q}{\partial_{t_{kl}}\partial_{t_{ij}}} &= 2\sum_{i=1}^{I}\left[\frac{\partial \hat{u}_i}{\partial_{t_{kl}}}\frac{\partial \hat{u}_i}{\partial_{t_{ij}}}\right] \\ &\quad -2\sum_{i=1}^{I}(u_i-\hat{u}_i)\frac{\partial \hat{u}_i}{\partial_{t_{kl}}\partial_{t_{ij}}} \\ [H(\boldsymbol{\gamma})]_{r,s} &= E\left(\frac{\partial Q}{\partial_{t_{kl}}\partial_{t_{ij}}}\right) = 2\frac{\partial \hat{\mathbf{u}}'}{\partial_{t_{kl}}}.\frac{\partial \hat{\mathbf{u}}}{\partial_{t_{ij}}} \\ &= 2\frac{\partial \hat{\mathbf{y}}^\star}{\partial_{t_{kl}}}\mathbf{W}^{\star-1}\frac{\partial \hat{\mathbf{y}}^\star}{\partial_{t_{ij}}} \end{aligned} \tag{4.1.11}$$

The element in row r and column s of $\mathbf{H}$ is the expected value of Q differentiated with respect to the element r and s of vecs$\mathbf{T}$. Suppose that γ_k is the k-th approximation to the $\hat{\gamma}$ which minimizes Q. Let

$$\mathbf{g}_k = \mathbf{g}(\boldsymbol{\gamma}_k), \mathbf{H}_k = \mathbf{H}(\boldsymbol{\gamma}_k), \text{and } \; Q_k = Q(\boldsymbol{\gamma}_k)$$

The next approximation is obtained from

$$\hat{\gamma}_{k+1} = \hat{\gamma}_k + \boldsymbol{\alpha}_k\boldsymbol{\delta}_k \tag{4.1.12}$$

Where $\boldsymbol{\delta}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k$ and $\boldsymbol{\alpha}_k$ is a step size parameter chosen initially as 1. This process of approximation is repeated, with the successive halving of $\boldsymbol{\alpha}_k$, until $|Q_{k+1} - Q_k| < \varepsilon$, with $\varepsilon = 10^{-6}$. The method described above is known as the Gauss-Newton method. Note that when estimating $\boldsymbol{\beta}$ it is assumed that $\boldsymbol{\Phi}$ is known, and when estimating $\boldsymbol{\Phi}$ it is assumed that $\boldsymbol{\beta}$ is known. Values of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Phi}}$ may be obtained iteratively as follows:

1. To obtain an initial estimate $\hat{\mathbf{T}}$ of $\mathbf{T}$ and thus of $\mathbf{\Phi} = \mathbf{TT}'$, the following steps are carried out: Calculate $\hat{\boldsymbol{\beta}}$ for each level 2 unit and store these $\hat{\boldsymbol{\beta}}$'s as the matrix $\mathbf{B}$, where

$$\mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1' \\ \boldsymbol{\beta}_2' \\ \vdots \\ \boldsymbol{\beta}_N' \end{bmatrix} \tag{4.1.13}$$

The sample covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained as

$$\mathrm{S} = \frac{1}{N-1}\mathrm{B}'(\mathrm{I}_N - \mathbf{jj}'/N)\mathrm{B} \tag{4.1.14}$$

where

$$\mathbf{j} : (N \times 1) = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

An initial estimate of the matrix $\mathbf{T}$ is obtained from the Cholesky decomposition of S, that is,

$$\mathrm{S} = \hat{\mathbf{T}}\hat{\mathbf{T}}' \tag{4.1.15}$$

where $\hat{\mathbf{T}}$ is a lower triangular matrix. This matrix S is used as an initial estimate of $\mathbf{\Phi}$.

2. Obtain an estimate of $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ using (3.1.3)

3. Obtain an estimate of $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ using the expression

$$\hat{\boldsymbol{\tau}} = \left[\sum_{i=1}^{N} \mathbf{X}_i^{\star\prime}\mathbf{W}_i^{\star -1}\mathbf{X}_i^{\star}\right]^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i^{\star\prime}\mathbf{W}_i^{\star -1}\mathbf{y}_i^{\star}\right), \tag{4.1.16}$$

4. Obtain a new matrix $\mathbf{T}$ by applying the Gauss-Newton method

5. Repeat steps (2) to (4) until convergence is obtained, for example $|\hat{\boldsymbol{\tau}}_{k+1} - \hat{\boldsymbol{\tau}}_k| < \boldsymbol{\varepsilon}; |\hat{\boldsymbol{\beta}}_{k+1} - \hat{\boldsymbol{\beta}}_k| < \boldsymbol{\varepsilon}$, with $\varepsilon_k = 10^{-6}$, where ε_k denotes a typical element of $\boldsymbol{\varepsilon}$.

An improvement to the above algorithm is to test, after obtaining $\hat{\boldsymbol{\tau}}$, whether $\hat{\mathbf{\Phi}}$ is positive definite. If $\hat{\mathbf{\Phi}}$ proves to be positive definite, the Gauss-Newton step will be unnecessary.

The theory described above can also by applied to the cumulative logit model and McCullagh's model.

4.2 Practical application

The same example as in the application of Section 3.7 is used. Now another variable is included namely national developement region. The total area from which a sample was taken can be divided into 11 development regions. These regions are level 2 units. In this example the cumulative logit model in the case of multilevel data is fitted. The theory used is as described in Section 4.1. The vector $\hat{\beta}$ and the matrix $\hat{\Phi}$ is obtained through SAS programs which are found, with descriptions, in Appendix C. The following output is obtained:

```
Iteration 1:                 BHAT
                        0.2007274
                        1.5506069
                        0.0834918
                        0.0341046
                        -0.009121
                         0.041709


        PHI
   0.082222 0.0337931 -0.001946 -0.011262 0.0005781  -0.00937
  0.0337931 0.0604979 -0.002845 0.0017675 0.0062546 0.0091414
  -0.001946 -0.002845 0.0074364 0.0017231   0.00134 -0.000475
  -0.011262 0.0017675 0.0017231 0.0084329 0.0005844 0.0021656
  0.0005781 0.0062546   0.00134 0.0005844 0.0050852 -0.001529
   -0.00937 0.0091414 -0.000475 0.0021656 -0.001529 0.0113368


Iteration 2:                 BHAT
                        0.1970199
                        1.5860334
                        0.0827798
                        0.0345069
                        -0.008851
                        0.0364795


        PHI
  0.0833146 0.0338964 -0.002251 -0.012224 0.0000506 -0.009871
  0.0338964 0.0610943 -0.002744 0.0032785  0.006791 0.0101682
  -0.002251 -0.002744 0.0082296 0.0011231 0.0012238 -0.000527
  -0.012224 0.0032785 0.0011231 0.0103892 0.0007265 0.0029409
  0.0000506  0.006791 0.0012238 0.0007265 0.0051638 -0.001187
  -0.009871 0.0101682 -0.000527 0.0029409 -0.001187 0.0139781
```

```
Iteration 6:                    BHAT
                             0.1967993
                             1.5861606
                             0.0820845
                             0.0344568
                             -0.009838
                             0.0353515
```

PHI

0.0832951	0.033919	-0.00226	-0.012202	0.0000193	-0.00986
0.033919	0.0609965	-0.002708	0.0031627	0.0068264	0.0101615
-0.00226	-0.002708	0.0082266	0.001119	0.0012209	-0.000526
-0.012202	0.0031627	0.001119	0.010252	0.0007278	0.0028545
0.0000193	0.0068264	0.0012209	0.0007278	0.0051706	-0.001172
-0.00986	0.0101615	-0.000526	0.0028545	-0.001172	0.0139672

ANALYSIS OF RESULTS

SUMSTAT PARAMETER	DF	ESTIMATE	CHI-SQ	PROB	STD
1	1	0.1967993	4.6125685	0.0317385	0.0916331
2	1	1.5861606	364.85659	0.0001	0.0830398
3	1	0.0820845	4.743367	0.0294115	0.0376893
4	1	0.0344568	0.5787942	0.4467853	0.0452911
5	1	-0.009838	0.0862266	0.7690302	0.0335046
6	1	0.0353515	0.5195253	0.4710443	0.0490461

After 6 iterations $\hat{\beta}$ is obtained as

$$\hat{\beta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\lambda}_1^{CHILD} \\ \hat{\lambda}_2^{CHILD} \\ \hat{\lambda}_1^{GENDER} \\ \hat{\lambda}_2^{GENDER} \end{bmatrix} = \begin{bmatrix} 0.196 \\ 1.586 \\ 0.082 \\ 0.034 \\ -0.010 \\ 0.035 \end{bmatrix}$$

The estimated parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ and $\hat{\lambda}_1^{CHILD}$ are significant on the 5% level of significance.

Inspection of $\hat{\Phi}$ shows that the level 2 variances and covariances are quite small which suggests low variation across the level 2 units (areas).

