SOME ASPECTS OF FITTING MULTINOMIAL MODELS IN A GLM FRAMEWORK

MUHAMMAD YOUNIS ALI

MSc. MPhil.

A thesis submitted in fulfilment of the requirements of the London Metropolitan University for the degree of Doctor of Philosophy

STORM, The Statistics Operational Research and Mathematics Research Centre

London Metropolitan University

June, 2010

Acknowledgements

I would like to thank my supervisor, Professor Robert Gilchrist Director of STORM Research Centre, London Metropolitan University. I owe a great deal of gratitude to him for his guidance and help through out my research. I can only say that I have enjoyed his supervision over the years, and he has never been less than totally encouraging in my endeavours to become a good statistician. It is largely due to his perseverance, and to his willingness to help with my numerous research problems, that I am in a position to submit this thesis successfully.

Many more people have contributed to my statistical education over the years it has taken to research this thesis. I acknowledge the valuable suggestions of Dr. D. Stasinopoulos, Dr. R. Rigby of the STORM Research Centre and Dr. P. Calay of Mathematics, London Metropolitan University in different stages of my research. I am also grateful to all members of the School of Computing, Communication Technology and Mathematics of London Metropolitan University, members of staff, secretaries and colleagues for all the help they have given to me.

Finally, I would like to thank my family, especially my mother for their continuing support and for encouraging me to pursue my studies.

I dedicate this work to the memory of my grand mother

Azmait Bibi

who will always be missed

To Love

Abstract

This thesis discusses how an iteratively re - weighted least squares algorithm can be used to fit a multinomial regression model, with logit link function or own link functions, with any number of explanatory variables. The responses of each individual can be aggregated and the data can then be represented in a contingency table as are given in examples used in this study.

A satisfying aspect of the iteratively re-weighted least squares (IRLS) algorithm gives for fitting a multinomial regression model is that the calculations only require a program which can handle ordinary least squares and hence can be handled by a range of standard statistical software.

The approach in this thesis applies an interesting and simple form of the Cholesky decomposition to a matrix that consists of diagonal sub - matrices to find the iterative weight matrix W_{ij}. This method requires no matrix algebra facilities as all the calculations are carried out in an array format. This makes it amenable to implementation in most statistical software, including GLIM and shows how to fit a multinomial logit model without recourse to the Poisson trick approach of Francis et al. (1992).

The method given here also allows us to find the 'hat - matrix' as is needed in the calculation of leverages and Cook's distances. These statistics in general can be used for diagnostic purposes or to detect the influential observations but in multinomial models the 'hat - matrix' may have a very little or no use at all for detecting any inappropriate observation. Our approach is in contrast to any potential naive use of the Poisson trick approach of Francis et al. (1992) model that would then produce inappropriate leverages and Cook's distances.

We check that our approach gives exactly the same scaled deviance with correct degrees of freedom, parameter estimates and standard errors as are obtained from the Poisson trick approach with some minor rounding errors. Our approach has the freedom to

consider any number of response levels or explanatory variables for fitting a multinomial regression model.

The method given here in this research for the multinomial data is quite general. It allows us to use different link functions, as is explained in more detail in chapter 4. We concentrate on use of so - called Box - Cox links. Interval estimates for different parameters in these user - defined 'own' link functions are also given in section 4.6, although, in the first three chapters the main concentration is on the logit link functions. Chapter 5 gives a suggestion as to how we can get sharper convergence by using re - parameterisation of the design matrix. This is useful as in some cases as without such a modified design matrix we may not get convergence even after 1000 iterations.. Chapter 6 indicates with examples that the theory developed in the previous chapters works well in a more general form of a multinomial data and discusses how the startup values must be considered with some knowledge of the data. Chapter 7 suggests how to calculate the likelihood influence measure of Cook (1986) by the idea of a single case *i* deletion for our fitting of multinomial models and for the Poisson trick approach of Francis et al. (1992). This can be further used to detect any influential observations for the regression parameter estimate of $\hat{\beta}$ only. Chapter 8 presents the conclusions and gives some guidelines for further extensions of the ideas presented in this thesis.