The predicted probabilities, analogous to the probabilities in the example of Section 3.7, are given in Table 4.1

Table 4.1

Ordered response categories

			Less	Same	More
	Yes,	Male	0.57	0.27	0.16
Subpopulation	Yes,	Female	0.57	0.26	0.17
	No,	Male	0.53	0.30	0.17
	No,	Female	0.53	0.29	0.18

Again, the majority of people within each subpopulation feel that there will be less violence. Now though, people with children feel more positive than people without children. Males and females within each of the two groups, people with children and people without children, feel exactly the same. Notice that when fitting the cumulative logit model with a "correction" for area in which the people stay produces different results than fitting the cumulative logit model (cf. Section 3.7) without "correcting" for area.

4.3 Summary

The logit model, discussed in Section 3.3, is written as a two-level model. Iterative generalized least squares estimation is used as a method for obtaining estimates of the unknown parameters. These estimates are $\hat{\beta}$ and $\hat{\Phi}$. Cholesky decomposition is applied in the Gauss-Newton approximation step to ensure a positive definite matrix, $\hat{\Phi}$. The theory of this chapter can also be applied to the previously discussed cumulative logit model and McCullagh's proportional odds model.

CHAPTER 5

5 Conclusions and suggestions for further research

Multilevel theory is an extremely powerful approach in various fields of statistical application. In this report general multilevel theory is given where the emphasis is in the field of categorical data analysis. Analysing data with an ordinal dependent variable in the context of multilevel theory is only one example of application. The analysis of multilevel models is a relatively new field and various other models, other than the three models discussed, need to be incorporated in the multilevel context.

Multilevel modelling was dicussed in the context of social sciences. More examples can be found in economics, in political science, repeated measures data analysis, in the estimation of variance components in complex sampling and multilevel non-linear modelling.

At the moment limited computer software is available for the analysis of multilevel models. The program given in this report is also limited in the sense that no provision is made for the analysis of three-level models. The runtime of this program can also be minimized, for example:

Recall that the covariance matrix of the vector of response functions for level i can be written as expression (4.1.2) that is,

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i') = \mathbf{X}_i \boldsymbol{\Phi} \mathbf{X}_i' + \mathbf{V}_i = \boldsymbol{\Sigma}_i$$

where

$\mathbf{y}_i$ is the $(n_i \times m)$ vector of response functions
$\mathbf{X}_i$ is the $(n_i \times m)$ design matrix
$\boldsymbol{\Phi}$ is the $(m \times m)$ covariance matrix of $\mathbf{u}_i$.

It is assumed that $\mathbf{V}_i$, the covariance matrix of $\mathbf{e}_i$, is of the form

$$\mathbf{V}_i = \sigma^2 \mathbf{I}_{n_i}$$

The following result from Browne (1991) is applied:

Let A: $p \times p$ and C: $m \times m$ be nonsingular and let $\mathbf{B}$ be an $m \times m$ matrix. Then

$$(\mathbf{A} + \mathbf{BCB}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}^{-1}$$

if

$$|\mathbf{C}^{-1} + \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}| \neq 0$$

Thus $\boldsymbol{\Sigma}_i^{-1}$ can be written as

$$\begin{aligned} \boldsymbol{\Sigma}_i^{-1} &= \mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i(\boldsymbol{\Phi}^{-1} + \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1} \\ &= \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{X}_i(\boldsymbol{\Phi}^{-1} + \mathbf{X}_i'\sigma^{-2}\mathbf{X}_i)^{-1}\mathbf{X}_i'\sigma^{-2} \end{aligned}$$

In the existing computer program, $\boldsymbol{\Sigma}_i^{-1}$, the inverse of a $n_i \times n_i$ matrix, is calculated. If $\boldsymbol{\Sigma}_i^{-1}$ is obtained as in the above expression, the inverse of an $m \times m$ matrix will be calculated. With $m < n$ this will imply less computer runtime.

Multilevel models for the analysis of ordinal data

by

Helena Nyikos

Supervisor: Professor S.H.C. Du Toit

Statistics Department

Submitted in part fulfilment of the requirements for the degree of Master of Science in the subject Mathematical Statistics

6 Summary

The scope for application of multilevel models is very wide. The term multilevel refers to a hierarchical relationship among units in a system. In an education system, for example, multi-level data is obtained from samples of randomly drawn students (level 1) from randomly drawn classes (level 2) from randomly drawn schools (level 3). Multilevel analysis allows characteristics of each group (for example the students of a specific class of a specific school) to be incorporated into models of individual behaviour.

General multilevel theory is discussed. The fixed parameter linear regression model is extended to a random parameter linear regression model. Marginal maximum likelihood and the E-M algorithm are given combined as a means for estimating the unknown model parameters. A general expression for the two-level model is obtained. Maximum likelihood estimation and iterative generalized least squares are discussed as estimation procedures. The multilevel logit model is emphasized as a form of the general two-level model, and illustrated with an example. The two-level model is then extended to the general three-level model.

Ordinal variables are often treated as qualitative, being analysed using methods for nominal variables. But, in many aspects ordinal variables more closely resemble interval variables. Often in analysis numerical scores are assigned to ordinal categories. This approach though is subjective. In a new approach, three models are described. These models are the logit model, the cumulative logit model and McCullagh's proportional odds model. To estimate the unknown model parameters, generalized least squares estimation is applied.

The three models used for analysing data with an ordinal dependent variable is described in the context of multilevel theory. Iterative generalized least squares is discussed in this new framework. In particular Cholesky decomposition is used to obtain a positive definite matrix estimate of the covariance matrix of the explanatory variables whose coefficients are random at level 2. Examples of the logit, cumulative logit and McCullagh's proportional odds models are used to illustrate the effect of the multilevel approach.

Meerpeilmodelle vir die analise van ordinale data

deur

Helena Nyikos

Studieleier: Professor S.H.C. Du Toit

Departement Statistiek

Voorgelê ter vervulling van 'n deel van die vereistes vir die graad Magister in Wiskundige Statistiek.

7 Samevatting

Daar bestaan baie toepassings vir meerpeilmodelle. Die term hiërargies verwys na 'n verwantskap tussen eenhede in 'n sisteem. In 'n onderwyssisteem, byvoorbeeld, word meerpeil data verkry vanuit steekproewe van ewekansig gekose leerlinge (peil 1) uit ewekansig gekose klasse (peil 2) uit ewekansig gekose skole (peil 3). Meerpeilanalise neem eienskappe van elke groep (byvoorbeeld leerlinge in 'n spesifieke klas in 'n spesifieke skool) in ag in die modellering .

Algemene meerpeilteorie word bespreek. Die vaste parameter lineêre regressie model word na die stogastiese parameter lineêre regressie model uitgebrei. Die metode van marginale maksimum aanneemlikheid en die E-M algoritme word as beramingsmetodes bespreek om die onbekende parameters te beraam. 'n Algemene uitdrukking vir die twee-peil model word verkry. Maksimum aanneemlikheid en iteratiewe veralgemeende kleinste kwadrate word as beramingsmetodes bespreek. Die meerpeil logit model word as 'n tipe twee-peil model beklemtoon en met 'n voorbeeld geïllustreer. Die twee-peil model word na die drie-peil model uitgebrei.

Ordinale veranderlikes word dikwels as kwalitatief beskou en geanaliseer volgens metodes vir nominale veranderlikes. Hierdie benadering is egter subjektief. In 'n nuwe benadering word drie modelle bespreek. Hierdie modelle is die logit model, die kumulatiewe logit model en McCullagh se model. Veralgemeende kleinste kwadrate word as beramingsmetode vir die onbekende parameters gebruik.

Die drie modelle vir die analise van data met 'n ordinale veranderlike word in die konteks van meerpeilteorie bespreek. Iteratiewe veralgemeende kleinste kwadrate word ook bespreek in hierdie konteks. Cholesky dekomposisie word gebruik om 'n matriks te verkry wat positief-definiet is vir die beraming van die kovariansie matriks van die onafhanklike veranderlikes met stogastiese koëffisiënte op peil twee. Die effek van die meerpeil benadering word geïllustreer met voorbeelde van die logit, kumulatiewe logit en McCullagh se model.

8 References

AITKIN, M. and LONGFORD, N. (1986). Statistical modeling issues in school effectiveness studies (with discussion) *J. Royal Statist. Soc., A. 149(1), 1-43.*

AGRESTI, A. (1990). *Categorical Data Analysis.* John Wiley, New York.

BOCK, R.D. (1989). *Multilevel analysis of educational data.* R.D. Bock (Ed.) San Diego: Academic Press.

BOCK, R.D. (1990). Marginal maximum likelihood estimation for the two-stage random regressions model with second-stage covariates. *Technical report,* Department of Psychology, University of Chicago, Chicago.

BROWNE, M.W. (1974). Generalized least squares estimators in the analysis of covariance structures. *S.A. Statistical Journal, Vol. 8. No. 1.*

BROWNE, M.W. (1991). *STA 401-Matrix methods in statistics.* Supplementary notes. University of South Africa, Pretoria.

BROWNE, M.W. and DU TOIT, S.H.C. (1992). Automated fitting of nonstandard models for mean vectors and covariances matrices. *Multivariate Behavioural Research. Vol. 27, No. 2. 269-300.*

DE KOCK, C. et al (1993). *The prospects for a free, democratic election.* Pretoria: Human Sciences Research Council.

DU TOIT, S.H.C. (1993). *Analysis of multilevel models.* Pretoria: Human Sciences Research Council.

DU TOIT, S.H.C. and BROWNE, M.W. (1991). *PC AUFIT User's guide for PC version of AUFIT program for automated fitting of non standard models.* University of Pretoria, Pretoria.