Contents

Abbreviations / Notations

The following abbreviations and notations are used in this study

i *ith* case, $i = 1, 2, 3, \ldots, m$

j *jth* level of response variable, $j = 1, 2, 3, \ldots, J$

r *rth* stacked observation, $r = 1, 2, 3, \ldots, n$

y_{ij} *ith* observation for *jth* level of response variable (in cell (i, j))

y_r *rth* observation in stacked data

z_i Cell counts, $i = 1, 2, 3, \ldots, n_c = mJ$

n Total stacked observations $= m(J-1)$

n_i *ith* case total, $n_i = \sum_{j=1}^{J} y_{ij}$

n_c Total number of cells, $n_c = mJ$

m Total number of cases

J Levels of the response variable

K Number of explanatory variables

p Number of parameters in the fitted model

p_{ij} Probability for cell (i, j)

h_r Leverage coefficient in our approach

r_r Residual coefficient in our approach

r_{pr} Pearson residual in our approach for y_r

r_r^P Modified Pearson residual

r_r^{PS} Standardized Pearson residual

$r_r^{PS'}$ Studentized residual

r_r^D Deviance residual

C_r Cook's distance (1977) for y_r

θ_{ij} Log - odds of *ith* case and *jth* level of response variable

η_{ij} Linear predictor for cell $(i, j) = \sum_{k=1}^{K} x_{ik}^{(j)} \beta_k^{(j)} = \eta^{(j)}$

L_i Likelihood function for case i , $i = 1, 2, 3, \ldots, m$

l_i Log - likelihood function for case i , $i = 1, 2, 3, \ldots, m$

η^T Row vector $(\eta^{(1)T}, \eta^{(2)T}, \ldots, \eta^{(J)T})$ of length mJ

β^T Row vector $(\beta^{(1)T}, \beta^{(2)T}, \ldots, \beta^{(J)T})$ of length KJ

$(\%\, lp)$ In our iterative fits it is assigned the linear predictor by the fit directive

$(\%\, fv)$ For non - iterative models $(\%\, lp) = (\%\, fv)$

A_{ii} Diagonal matrix in Cholesky decomposition

$\boldsymbol{D}$ Matrix of $\dfrac{\partial \eta_i}{\partial \beta_j}$ in our notations

$\boldsymbol{H}$ Hat - matrix

$\boldsymbol{N}$ New design matrix in our notation

$\boldsymbol{W}$ Weight matrix in our notation

$x_i^{(j)}$ *ith* explanatory variable for response level j

$\boldsymbol{X}^{(j)}$ Matrix of explanatory variable (with 1's in the 1st column)

GLIM Generalized linear iteractive modelling

GLM Generalized linear model

IRLS Iteratively re - weighted least squares

General multinomial data notations are given as follows,

Case	Level of response variable						Totals
	1	2	.....	j	.....	J	
	Observations						
1	y_{11}	y_{12}	.....	y_{1j}	.....	y_{1J}	n_i
2	y_{21}	y_{22}	.....	y_{2j}	.....	y_{2J}	n_2
:	.....						
i	y_{i1}	y_{i2}	.....	y_{ij}	.....	y_{iJ}	n_i
:	.....						
m	y_{m1}	y_{m2}	.....	y_{mj}	.....	y_{mJ}	n_m

The stacking of data is adopted as follows,

$$\text{Response variable level 1} \quad \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{i1} \\ \vdots \\ y_{m1} \end{bmatrix}$$

$$\text{Response variable level 2} \quad \begin{bmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{i2} \\ \vdots \\ y_{m2} \end{bmatrix}$$

$$\vdots$$

$$\vdots$$

$$\text{Response variable level } j \quad \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{mj} \end{bmatrix}$$

$$\vdots$$

$$\vdots$$

$$\text{Response variable level } J-1 \quad \begin{bmatrix} y_{1J-1} \\ y_{2J-1} \\ \vdots \\ y_{iJ-1} \\ \vdots \\ y_{mJ-1} \end{bmatrix}.$$

List of Tables

List of Figures

CHAPTER 1

Review and Some Basic Theory

1.1 Introduction

In this chapter we will review and give some basic theory for a new method developed in this thesis to fit a multinomial logit model using the direct iteratively re - weighted least squares (IRLS) algorithm. Our approach of fitting the model is based on using GLIM (an acronym for Generalized Linear Interactive Modelling), a statistical modelling package developed by the Royal Statistical Society's GLIM working party and it can be formulated using some other statistical packages. An important aspect of our algorithm is that the calculations only require a simple program, which can handle using ordinary least squares. The method uses Cholesky decomposition applied to a matrix that has diagonal sub - matrices and makes the required matrix inverse straightforward to evaluate using standard statistical software that can only handle arrays (*e.g.* GLIM).