DU TOIT, S.H.C. and BROWNE, M.W. (1982) *AUFIT (Version 1): A computer program for the automated fitting of non-standard models for means and covariances.* Research finding WS-27. Pretoria: Human Sciences Research Council.

DU TOIT, S.H.C. and BROWNE, M.W. (1989). *AUFIT (Version 2): PC version of computer program AUFIT,* University of Pretoria, Pretoria.

DU TOIT, S.H.C. and LAMPRECHT, T. (1984). *Analyzing data with an ordinal dependent variable.* Occasional paper no 12. Pretoria: Human Sciences Research Council

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika 73, 43-56.*

GOLDSTEIN, H. (1987) *Multilevel models in educational and social research.* C. Griffin & Co., London, Oxford University Press, New York.

GOLDSTEIN, H. (1989). Restricted unbiased iterative generalized least squares estimation. *Biometrika 76, 662-3.*

GOLDSTEIN, H. (1991). Multilevel modelling of survey data. *The Statistician, 40, 235-244.*

GOLDSTEIN, H. and McDONALD, R,P. (1988). A general model for the analysis of multilevel data. *Psychometrika 53, 455-467.*

HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Assoc. 72, No. 358, 320-340.*

McCULLAGH, P. (1980). Regression models for ordinal data. *J. Royal statist. Soc., B, 42, 109-142.*

PROSSER, R. RASBASH, J. & GOLDSTEIN, H. (1990). *ML3 Software for three-level analysis.* London: Institute of Education.

RASBASH, J., PROSSER, R. & GOLDSTEIN, H. (1989). *ML2 and ML3: Software for two-level and three-level analysis.* Users guide. London: Institute of education.

SEBER, G.A.F. and WILD, C.J. (1989).*Nonlinear regression.* John Wiley and Sons.

STOKER, D.J. (1982) *The analysis of categorical data with ordered categories A review.* Paper PR-82. Pretoria: Human Sciences Research Council.

WIGGINS, R.D. et al (1990). Multilevel analysis of attitudes to abortion. *The Statistician, 40, 225-234.*

9 Appendix A1

9.1 The SAS Program

A dataset with the necessary variables is created using SAS. Each observation (record) must have a code for cluster and category, and must also have the frequency of "yes-answers" (m) and total frequency of people (n) in that specific cluster and category combination. There will be a maximum of 28 × 5 observations. These observations are the level 1 units. To create such a dataset the following SAS program can be used (the initial data step is omitted).

```
DATA D1;SET MASTER.TOTAL;

*DEFINE THE CATEGORIES;

CAT=DAYS;

*FORM THE 32 CLUSTERS;

IF LANG=1 AND EDUC=1 AND GENDER=1 THEN GROUP=1;
IF LANG=1 AND EDUC=1 AND GENDER=2 THEN GROUP=2;
IF LANG=1 AND EDUC=2 AND GENDER=1 THEN GROUP=3;
IF LANG=1 AND EDUC=2 AND GENDER=2 THEN GROUP=4;
IF LAND=1 AND EDUC=3 AND GENDER=1 THEN GROUP=5;
IF LANG=1 AND EDUC=3 AND GENDER=2 THEN GROUP=6;
IF LANG=1 AND EDUC=4 AND GENDER=1 THEN GROUP=7;
IF LANG=1 AND EDUC=4 AND GENDER=2 THEN GROUP=8;

IF LANG=2 AND EDUC=1 AND GENDER=1 THEN GROUP=9;
IF LANG=2 AND EDUC=1 AND GENDER=2 THEN GROUP=10;
IF LANG=2 AND EDUC=2 AND GENDER=1 THEN GROUP=11;
IF LANG=2 AND EDUC=2 AND GENDER=2 THEN GROUP=12;
IF LAND=2 AND EDUC=3 AND GENDER=1 THEN GROUP=13;
IF LANG=2 AND EDUC=3 AND GENDER=2 THEN GROUP=14;
IF LANG=2 AND EDUC=4 AND GENDER=1 THEN GROUP=15;
IF LANG=2 AND EDUC=4 AND GENDER=2 THEN GROUP=16;

IF LANG=3 AND EDUC=1 AND GENDER=1 THEN GROUP=17;
IF LANG=3 AND EDUC=1 AND GENDER=2 THEN GROUP=18;
IF LANG=3 AND EDUC=2 AND GENDER=1 THEN GROUP=19;
IF LANG=3 AND EDUC=2 AND GENDER=2 THEN GROUP=20;
IF LAND=3 AND EDUC=3 AND GENDER=1 THEN GROUP=21;
IF LANG=3 AND EDUC=3 AND GENDER=2 THEN GROUP=22;
IF LANG=3 AND EDUC=4 AND GENDER=1 THEN GROUP=23;
IF LANG=3 AND EDUC=4 AND GENDER=2 THEN GROUP=24;

IF LANG=4 AND EDUC=1 AND GENDER=1 THEN GROUP=25;
IF LANG=4 AND EDUC=1 AND GENDER=2 THEN GROUP=26;
```

```
 IF LANG=4 AND EDUC=2 AND GENDER=1 THEN GROUP=27;
 IF LANG=4 AND EDUC=2 AND GENDER=2 THEN GROUP=28;
 IF LAND=4 AND EDUC=3 AND GENDER=1 THEN GROUP=29;
 IF LANG=4 AND EDUC=3 AND GENDER=2 THEN GROUP=30;
 IF LANG=4 AND EDUC=4 AND GENDER=1 THEN GROUP=31;
 IF LANG=4 AND EDUC=4 AND GENDER=2 THEN GROUP=32;

 DATA D2;SET D1 (KEEP= GROUP CAT VOTE);
  PROC SORT;
 BY GROUP;                                *CLUSTER;

 PROC FREQ ;
  TABLES VOTE*CAT/OUT=A NOPRINT;
 BY GROUP;

 DATA E;SET A ;
  IF CAT^=.;                               *CATEGORY;
  IF VOTE=2;
  VAR1=COUNT;
 PROC SORT;
 BY GROUP CAT;

 DATA F;SET A ;
  IF CAT^=.;
  IF VOTE=1;
  VAR2=COUNT;
 PROC SORT;
 BY GROUP CAT;

 DATA G;MERGE E F;
 BY GROUP CAT;
  TOTAL=SUM(VAR1,VAR2);
  IF TOTAL=VAR2 THEN VAR2=TOTAL-0.25;
  IF VAR2=. THEN VAR2=0.25;

PUT (GROUP CAT) (3.) VAR2 7.2 TOTAL 6.;
```

10 Appendix A2

10.1 The ML3 Package

ML3 (Prosser, Rasbash and Goldstein, 1989 and 1990) is the updated version of ML2. This package is easily installed on a personal computer. Data are stored by the program in a memory segment called a worksheet. This worksheet can be thought of as a matrix. A row is assigned to each level 1 unit and a column contains values for a given variable. The columns are numbered from 1 to 100 and are referred to as $C1, C2, \cdots, C100$. ML3 is invoked by typing ML3 at the DOS prompt. The command STOP is used to end a worksession.

Every ML3 instruction must begin with a keyword. At least the first four characters of the keyword must be specified. Typing

HELP

lists all the available keywords, and typing

HELP keyword

gives the syntax for the specific keyword.

Typing LOGO turns recording of your worksession on (and off). When the command is used first, ML3 requests the name of a log file in which screen output is to be stored. In order to create more than one log file during a worksession the commmand LOGO 1 closes a current log file and turns logging off. Typing LOGO turns logging on again and you will be prompted for a new log filename. In this way, different parts of screen output can be stored in separate files.

The command

DIREctory

shows a listing of files in a particular DOS diretory, without having to end the worksession. To see the contents of a file use the command

VIEW

The command

SAVE

records the data and settings of the current worksheet during a worksession. ML3 prompts for the name of a worksheet storage file. The command

SAVE 1

records only the data in the worksheet.

The command

RETRieve

can be used to restore data and settings of a previously saved worksheet. The command

RETR 1

is used to retrieve only the data of a previously saved worksheet.

To remove all data from the worksheet use the command

WIPE

The command

ERASE $C_{-} \ldots C_{-}$

clears the specified columns.

Some commands in a certain sequence might be used in many analyses. It is possible to store these commands in a file that is to be called up during a worksession. Use the command

OBEY

to call up a file containing certain commands. You will be prompted to specify a filename.

Conducting an analysis can roughly be divided into five stages:

1. Input
2. Data manipulation
3. Model specification

4. Run

5. Output

These five stage are discussed separately.

1. Input

The data which is to be used in the analysis can either be read from an ASCII-format file or entered during the worksession. All variables must be numerical.

1.1 Input from an existing file

To input free format data use the command

DINPut $C_{-} \cdots C_{-}$

where $C_{-} \cdots C_{-}$ refers to the columns into which the data is to be read. If the dataset to be inputted contains five variables, for example, the command is

DINP C1 - C5

note: any range of columns can be specified as long as five columns are specified. The following input commands are also correct:

	DINP	C2 - C6
or	DINP	C2 C4 C5 C3 C1

To input data which is to be formatted use the command

FDINput $C_{-} \cdots C_{-}$

You will be prompted for the format statement. The format statement must be typed on a single line and begins with a "(" and ends with a ")". Skipping characters is indicated by -N where N is the number of characters to be skipped. If observations have more than one line of data (more than one card) indicate the beginning of a new line by a "/". It is not possible to indicate the locations of decimal points. Decimal points can either be put in the original data or the variables can be multiplied/divided by appropriate powers of 10 after inputting.