We can extract easily all the appropriate statistics using GLIM codes and will investigate the hat - matrix, leverages, Cook's distances, parameter estimates, standard errors, fitted values, Pearson's coefficient of correlation and residuals for fitting the multinomial logit model. It is noted that even when we find the appropriate hat - matrix, leverages, Cook's distances and residuals but cannot be defined easily for the multinomial data. We will compare our results of fitting a multinomial logit model and the theory with the Poisson trick approach of Francis et al. (1992).

In our approach for fitting the multinomial model we can use our own link functions with some knowledge about the start - up values in the macros. The modified design matrix may be used if the convergence is hard to achieve for lots of 0's and 1's in the response variable or for the data in an individuals each case i.

It may be considered that it is possible in our approach to use standard statistical techniques to detect any influential case i as is defined in section (1.8) for the multinomial data by investigating the leverages and Cook's distances. The regression diagnostics for the normal linear models using leverages and Cook's distance (1977) are well established in the literature and have been surveyed comprehensively by Cook and Weisberg (1982), Chatterjee and Hadi (1986), Belsley et al. (1980) and Wetherill (1986).

Many of these diagnostics use statistics that measure the effects of deleting a single case from the data. These statistics exploit the exact algebraic relationship between the least squares fit of the linear model to a complete set of m cases, and the fit of the $m-1$ cases remaining after the deletion of a single case. The maximum likelihood (ML) estimation of most generalized linear models (GLMs) requires iterative methods. The maximum likelihood estimates from $m-1$ cases cannot then be obtained as an explicit function of the results of the fit of all the m cases. Pregibon (1981) derives a useful one step approximation for the changes in the maximum likelihood estimate and deviance of the model when the single case is deleted, and he discusses some diagnostic methods that use these approximations.

Cook and Weisberg (1982) discuss GLM diagnostics briefly in section (5.4) and they make some use of the Pregibon's results. McCullagh and Nelder (1989) discuss diagnostics in model checking. Williams (1987) described GLM model diagnostics using the deviance and single case deletions.

Unfortunately, for the multinomial data with a J - level of response variable, the position is not so simple, as a 'case' typically depends upon $J-1$ observations. Hence a single point deletion is not adequate; in particular, Cook's distances or the hat - matrix cannot be used in their usual way for detecting the influential cases and in chapter 7 we suggest the ways to find the likelihood influence measures of Cook (1986) for detecting any influential case i, $i=1, 2, \ldots, m$.

In chapter 2, we discuss how an iteratively re - weighted least squares (IRLS) algorithm can be used to fit a multinomial regression model and illustrated this at each stage using

two different examples of multinomial data. We considered in this chapter a multinomial logit model for three level of response variable only. The extensions of the multinomial regression models those are given in chapter 2 are followed in chapter 3, for more then three level of response variable with any number of explanatory variables or for the different explanatory variables at each level of response variable. The design matrix and the $\mathbf{y}$ - variable for $J = k$ any arbitrary level of response variable are also given to fit a multinomial logit model.

In chapter 4, the multinomial model is extended further for an 'own link function' instead of the logit link function. Confidence limits for own link function parameters are also given in this chapter. The link function proposed is shown to be equivalent to the logit link function as the extra parameter tends to zero. This idea is quite useful for fitting any appropriate our own link function in a multinomial model and the extension is easily applicable for different own link functions with different parameters for each level of a response variable.

In chapter 5, an alternative form of a design matrix is derived using the idea of spectral decomposition of the covariance matrix of the parameter estimates from the fitting of a multinomial logit model. This alternative form of a design matrix improves convergence in fitting the multinomial logit model. Indeed, in some cases we need only a few iterations to get the same scaled deviance as can be obtained from the Poisson trick approach of Francis et al. (1992). The framework of this alternative design matrix is easy to generalize for any level of response variable.