Example

Suppose part of the information is arranged as follows:

line 1
a level 1 ID code in columns 1-3
a number series ability score in columns 5-8

line 2
a level 2 ID code in columns 1-3
a mathematics score in columns 4-6

Assuming that there will be only two lines per observation the required format is

$$(3, -1, 4/3, 3)$$

1.2 Input data during worksession

This command, usually used to input vectors and small matrices, has the syntax

$$\text{INPUt } C_{-} \cdots C_{-}$$

and is followed by one or more rows of numbers. The total number of numbers must be a multiple of the number of columns specified. To input the matrix

$$\begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{matrix}$$

into columns C1-C3,three of the possible ways are;

INPU C1-C3	INPU C1-C3	INPU C1-C3
1 2 3	1 2 3	1 2
4 5 6	4 5	3 4
	6	5
		6

1.3 Naming columns

It is possible to give names to one or more columns with the command

$$\text{NAME } C_{-} \text{ 'name1' } C_{-} \text{ 'name2' } \cdots$$

Example

If mathematics score is contained in column 1 and number series ability in column 3, name columns with the command

NAME C1 'MATH' C3 'ABILITY'

The first eight characters of a name are used.

1.4 Identifiers

Each case should be labelled with a case identifier (level 1 units), group identifier (level 2 units) and a third level identifier in the case of three-level modelling. If, for example, the level 1 identifier is in a column with the name 'ID1' and the level 2 identifier is in a column with the name 'ID2' specify these identifiers with the commands

IDEN 1 'ID1'
and IDEN 2 'ID2'

1.5 Checking entered data

The command

NAMEs

displays column identifiers, each column's length, number of missing values and the maximum and minimum value within a column.

The command

WRITe $C_{-} \cdots C_{-}$

shows the contents of each specified column.

The command

SUMMary

displays each group identification code (level 2 and level 3 groups), its number and percentage of level 1 units, and the total number of units at each level. Use this command after specifying level identification codes with the IDEN command.

Suppose an error is detected in column 3 case 17. A value of 32 was inputted whereas the value should have been 44. The command

EDIT 17 C3 44

will overwrite the value of 32 in row 17 with the value of 44.

2. Data manipulation

ML3 has commands for almost any kind of data manipulation. Only a few basic commands are discussed here. Data manipulations for the practical application in chapter 5 were mainly done using SAS.

2.1 Transformations

Five basic one-operand transformations are

LOGT	$C\_$	$C\_$	(log to the base 10)
LOGE	$C\_$	$C\_$	(log to the base e)
LOGI	$C\_$	$C\_$	($ln(p/(1-p))$ for $0 < p < 1$)
EXPO	$C\_$	$C\_$	(exponential)
SQRT	$C\_$	$C\_$	(square root)

For example

LOGT C1 C3

will write the logarithms of values in C1 in C3.

Specifying

LOGT C1 C1

will overwrite the values in C1 with their respective logarithms.

Arithmic functions and/or one-operand transformations can be combined into one single operation with the command

CALCulate $C\_$ = expression involving columns/numbers

For example to transform the values of x in C9 with the transformation

$$(x - \sqrt{2})(\sqrt{x} + 1)$$

and to write the transformed values into C10, type the command

CALC C10 = (C9 - SQRT(2))*(SQRT(x)+1)

2.2 Coding

The CODEs command has several uses in multilevel modelling. The syntax is

CODEs from one to M in blocks of N, P times to be put in C_{-}

Suppose a datafile containes 100 cases. To create a column C3 which contains 1's type

CODE 1 1 100 C3

This command can be used to create an intercept column. Suppose the 100 cases must be numbered from 1 to 100 and these numbers must be written to C2. Type

CODE 100 1 1 C2

Values in C2 can be used as level 1 identifiers.

2.3 Dummy variables

Suppose that each case in an analysis belongs to one of three categories of the variable AGEG. Dummy variables, two in this example, can be formed with the DUMMies command. If the three categories are 1, 2 and 3

DUMM 'AGEG' C8 C9

writes dummy vectors into C8 and C9. Cases with AGEG equal to 1 are in the base category, and coefficients for the dummy variables represent differences of categories 2 and 3 from category 1. The general form is

DUMMies using categorical variable C_{-} to be written in $C_{-} \cdots C_{-}$

and the number of destination columns must be one less than the number of categories of the categorical variable.

3. Model specification

This stage involves a set of commands which specifies the roles of the variables and the parameters to be estimated. As mentioned in section 4.1.4, identifiers for the various levels are necessary. These identifiers are specified with the IDEN command.

The RESP command declares one variable to be the response variable. The syntax is

RESPonse variable is in C_{-}

To change the response variable, simply specify the command again with the new variable.

The EXPL command gives one or more variables the role of being explanatory variables. The syntax is

EXPLanatory variable candidates in $C_{-} \cdots C_{-}$

This is a toggle command. To undeclare some variable, retype the EXPL command specifying the variable concerned. By default all variables specified with EXPL are in the fixed part of the model.

Using the command

FPAR explanatory variables in $C_{-} \cdots C_{-}$

will remove the specified variables from the fixed part of the model. These variables are now in the random part of the model.

The command

SETVariance-covariance matrix at level N

requests estimation of all variances and covariances of the coefficients of all explanatory variables in the random part of the model. Specifying

SETV N $C_{-} \cdots C_{-}$

restricts the request to all coefficient variances related to the variables named in the command.

Use

CLRV N

to undo the request given with SETV, and

CLRV N $C_{-} \cdots C_{-}$

to request that certain variances and/or covariances not be estimated.

To request the estimation of parameters one at a time, use

SETElement at level N $C_{-} \cdots C_{-}$

To undo this command type

CLRE N $C_{-} \cdots C_{-}$

Example

In section 1.12 one of the goals is to estimate the level 2 variance. It is assumed that

$$b_{io} = u_{io}$$

If "CONS" is the explanatory variable associated with the intercept b_{io}, request estimation of σ_o^2 with the command

SETE 2 "CONS" "CONS"
or SETV 2 "CONS" "CONS"

Use the command

SETT

to see the current model specifications.

4. Run

A number of features in the estimation process can be controlled.

The command

MAXIterations N

changes the default number of iterations of five to N. A test for convergence is made at the end of each iteration. Convergence is reached if a certain estimated value is smaller than the tolerance. The tolerance is a number of the form 10^{-N} where N is an iteger in the range $1, 2, \cdots, 9$. The default tolerance is three. To change the tolerance to 10^{-N} use the command

TOLErance N

The default estimation procedure used is iterative generalized least squares (IGLS). The command

METHod of estimation is XXXX

with XXXX either equal to IGLS or RIGLS (restrictive iterative generalized least squares) is used to control the estimation procedure.

The command

BATChmode

turns inter-iteration on/off. Default operation is interrupted mode. The estimation process is started with the command

STARt

When the program pauses at the end of an iteration (this will happen if inter-iteration is off), the command

NEXT

will start the next iteration.

5. Output

The estimates of the fixed parameters and their standard errors are shown if the command

FIXEd

is specified. The command

RANDom

displays estimates of the random parameters and their standard errors.

Fixed parameter contrasts can be examined with the command

FTESt using information in C_{-}

Consider the contrast $\mathbf{C}\gamma = \mathbf{k}$

that is

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ \vdots & \vdots & \vdots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pq} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_q \end{bmatrix} = \begin{bmatrix} k_1 \\ \vdots \\ k_q \end{bmatrix}$$

The vector to be specified in the FTES command is

$$\mathbf{f} = \begin{bmatrix} c_{11} \\ \vdots \\ c_{1q} \\ k_1 \\ \vdots \\ c_{p1} \\ \vdots \\ c_{pq} \\ k_q \end{bmatrix}$$

The INPU command can be used to enter this vector.

Example

If the vector of fixed parameters is $\gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}$ and the following

hypotheses are to be tested

$$\begin{aligned} \gamma_1 \quad - \quad \gamma_3 &= 0 \\ \gamma_2 &= 0 \end{aligned}$$

vector **f** must first be created. Read **f** into a column, say C20. The following can be used:

INPU C20

1 0 -1 0
0 1 0 0

To examine the contrasts specify the command

FTES C20

The outputs from this command are as follows:

- ⋆ the value of each contrast;
- ⋆ a χ^2 value for a simultaneous test of the hypotheses;
- ⋆ a χ^2 value for a test of each hypothesis individually; and
- ⋆ individual and simultaneous 95% confidence intervals.

11 Appendix A3

11.1 The ML3 Program

The datafile VOTING.DAT as created in Appendix A2 is an ASCII-format file with the four variables per observation (record) in free format. Apply ML3 to analyse the data in order to achieve goals (a) and (b). The ML3 commands used is shown below. Two previously created files MUL.COM and MUL5.COM are invoked with the OBEY command. These files contain certain commands, which could also be typed in. These files have the following commands:

1. MUL.COM:

```
name c1 'id2' c2 'id1' c3 'category' c4 'm' c5 'n'
iden 1 'id1'
iden 2 'id2'
calc c6=loge('m'/('n'-'m'))
resp c6
```

2. MUL5.COM

```
name c8 'cons'
dumm c3 c10-c13
calc c9='cons'-(c10+c11+c12+c13)
name c9 'fp1' c10 'fp2' c11 'fp3' c12 'fp4' c13 'fp5'
calc c7=1/sqrt('n')
mult c7 c9 c17
mult c7 c10 c18
mult c7 c11 c19
mult c7 c12 c20
mult c7 c13 c21
name c17 'rp1' c18 'rp2' c19 'rp3' c20 'rp4' c21 'rp5'
expl 'cons' c9-c13 c17-c21
fpar c17-c21 'cons'
setv 2 'cons'
sete 1 'rp1' 'rp1'
sete 1 'rp2' 'rp2'
sete 1 'rp3' 'rp3'
sete 1 'rp4' 'rp4'
sete 1 'rp5' 'rp5'
batc
sett
```

The ML3 statements are as follows:

```
dinp c1 c3-c5
80000 spaces left on worksheet
```

```
Type file name
->
voting.dat
  1  1   0.75     1
  1  2   6.00     9
  1  3   6.00     7
  1  4   6.75     7
  1  5  46.00    52
  2  1   4.00     6
  2  2  16.00    19
  2  3   8.00    13
  2  4  13.00    14
  2  5  28.00    32
  3  1  28.00    34
  3  2  49.00    58
  3  3  46.00    51
  3  4  42.00    50
  3  5 105.00   114
  4  1  24.00    31
  4  2  40.00    51
  4  3  51.00    55
  4  4  71.00    82
  4  5 118.00   130
  6  1  11.00    20
  6  2  72.00    82
  6  3  74.00    82
  6  4  67.00    73
  6  5 143.00   157
  7  1   5.00     6
  7  2  13.00    14
  7  3  11.75    12
  7  4  23.75    24
  7  5  38.00    43
  8  1   3.00     4
  8  2   8.00    11
  8  3  14.00    15
  8  4  17.00    21
  8  5  35.00    39
  9  1   6.00    10
  9  2  16.00    20
  9  3  13.00    15
  9  4  18.00    19
  9  5  23.00    28
 10  1  12.00    20
 10  2  26.00    32
 10  3  24.00    31
 10  4  27.00    35
 10  5  40.00    55
```

11	1	16.00	21
11	2	37.00	43
11	3	27.00	33
11	4	34.00	37
11	5	73.00	82
12	1	23.00	34
12	2	57.00	70
12	3	44.00	49
12	4	61.00	67
12	5	96.00	109
14	1	36.00	44
14	2	69.00	91
14	3	57.00	70
14	4	54.00	62
14	5	90.00	111
15	1	13.00	16
15	2	26.00	29
15	3	24.00	28
15	4	16.75	17
15	5	50.00	54
16	1	10.00	11
16	2	31.00	37
16	3	24.00	26
16	4	22.00	28
16	5	53.00	57
17	1	4.00	11
17	2	17.00	20
17	3	11.00	14
17	4	11.00	15
17	5	19.00	24
18	1	13.00	24
18	2	23.00	35
18	3	26.00	43
18	4	25.00	34
18	5	42.00	65
19	1	20.00	29
19	2	40.00	49
19	3	44.00	48
19	4	33.00	35
19	5	67.00	78
20	1	22.00	34
20	2	74.00	95
20	3	50.00	61
20	4	62.00	71
20	5	73.00	100
22	1	34.00	50
22	2	80.00	103
22	3	48.00	66
22	4	67.00	83

```
  22  5 102.00   124
  23  1   8.00    12
  23  2  30.00    33
  23  3  34.00    36
  23  4  27.00    32
  23  5  65.00    67
  24  1   8.00    10
  24  2  24.00    33
  24  3  27.00    32
  24  4  32.00    40
  24  5  49.00    58
  25  1   7.00    18
  25  2  19.00    31
  25  3  11.00    24
  25  4  15.00    17
  25  5  32.00    54
  26  1   8.00    41
  26  2  17.00    40
  26  3   9.00    16
  26  4  15.00    20
  26  5  28.00    52
  27  1  28.00    58
  27  2  53.00    83
  27  3  70.00    92
  27  4  53.00    69
  27  5  81.00    98
  28  1  40.00    91
  28  2  70.00   129
  28  3  61.00    92
  28  4  35.00    57
  28  5  67.00   100
  30  1  44.00    80
  30  2  75.00   123
  30  3  83.00   106
  30  4  69.00    87
  30  5  74.00    95
  31  1  20.00    26
  31  2  27.00    37
  31  3  30.00    32
  31  4  37.00    40
  31  5  48.00    54
  32  1  14.00    23
  32  2  22.00    40
  32  3  28.00    33
  32  4  19.00    20
  32  5  25.00    36
code 140 1 1 c2
obey mul.com
name c1 'id2' c2 'id1' c3 'category' c4 'm' c5 'n'
```

```
iden 1 'id1'
iden 2 'id2'
calc c6=loge('m'/('n'-'m'))
resp c6
code 1 1 140 c8
obey mul5.com
name c8 'cons'
dumm c3 c10-c13
calc c9='cons'-(c10+c11+c12+c13)
name c9 'fp1' c10 'fp2' c11 'fp3' c12 'fp4' c13 'fp5'
calc c7=1/sqrt('n')
mult c7 c9 c17
mult c7 c10 c18
mult c7 c11 c19
mult c7 c12 c20
mult c7 c13 c21
name c17 'rp1' c18 'rp2' c19 'rp3' c20 'rp4' c21 'rp5'
expl 'cons' c9-c13 c17-c21
fpar c17-c21 'cons'
setv 2 'cons'
sete 1 'rp1' 'rp1'
sete 1 'rp2' 'rp2'
sete 1 'rp3' 'rp3'
sete 1 'rp4' 'rp4'
sete 1 'rp5' 'rp5'
batc

BATCh mode is ON
sett
EXPLanatory variables in       CONS     FP1      FP2      FP3      FP4
FPARameters                             FP1      FP2      FP3      FP4
FMEAns
RMEAns
RESPonse variable in           C6
IDENtifying codes for level 1: ID1      level 2: ID2      level 3:
RESEtting covariances level 1: ON       level 2: ON       level 3: ON
MAXIterations   5     TOLErance   2     METHod is IGLS    BATCh is ON


LEVEL 3 RANDOM PARAMETER MATRIX unspecified
LEVEL 2 RANDOM PARAMETER MATRIX
        CONS
CONS    1
LEVEL 1 RANDOM PARAMETER MATRIX
        RP1      RP2      RP3      RP4      RP5
RP1     1
RP2     0        1
RP3     0        0        1
RP4     0        0        0        1
```

```
RP5     0       0       0       0       1
star


Iteration number 1 in progress
Iteration number 1 in completed

Iteration number 2 in progress
Iteration number 2 in completed

Iteration number 3 in progress
Iteration number 3 in completed

Iteration number 4 in progress
Iteration number 4 in completed

Iteration number 5 in progress
Iteration number 5 in completed
Convergence not achieved
next


Iteration number 6 in progress
Iteration number 6 in completed
Convergence achieved
rand
```

LEVEL 3

PARAMETER		ESTIMATE	S. ERROR	PREV. ESTIMATE	NCONV

LEVEL 2

PARAMETER		ESTIMATE	S. ERROR	PREV. ESTIMATE	NCONV
CONS	/CONS	0.3969	0.1166	0.3972	2

LEVEL 1

PARAMETER		ESTIMATE	S. ERROR	PREV. ESTIMATE	NCONV
RP1	/RP1	5.163	1.587	5.163	3
RP2	/RP2	4.59	1.718	4.574	1
RP3	/RP3	7.011	2.215	6.981	1
RP4	/RP4	16.75	4.805	16.8	3
RP5	/RP5	7.462	2.813	7.494	1

fixe

PARAMETER	ESTIMATE	S. ERROR	PREV. ESTIMATE
FP1	0.7003	0.1468	0.7004

```
FP2                          1.21        0.1337              1.21
FP3                          1.676       0.1434              1.676
FP4                          1.854       0.1705              1.854
FP5                          1.696       0.1343              1.696
save voting.ws
77433 spaces left on worksheet
stop
```

- By default there are five iteration steps. Convergence is not achieved after these five iterations. Iteration six is done after specifying the NEXT command. From the output given after specifyng the RAND and the FIXE commands respectively it is clear that convergence is achieved.The variable NCONV gives the number of steps at which the specific estimate has achieved convergence.