In chapter 6, we will apply the theory and the methods developed in this study to fit a multinomial logit model for an appropriate selected model. The model may have been chosen from a forward or backward selection method or from any available criteria. We have fitted the selected models using the logit link function, our own link function and with the improved design matrix method. These different models results are compared with each other and it is found that for response levels 0's and 1's, the Poisson trick approach of Francis et al. (1992) is not a good one and gives unstable parameter estimates with very high standard errors. The multinomial models fitted using our approach here as is expected gives appropriate results even with the response levels 0's and 1's.

In chapter 7 we illustrated the method of finding the likelihood influence measure of Cook (1986) on the parameter estimate $\hat{\beta}$ only (as defined in section (1.13)) by using a single case deletion for a multinomial logit model. The estimates for the deleted cases are

needed in section (7.5) to find the likelihood influence measures $LD(\hat{\beta}_{(i)})$. The method is easy to extend for any levels of response variable or with any number of explanatory variables. The macros are given in Appendix A with an explicit expression for calculating $l(\hat{\beta})$ and $l(\hat{\beta}_{(i)})$. The likelihood influence measures of Cook (1986) for different other regression parameter estimates as are defined briefly in section (1.10) can be found with the same ideas given in this chapter.

Chapter 8 gives the conclusions and some suggestions for further areas of investigation for multinomial regression models with or without an own link function.

1.2 Multinomial Distribution

The multinomial distribution is in many ways the most natural distribution to model a multi - level response variable. The application occurs in social survey data and as such is given in McCullagh and Nelder (1989) for an infinitely large population and a simple random sample of size n is taken, the question is raised as to how many individuals will be observed to have attribute A_j ? The answer is given by the multinomial distribution

$$P_r(Y_1 = y_1, \ldots, Y_J = y_J ; n, p) = \binom{n}{y} p_1^{y_1} \ldots p_J^{y_J}, \qquad (1.1)$$

where $p_1, \ldots, p_J$ $(\sum_{j=1}^{J} p_j = 1)$ are the attribute frequencies in an infinite population of interest that possesses one and only of the J attributes $A_1, \ldots, A_J$ and

$$\binom{n}{y} = \frac{n!}{y_1 ! \ldots y_J !}, \qquad \sum_{j=1}^{J} y_j = n, \qquad 0 \le y_j \le n.$$

Another way to express the multinomial distribution is that, if $Y_1, Y_2, \ldots, Y_J$ are independent Poisson random variables with means $\mu_1, \mu_2, \ldots, \mu_J$, then the joint

conditional distribution of $Y_1, Y_2, \ldots, Y_J$ given that $Y_1 + Y_2 + \ldots + Y_J = Y_\bullet = n$ is given by equation (1.1) with $p_j = \frac{\mu_j}{\mu_\bullet}$, $\mu_1 + \mu_2 \ldots + \mu_J = \mu_\bullet$.

The multinomial distribution equation (1.1) can further be re - written as

$$f_Y(y/p) = \binom{n}{y} p_1^{y_1} \ldots p_J^{y_J}$$

$$= \binom{n}{y} \exp\{ y_1 \log p_1 + y_2 \log p_2 + \ldots + y_J \log p_J \}$$

$$= \binom{n}{y} \exp\left\{ \sum_{j=2}^{J} y_j \log \left(\frac{p_j}{p_1}\right) + n \log p_1 \right\}. \qquad (1.2)$$

The above equation (1.2) is basically a vector parameter $\mathbf{p} = (p_1, p_2, \ldots, p_J)^T$ exponential family of the multinomial model equation (1.1) and can be denoted as

$$f_Y(y/p) = h(y) \exp\left\{ \sum_{j=2}^{J} \eta_j(p) T_j(y) - A(p) \right\}. \qquad (1.3)$$

We define from above equations

- natural parameter $\eta_j(p) = \log \left(\frac{p_j}{p_1}\right)$, $j \geq 2$
- sufficient statistics $T_j(y) = y_j$ $j \geq 2$
- log - partition function $A(p) = - n \log (p_1)$
- base measure $h(y) = \binom{n}{y}$.

- $E(y_j) = n p_j \qquad Var(y_j) = n\, p_j\, (1 - p_j)$

and $\qquad Cov(y_j, y_{j^*}) = n\, p_j\, p_{j^*} \qquad j \neq j^*$.

In our approach for the multinomial data with J - level of response variable of m 'cases' we need to expand our notation to consider the corresponding quantities for case $i = 1, 2, \ldots, m$ and we also need to define $y_{i1}, y_{i2}, \ldots, y_{iJ}$ as independent Poisson variables, conditional on $\sum_{j=1}^{J} y_{ij} = n_i$.

For further explanations to motivate the theoretical development of our research we consider the sort of data considered by Collier et al. (2003), being data on whether young people intended to enter UK Higher Education (HE). The response is a three level variable; the levels are as (*i*) definitely plans to enter HE, (*ii*) may possible enter HE and (*iii*) definitely will not enter HE. In general, the response variable is to be explained by a number of explanatory variables but in this social survey the explanatory variables gender and age are nominal variables; we refer to these variables as explanatory factors. In this situation, the responses of each individual can be aggregated, so the data can be represented as a 3 - way contingency table. The following table is an illustrative example of a three level of response variable. (The actual data in Collier et al. (2003) is much larger).

Gender	Age	Plans to Enter HE			Total
		Definitely Yes	Possibly	Definitely No	
Male	<21	6	9	5	20
	21+	5	4	1	10
Female	<21	1	3	11	15
	21+	6	9	6	21

Table 1.1: 3 - way contingency table

The observations 6, 9, 5, . . . , 9, 6 are the observed cell counts but the data collection can be viewed as a multinomial response model (1.1). The first row of the Table 1.1 may be viewed as 20 males < 21 years old, of whom 6 definitely plan to enter HE, 9 possibly plan

to enter HE and 5 definitely do not plan to enter HE. Similarly we interpret the other rows. It is important to note here that the explanatory variables (Age and Gender) are regarded as nominal; they are often called as factors.

The straightforward way to model the data in Table 1.1 as a so - called multinomial logit model is via the so - called Poisson trick approach (McCullagh and Nelder, 1989). In this technique, the cell counts z_i , $i = 1, 2, \ldots, n_c$ in the multi - way table are treated as independent Poisson response variates, with a constraint added that the fitted values must equal the observed values in the marginal table of the explanatory factors.

For the Poisson approach, the log - link is usually used *i.e.* $\eta_i = \log \mu_i$, $i = 1, 2, \ldots, n_c$, where the linear predictors η_i are linear combinations of the explanatory factors and $\mu_i = E(z_i)$. The Poisson assumption with the added constraints and the assumption of log - link results is the so - called multinomial logit model.

The required 'fixing of the margins' is achieved by ensuring that all linear predictors contain the multi - level interaction of all the explanatory factors *e.g.* in Table 1.1, we must include Gender * Age. The Poisson trick method can be used (Aitkin and Francis, 1992) when some of the factors are continuous variates by expanding the data to the individual level, with a nuisance parameter for each distinct set of covariates, although the number of cases can get large.

Our approach presented here in this study will overcome this problem directly.

1.3 Notation

In general, we shall assume we have m observed values of a J - level of response variable, which we denote by y_{ij}, $i=1, 2, \ldots, m$, $j=1, 2, \ldots, J$, as is shown in Table 1.2 and in the formulation of Table 1.1, we have $m = 4$ and $J=3$. Following Payne et al. (1993), we can use the neutral name 'case' to describe m rows in the Table 1.2 and the *jth* column represents the *jth* level of response for each case. The m - elements of vectors $\mathbf{y}_j$, $j=1,2,\ldots,J$ are defined as columns of the Table 1.2, as are shown below;

Responses				Totals
y_{11}	y_{12}	.....	y_{1J}	n_1
y_{21}	y_{22}	.....	y_{2J}	n_2
.....				.....
.....				.....
y_{m1}	y_{m2}		y_{mJ}	n_m

Table 1.2: Table of counts

A basic assumption is that the totals n_i, $i = 1, 2, \ldots, m$ are fixed by design, where $\sum_{j=1}^{J} y_{ij} = n_i$. We may note that it is possible that the data are arranged so that each $n_i = 1$, $i = 1, 2, \ldots, m$, although it is often convenient to aggregate the data so that $n_i \neq 1$. We shall assume that the $y_{i1}, y_{i2}, \ldots, y_{iJ}$ follow a multinomial distribution for each i, $i = 1, 2, \ldots, m$. We therefore assume that corresponding to each y_{ij} in Table 1.2, there is a probability p_{ij} which we wish to estimate, where $\sum_{j=1}^{J} p_{ij} = 1$, for $i = 1, 2, \ldots, m$, as is in Table 1.3,