12 Appendix B

12.1 SAS Program with Description

The following SAS program is used to fit the models described is Sections 3.3, 3.4 and 3.5

```
PROC IML;

number=" ";
msg=" ";
window lookup color='gray' cmndline=cmnd msgline=msg group=grp
       #2 "type the number of the model wanted:"
       #3  "1=logit"
       #4  "2=cumulative logit"
       #5  "3=McCullagh's model"
       #6  "number:" number
       #8  "enter exit on the command line to exit";

create lookup var{number};
display lookup.grp;
if cmnd="exit" then append;
window close=lookup;

*DATA;
G={1142 610 329,
   1501 962 493,
   882 447 280,
   739 368 264};
R=NROW(G);
C=NCOL(G);

if number='1' then do;
  Z={1  1  1,
     1  1 -1,
     1 -1  1,
     1 -1 -1};
  X=Z@I(C-1);
  A=I(C);
  K={1 0 -1,0 1 -1};
end;
if number='2' then do;
  Z={1  1  1,
     1  1 -1,
     1 -1  1,
     1 -1 -1};
  X=Z@I(C-1);
  A={ 1 0 0,0 1 1,1 1 0,0 0 1};
  K={1 -1 0 0,0 0 1 -1};
```

```
 end;
 if number='3' then do;
   Z={1 0 0 1 1,0 1 0 1 1,0 0 1 1 1,
      1 0 0 1 -1,0 1 0 1 -1,0 0 1 1 -1,
      1 0 0 -1 1,0 1 0 -1 1,0 0 1 -1 1,
      1 0 0 -1 -1,0 1 0 -1 -1,0 0 1 -1 -1};
   X=Z;
   A={ 1 0 0,0 1 1,1 1 0,0 0 1};
   K={1 -1 0 0,0 0 1 -1};
 end;
 QS=(C-1)*R;
 SINV=J(QS,QS,0);
 HS=J(C,C-1,0);
 F=J(QS,1,0);
 FF=J(QS,1,0);

 *VECTOR OF PROBABILITIES;
 P1=G[ ,+];
 P2=G/(P1*REPEAT(1,1,C));
 P=SHAPE(P2,C*R,1);

 *VECTOR OF RESPONSE FUNCTIONS;
 DO SS=1 TO R;
   LL=((SS-1)*C)+1;
   UL=SS*C;
   ML=((SS-1)*(C-1))+1;
   KL=SS*(C-1);
   PS=P[LL:UL, ];
   FS=K*LOG(A*PS);
   FF[ML:KL, ]=FS;
 END;

 DO ITER=1 TO 15;

   *COVARIANCE MATRIX OF VECTOR OF PROBABILITIES;
   S1=J(C,C,1);
   S2=INV(DIAG(P1));
   V=(DIAG(P)-P*P`)#(S2@S1);

   *COVARIANCE MATRIX OF VECTOR RESPONSE FUNCTIONS;
   DO SS=1 TO R;
     LL=((SS-1)*C)+1;
     UL=SS*C;
     ML=((SS-1)*(C-1))+1;
     KL=SS*(C-1);
     PS=P[LL:UL, ];
     VS=V[LL:UL,LL:UL];
     FS=FF[ML:KL, ];
     EPS=.1E-6;
```

```
      DO L=1 TO C;
        PS[L, ]=PS[L, ]+EPS;
        FDS=K*LOG(A*PS);
        HS[L, ]=(FDS-FS)`/EPS;
        PS[L, ]=PS[L, ]-EPS;
      END;
      SINV[ML:KL,ML:KL]=INV(HS`*VS*HS);
    END;

    *ESTIMATING BETA;
    COV=INV(X`*SINV*X);
    BHAT=COV*X`*SINV*FF; PRINT BHAT;

    *OBTAIN A NEW P;
    IF NUMBER='1' THEN DO;
      VECT=EXP(X*BHAT);
      PI=J(R,C,0);
      DO J=1 TO R;
        ML=((J-1)*(C-1))+1;
        KL=J*(C-1);
        VT=VECT[ML:KL,];
        TOTL=VT[+, ];
        PII=1/(TOTL+1);
        DO I=1 TO (C-1);
          PI[J,I]=PII*VT[I, ];
        END;
      PI[J,C]=PII;
      END;
    END;
    IF (NUMBER='2'| NUMBER='3') THEN DO;
      VECT=EXP(X*BHAT);
      PI=J(R,C,0);
      DO J=1 TO R;
        ML=((J-1)*(C-1))+1;
        KL=J*(C-1);
        VT=VECT[ML:KL,];
        DO I=1 TO (C-1);
          PROP=VT[I, ];
          PII=PROP/(PROP+1)-PI[J,+];
          PI[J,I]=PII;
        END;
      PI[J,C]=1-PI[J,+];
      END;
    END;
   P=SHAPE(PI,C*R,1);
END;

*OTHER OUTPUT;
VECCOV=VECDIAG(COV);
```

```
SE=SQRT(VECCOV);
FEST=X*BHAT;
RESID=FF-FEST;
NPAR=NCOL(X);
DF=J(NPAR,1,1);
SSLB=J(NPAR,1,0);
PAR=J(NPAR,1,0);
WH=J(NPAR,1,0);
DO NP=1 TO NPAR;
  PAR[NP, ]=NP;
  L=J(1,NPAR,0);
  L[ ,NP]=1;
  LB=L*BHAT;
  SSLB[NP, ]=LB'*INV(L*COV*L')*LB;
  WH[NP, ]=1.0-PROBCHI(SSLB[NP, ],1);
  IF WH[NP, ]<= .1E-3 THEN WH[NP, ]=.1E-3;
  L[ ,NP]=0;
END;

*PRINT RESULTS;
COLN4={"PARAMETER" "DF" "ESTIMATE" "CHI-SQ" "PROB" "STD"};
COLN5={"ACTUAL" "PREDICTED" "RESIDUAL"};
SUMSTAT=PAR||DF||BHAT||SSLB||WH||SE;
PRINT,'ANALYSIS OF RESULTS';
PRINT SUMSTAT [COLNAME=COLN4];
DESIGN=FF||FEST||RESID;
PRINT,'RESPONSE FUNCTION';
PRINT DESIGN [COLNAME=COLN5];
run;
```

Description

Step 1

The user is prompted with a window from which the model to be fitted is chosen.

Step 2

The data is entered in the form of a matrix called **G**. According to the model chosen to be fitted a design matrix **X** is calculated. These design matrices are as described in Sections 3.3, 3.4, and 3.5 respectively. Two matrices **A** and **K** are also entered according to the model chosen to be fitted. These two matrices will be used to calculate the specific vector of response functions.

Step 3

The vector of probabilities is calculated analogous to the vector **p** in Section 3.1.

Step 4

The vector of response functions is calculated. For each value of SS the vector of response functions of subpopulation SS is calculated. All these vectors are then combined to form the vector with all the response functions, which is called **FF** in the program.

Within a $ITER$-loop steps 5 to 8 are repeated to obtain $\hat{\beta}_r$ iteratively:

Step 5

The covariance matrix **V** of the vector of probabilities is calculated

Step 6

The covariance matrix of the vector of response functions is calculated according to the theory described in Section 3.6.1. For each value of SS, the covariance matrix for the vector of response functions for subpopulation SS is calculated. For subpopulation SS, the matrix $\triangle$ is calculated as $\mathbf{HS}'$, the derivatives being obtained numerically. **VS** is the specific covariance matrix for the probabilities of subpopulation SS. The matrix **SINV** contains the inverse covariance matrices of the vectors of response functions for all the subpopulations.

Step 7

The estimate of the vector of unknown parameters is calculated as **BHAT**

Step 8

In this step, a new vector of probabilities is obtained. Again this is done according to the model chosen to be fitted. If it is the logit model (number = '1') the algorithm is:

For each value of J , the three probabilities ($c = 3$) of subpopulation J is obtained.

$$\text{TOTL} = \frac{p_{11}}{p_{13}} + \frac{p_{12}}{p_{13}} = \frac{1 - p_{13}}{p_{13}}$$

so that

$$\text{PII} = p_{13}$$

Then

for $I = 1$, p_{11} is obtained as $p_{13} \star \frac{p_{11}}{p_{13}}$

and for

$I = 2$, p_{12} is obtained as $p_{13} \star \frac{p_{12}}{p_{13}}$

The matrix **PI** contains the probabilities of all the subpopulations

If the cumulative logit model or McCullagh's model was fitted (number $=' 2'$ or number $=' 3'$) the algorithm is:

For each value of J, the probabilities of subpopulation J is calculated. For example, if each subpopulation has three probabilities ($C = 3$) then for subpopulation 1:

For $I = 1$,

$$\begin{aligned}
\text{PROP} &= \frac{p_{11}}{p_{12} + p_{13}} \\
\text{PII} &= \left(\frac{p_{11}}{p_{12} + p_{13}} \star \frac{p_{12} + p_{13}}{p_{11} + p_{12} + p_{13}}\right) - 0 \\
&= p_{11}
\end{aligned}$$

For $I = 2$,

$$\begin{aligned}
\text{PROP} &= \frac{p_{11} + p_{12}}{p_{12}} \\
\text{PII} &= \left(\frac{p_{11} + p_{12}}{p_{13}} \star \frac{p_{13}}{p_{11} + p_{12} + p_{13}}\right) - p_{11} \\
&= p_{12}
\end{aligned}$$

After the I-loop,

$$p_{13} = 1 - (p_{11} + p_{12})$$

The matrix **PI** contains the probabilities of all the subpopulations.

After obtaining **PI** the vector **P** of new probabilities is obtained by shaping the matrix **PI**. This is also the end of the steps within the *ITER*-loop. This new vector of probabilities is now used to calculate a new **BHAT** . This process is repeated until convergence for **BHAT** is achieved (it is assumed that convergence will be achieved after 15 iterations in this case).

Step 9

Estimated response functions are obtained, and residuals are calculated. To test the significance of each estimated parameter the Wald statistic with a p-value is obtained within the NP-loop.

Step 10

The results are printed.

13 Appendix C

13.1 SAS Programs with Description

The following SAS program is used to obtain an initial value for **T**.

```
*OBTAIN STARTING VALUES FOR T WHERE PHI=T*T';


 PROC IML;

 *DATA;
 DAT={67 26 19,
     90 46 29,
     50 36 13,
     43 10  9,
     20  8  5,
     26 28 11,
      8 9  5,
     14  6 4,
    129 38 29,
    155 93 44,
    100 57 24,
     70 48 18,
     71 30 15,
    107  44  37,
     48 25 14,
     67 27 16,
    245 275 116,
    245 297 136,
    146 138  70,
     93  100 76,
    104   40  34,
    120 103  41,
     91 34 22,
     61 27 31,
     65 31 22,
    180  88  55,
    145  39  46,
    183  63  51,
    309 115  51,
    370 163  57,
    158  56  38,
    101  38  20,
     45  25  5,
     61  45 20,
    30  23  10,
    22   8   8,
```

```
     38 9 10,
     71 25  29,
     58 14 14,
     47 22 16,
     49 13 24,
     76 30 34,
     48 16 25,
     38 19 15};
SUB=4;

R=NROW(DAT);
C=NCOL(DAT);
LEVEL2_N=R/SUB;

*DESIGN;
Z={1  1  1,
   1 -1  1,
   1  1 -1,
   1 -1 -1};
X=Z@I(C-1);
RX=NROW(X);
CX=NCOL(X);

*TRANSFORMATION;
A={ 1 0 0,
    0 1 1,
    1 1 0,
    0 0 1};
K={1 -1 0 0,
   0  0 1 -1};

QS=(C-1)*SUB;
SINV=J(QS,QS,0);
HS=J(C,C-1,0);
F=J(QS,1,0);
FF=J(QS*LEVEL2_N,1,0);

*VECTOR OF PROBABILITIES;
P1=DAT[ ,+];
P2=DAT/(P1*REPEAT(1,1,C));
P=SHAPE(P2,C*R,1);

*COVARIANCE MATRIX OF VECTOR OF PROBABILITIES;
START COVAR;
   S1=J(C,C,1);
   S2=INV(DIAG(P1));
   V=(DIAG(P)-P*P`)#(S2@S1);
FINISH;
```

```
START LIM1;
   LL=((SS-1)*C)+1;
   UL=SS*C;
   ML=((SS-1)*(C-1))+1;
   KL=SS*(C-1);
   PS=PP[LL:UL, ];
   VS=ES[LL:UL,LL:UL];
   FS=K*LOG(A*PS);
FINISH;

START LIM2;
   DD=(I-1)*(C-1)*SUB+1;
   EE=(C-1)*SUB*I;
   OO=((I-1)*(C*SUB))+1;
   BB=I*C*SUB;
   PP=P[OO:BB,];
   ES=V[OO:BB,OO:BB];
FINISH;

*COVARIANCE MATRIX OF VECTOR OF RESPONSE FUNCTIONS;
START RESP;
   DO SS=1 TO SUB;
     RUN LIM1;
     EPS=.1E-6;
     DO L=1 TO C;
       PS[L, ]=PS[L, ]+EPS;
       FDS=K*LOG(A*PS);
       HS[L, ]=(FDS-FS)`/EPS;
       PS[L, ]=PS[L, ]-EPS;
     END;
     SINV[ML:KL,ML:KL]=INV(HS`*VS*HS);
   END;
FINISH;

*FIXED F WITHIN EACH LEVEL;
RUN COVAR;
DO I=1 TO LEVEL2_N;
  RUN LIM2;
  DO SS=1 TO SUB;
    RUN LIM1;
    F[ML:KL, ]=FS;
  END;
  FF[DD:EE]=F;
END;

*BHAT;
BETA=J(LEVEL2_N,CX,0);
DO I=1 TO LEVEL2_N;
  RUN LIM2;
```

```
  RUN RESP;
  FFF=FF(]DD:EE]);
  BHAT=INV(X'*SINV*X)*X'*SINV*FFF;
  BETA[I, ]=BHAT';
END;
JJ=J(LEVEL2_N,1,1);
S=((BETA'*(I(LEVEL2_N)-(JJ*JJ')/LEVEL2_N)*BETA)/(LEVEL2_N-1));
U=ROOT(S);
T=U';
LRINT T;
```

Description

Step 1

The data is entered in the form of a matrix called **DAT**. The expression SUB=4 denotes the number of subpopulations, in this case 4. The number of level 2 units is calculated as LEVEL2-N. The design matrix **X** is calculated in the case of the cumulative logit model, and two matrices **A** and **K** are entered which will be used to calculate the vector of response functions.

Step 2

The vector of probabilities for matrix **DAT** is calculated as the vector **P**.

Step 3

The vector of response functions is calculated. For each value of I the vector of response functions of all the subpopulations within the I-th level 2 unit is obtained. The vector with all the response functions is called **FF**.

Step 4

For each level 2 unit, $\hat{\beta}$ is obtained as **BHAT**. For each value of I the inverse covariance matrix of the vector of response functions for the I-th level 2 unit is calculated as the matrix **SINV** (see the description of the program in Appendix B). At the end of the I-loop. the matrix **BETA** contains rowwise all the **BHAT** vectors which were obtained for each level 2 unit.

Step 5

The matrix **S** is obtained according to (4.1.14). From **S** the lower triangular matrix **T** is obtained.

The following SAS program is used to fit the multilevel model described in Section 4.1.

```
PROC IML;

*DATA;
```

```
DAT={ 67 26 19,
    90 46 29,
    50 36 13,
    43 10  9,
    20  8  5,
    26 28 11,
     8 9  5,
    14  6 4,
   129 38 29,
   155 93 44,
   100 57 24,
    70 48 18,
    71 30 15,
   107  44  37,
    48 25 14,
    67 27 16,
   245 275 116,
   245 297 136,
   146 138  70,
    93  100 76,
   104   40  34,
   120 103  41,
    91 34 22,
    61 27 31,
    65 31 22,
   180  88  55,
   145  39  46,
   183  63  51,
   309 115  51,
   370 163  57,
   158  56  38,
   101  38  20,
    45  25  5,
    61  45 20,
   30  23  10,
   22   8   8,
   38 9 10,
   71 25  29,
   58 14 14,
   47 22 16,
   49 13 24,
   76 30 34,
   48 16 25,
   38 19 15};
R=NROW(DAT);
C=NCOL(DAT);
SUB=4;
LEVEL2_N=R/SUB;
```

```
*DESIGN;
Z={1  1  1,
   1 -1  1,
   1  1 -1,
   1 -1 -1};
X=Z@I(C-1);
RX=NROW(X);
CX=NCOL(X);

*TRANSFORMATION;
A={ 1 0 0,
    0 1 1,
    1 1 0,
    0 0 1};
K={1 -1 0 0,
   0  0 1 -1};

QS=(C-1)*SUB;
QSS=(C-1)*SUB*LEVEL2_N;
VI=J(QS,QS,0);
COVFF=J(QSS,QSS,0);
HS=J(C,C-1,0);
F=J(QS,1,0);
FF=J(QS*LEVEL2_N,1,0);
T={0.2894001         0         0         0         0         0,
   0.1146744 0.2264962         0         0         0         0,
   -0.006596 -0.009131 0.0985589         0         0         0,
   -0.034332 0.0528294 0.0323078 0.1233098         0         0,
   -0.000395 0.0211859 0.0159665 -0.000749 0.0645968         0,
   -0.051754 0.0658845 0.0102423 0.0342904 -0.028799 0.1036737};
CXX=0.5*CX*(CX+1);
RXX=0.5*RX*(RX+1);
SVECYY=J(RXX,LEVEL2_N,0);
G=J(CXX,1,0);
H=J(CXX,CXX,0);
STOOR=J(CXX,CXX,0);

*VECTOR OF PROBABILITIES;
P1=DAT[ ,+];
P2=DAT/(P1*REPEAT(1,1,C));
P=SHAPE(P2,C*R,1);

*COVARIANCE MATRIX OF VECTOR OF PROBABILITIES;
START COVAR;
   S1=J(C,C,1);
   S2=INV(DIAG(P1));
   V=(DIAG(P)-P*P‘)#(S2@S1);
FINISH;
```

```
 START LIM1;
    LL=((SS-1)*C)+1;
    UL=SS*C;
    ML=((SS-1)*(C-1))+1;
    KL=SS*(C-1);
    PS=PP[LL:UL, ];
    VS=ES[LL:UL,LL:UL];
    FS=K*LOG(A*PS);
 FINISH;

 START LIM2;
    DD=(I-1)*(C-1)*SUB+1;
    EE=(C-1)*SUB*I;
    OO=((I-1)*(C*SUB))+1;
    BB=I*C*SUB;
    PP=P[OO:BB,];
    ES=V[OO:BB,OO:BB];
 FINISH;

 *COVARIANCE MATRIX OF VECTOR OF RESPONSE FUNCTIONS;
 START RESP;
    DO SS=1 TO SUB;
      RUN LIM1;
      EPS=.1E-6;
      DO L=1 TO C;
        PS[L, ]=PS[L, ]+EPS;
        FDS=K*LOG(A*PS);
        HS[L, ]=(FDS-FS)'/EPS;
        PS[L, ]=PS[L, ]-EPS;
      END;
      VI[ML:KL,ML:KL]=HS'*VS*HS;
    END;
FINISH;

*FIXED F WITHIN EACH LEVEL;
RUN COVAR;
DO I=1 TO LEVEL2_N;
  RUN LIM2;
  DO SS=1 TO SUB;
    RUN LIM1;
    F[ML:KL, ]=FS;
  END;
  FF[DD:EE]=F;
END;

*XISTER;
XISTER=J(0.5*RX*(RX+1),0.5*CX*(CX+1),0);
IJ=0;
DO IL=1 TO RX;
```

```
  DO JL=1 TO IL;
    IJ=IJ+1;
    KL=0;
    DO KK=1 TO CX;
      DO LL=1 TO KK;
        KL=KL+1;
        IF KK=1 THEN XISTER[IJ,KL]=X[IL,KK]*X[JL,KK];
        ELSE DO;
        IF LL=KK THEN XISTER[IJ,KL]=X[IL,KK]*X[JL,LL];
        ELSE XISTER[IJ,KL]=X[IL,KK]*X[JL,LL]+X[IL,LL]*X[JL,KK];
        END;
      END;
    END;
  END;
END;

*W-INVERSE;
START WINVE;
   RXI=RX*(RX+1)/2;
   WINV=J(RXI,RXI,0);
   IJ=0;
   DO ILOOP=1 TO RX;
   DO JLOOP=1 TO ILOOP;
     IJ=IJ+1;
     CDIJ=1.0;
     IF ILOOP=JLOOP THEN CDIJ=0.5;
     KL=0;
     DO KLOOP=1 TO RX;
       VIK=COVF[ILOOP,KLOOP];
       VJK=COVF[JLOOP,KLOOP];
       DO LLOOP=1 TO KLOOP;
         KL=KL+1;
         CDKL=2.0;
         IF KLOOP=LLOOP THEN CDKL=1.0;
         WINV[IJ,KL]=CDIJ*CDKL*(VIK*COVF[JLOOP,LLOOP]
                                 +COVF[ILOOP,LLOOP]*VJK);
         WINV[KL,IJ]=WINV[IJ,KL];
       END;
     END;
   END;
   END;
FINISH;

DO ITER=1 TO 6;

  *BHAT;
  XVX=J(CX,CX,0);
  XVF=J(CX,1,0);
  RUN COVAR;
```

```
PHI=T*T';
DO I=1 TO LEVEL2_N;
  RUN LIM2;
  RUN RESP;
  COVF=INV(VI+X*PHI*X');
  COVFF[DD:EE,DD:EE]=COVF;
  XVX=XVX+X'*COVF*X;
  FFF=FF[DD:EE];
  XVF=XVF+X'*COVF*FFF;
END;
BHAT=INV(XVX)*XVF;  PRINT BHAT;

*PHI;
RXI=RX*(RX+1)/2;
WINVV=J(RXI*LEVEL2_N,RXI*LEVEL2_N,0);
STQ=0;
PHI=T*T';
SVECP=SYMSQR(PHI);
YHATST=XISTER*SVECP;
DO I=1 TO LEVEL2_N;
  DD=(I-1)*(C-1)*SUB+1;
  EE=(C-1)*SUB*I;
  COVF=COVFF[DD:EE,DD:EE];
  FFF=FF[DD:EE];
  FSTAR=(FFF-X*BHAT)*(FFF-X*BHAT)';
  SVECY=SYMSQR(FSTAR);
  SVECYY[ ,I]=SVECY;
  RUN WINVE;
  STQ=STQ+(SVECY-YHATST)'*WINV*(SVECY-YHATST);
END;
Q=STQ;
NEWQ=0;
J1=J(CX,CX,0);
J2=J(CX,CX,0);
DO ITE=1 TO 15 WHILE(ABS(Q-NEWQ)>EPS);
  IJJ=0;
  PHI=T*T';
  SVECP=SYMSQR(PHI);
  YHATST=XISTER*SVECP;
  *CALCULATE G;
  DO K1=1 TO CX;
    DO L1=1 TO K1;
      ST1=0;
      IJJ=IJJ+1;
      J1[K1,L1]=1;
      J2[L1,K1]=1;
      DIFF=SYMSQR(J1*T'+T*J2);
      DO I=1 TO LEVEL2_N;
        DDD=(I-1)*RXI+1;
```

```
        EEE=RXI*I;
        RUN LIM2;
        RUN RESP;
        COVF=INV(VI+X*PHI*X');
        COVFF[DD:EE,DD:EE]=COVF;
        SVECY=SVECYY[ ,I];
        RUN WINVE;
        WINVV[DDD:EEE,DDD:EEE]=WINV;
        ST1=ST1+(SVECY-YHATST)'*WINV*XISTER*DIFF;
      END;
      G[IJJ, ]=-2*ST1;
      J1[K1,L1]=0;
      J2[L1,K1]=0;
      STOOR[ ,IJJ]=DIFF;
   END;
 END;

 *CALCULATE H;
 DO K2=1 TO CXX;
   DO L2=1 TO CXX;
     J3=STOOR[ ,K2];
     J4=STOOR[ ,L2];
     ST2=0;
     DO I=1 TO LEVEL2_N;
       DDD=(I-1)*RXI+1;
       EEE=RXI*I;
       WINV=WINVV[DDD:EEE,DDD:EEE];
       ST2=ST2+(XISTER*J3)'*WINV*(XISTER*J4);
     END;
     H[K2,L2]=2*ST2;
 END;
END;

*OBTAIN Q;
STQ=0;
DO I=1 TO LEVEL2_N;
  DDD=(I-1)*RXI+1;
  EEE=RXI*I;
  WINV=WINVV[DDD:EEE,DDD:EEE];
  SVECY=SVECYY[ ,I];
  STQ=STQ+(SVECY-YHATST)'*WINV*(SVECY-YHATST);
END;
Q=STQ;
DELTA=-INV(H)*G;
NEWDEL=J(CX,CX,0);
B=1;
DO II=1 TO CX;
  DO JJ=1 TO II;
    NEWDEL[II,JJ]=DELTA[B,1];
```

```
        B=B+1;
      END;
    END;
    NEWT=T+NEWDEL;

    *OBTAIN NEW Q;
    STNQ=0;
    PHI=NEWT*NEWT';
    SVECP=SYMSQR(PHI);
    YHATST=XISTER*SVECP;
    DO I=1 TO LEVEL2_N;
      RUN LIM2;
      RUN RESP;
      COVF=INV(VI+X*PHI*X');
      SVECY=SVECYY[ ,I];
      RUN WINVE;
      STNQ=STNQ+(SVECY-YHATST)'*WINV*(SVECY-YHATST);
    END;
    NEWQ=STNQ;
    T=NEWT;
    GGG=ABS(Q-NEWQ);
  END; PRINT PHI;

 *P FROM BHAT;
 VECT=EXP(X*BHAT);
 PI=J(SUB,C,0);
 DO J=1 TO SUB;
   ML=((J-1)*(C-1))+1;
   KL=J*(C-1);
   VT=VECT[ML:KL,];
   DO II=1 TO (C-1);
     PROP=VT[II, ];
     PII=PROP/(PROP+1)-PI[J,+];
     PI[J,II]=PII;
   END;
   PI[J,C]=1-PI[J,+];
 END;
 P=SHAPE(PI,C*R,1);
END;
```

Description

Step 1 to **Step 3** are as described previously with the entering of the matrix **T** as obtained in the previous program.

Step 4

The matrix $\mathbf{X}^\star$ (c.f. (4.1.6)) is obtained as the matrix **XISTER**. The algorithm used minimizes computer runtime. This method eliminates the need for calculating the matrices **H** and **G**.

Within an $ITER$-loop steps 5 to 7 are repeated ($\hat{\beta}$ and $\hat{\Phi}$ is obtained iteratively):

Step 5

$\hat{\beta}$ is obtained as the matrix **BHAT** according to expression (4.1.3). The design matrix $\mathbf{X}_i$ in the expression is the matrix **X** in the program, and Σ_i^{-1} is **COVF** in the program.

Step 6

The matrix $\hat{\Phi}$ or **PHI** in the program is obtained iteratively. The value of Q (cf.(4.1.8)) is obtained, where $\mathbf{y}^\star$ is **SVECY**, $\hat{\mathbf{y}}^\star$ is **YHATST** and $\mathrm{W}^{\star-1}$ is **WINV** in the program respectively. Within an ITE-loop the following is repeated:

The vector **G** as described by expression (4.1.10) is obtained. The matrix **H** is obtained as described by expression (4.1.11). A new value of **T** is obtained by applying expression (4.1.12). This value is called **NEWT**. A new value of Q is obtained by substituting **NEWT** in expression (4.1.8). The new value of Q is called NEWQ. These steps are repeated while the absolute difference between Q and NEWQ exceeds EPS which is equal to 10^{-6}. At the end of the loop **PHI** is obtained.

Step 7

The vector of new probabilities **P** is obtained as described in **step 8** of the SAS program in Appendix B. This new **P** is used to obtain a new **BHAT** etc.

In this case $\hat{\beta}$ and $\hat{\Phi}$ are obtained after 6 iterations (convergence is obtained after 6 iterations). The number of iterations necessary for convergence will be different for each problem.