Responses				Totals
p_{11}	p_{12}	.....	p_{1J}	1.0
p_{21}	p_{22}	.....	p_{2J}	1.0
.....				.....
.....				.....
p_{m1}	p_{m2}		p_{mJ}	1.0

Table 1.3: Table of probabilities

since the sum of probabilities in each row is 1.0. Therefore, there are $J-1=J^{*}$ distinct probabilities to estimate for each case $i=1, 2, \ldots, m$, so it is convenient to write the likelihood for each case i in terms of the J log - odds $\theta_{i2}, \theta_{i3}, \ldots, \theta_{iJ}$, where

$$\theta_{i2} = \log\left(\frac{p_{i2}}{p_{i1}}\right), \quad \theta_{i3} = \log\left(\frac{p_{i3}}{p_{i1}}\right), \ldots, \theta_{iJ} = \log\left(\frac{p_{iJ}}{p_{i1}}\right).$$

For each $i=1, 2, \ldots, m$, the likelihood is given after ignoring the constant term as

$$L_i = p_{i1}^{y_{i1}} \cdot p_{i2}^{y_{i2}} \cdot p_{i3}^{y_{i3}} \ldots p_{iJ}^{y_{iJ}} . \tag{1.4}$$

Now for clarity and simplification, we illustrate this technique for considering a special case for $J=3$ and the more general case follows similarly. The likelihood for case i for $J=3$, after ignoring the constant term is given by

$$L_i = p_{i1}^{y_{i1}} \cdot p_{i2}^{y_{i2}} \cdot p_{i3}^{y_{i3}} . \tag{1.5}$$

Then the log - likelihood for case i is

$$l_i = \log L_i = y_{i1} \log p_{i1} + y_{i2} \log p_{i2} + y_{i3} \log p_{i3} \tag{1.6}$$

$$= (n_i - y_{i2} - y_{i3}) \log p_{i1} + y_{i2} \log p_{i2} + y_{i3} \log p_{i3}$$

$$= (n_i - y_{i2} - y_{i3}) \log (1 - p_{i2} - p_{i3}) + y_{i2} \log p_{i2} + y_{i3} \log p_{i3}$$

$$= n_i \log (1 - p_{i2} - p_{i3}) - y_{i2} \log (1 - p_{i2} - p_{i3}) +$$

$$y_{i2} \log p_{i2} - y_{i3} \log (1 - y_{i2} - y_{i3}) + y_{i3} \log p_{i3}$$

$$= n_i \log (1 - p_{i2} - p_{i3}) + y_{i2} \log \left(\frac{p_{i2}}{1 - p_{i2} - p_{i3}}\right) + y_{i3} \log \left(\frac{p_{i3}}{1 - p_{i2} - p_{i3}}\right)$$

$$= n_i \log (1 - p_{i2} - p_{i3}) + y_{i2} \log \left(\frac{p_{i2}}{p_{i1}}\right) + y_{i3} \log \left(\frac{p_{i3}}{p_{i1}}\right)$$

$$= n_i \log (p_{i1}) + y_{i2} \log \left(\frac{p_{i2}}{p_{i1}}\right) + y_{i3} \log \left(\frac{p_{i3}}{p_{i1}}\right). \qquad (1.7)$$

Using the log - odds form, $\theta_{i2} = \log\left(\frac{p_{i2}}{p_{i1}}\right) \Rightarrow p_{i2} = p_{i1} e^{\theta_{i2}}$,

$$\theta_{i3} = \log\left(\frac{p_{i3}}{p_{i1}}\right) \Rightarrow p_{i3} = p_{i1} e^{\theta_{i3}}.$$

Therefore $p_{i1} + p_{i2} + p_{i3} = p_{i1} (1 + e^{\theta_{i2}} + e^{\theta_{i3}}) = 1.$

We can rewrite the equation (1.5) in terms of the log - odds as follows,

$$l_i = \log L_i = -n_i \log (1 + e^{\theta_{i2}} + e^{\theta_{i3}}) + y_{i2}\theta_{i2} + y_{i3}\theta_{i3}. \qquad (1.8)$$

The overall likelihood is given by

$$l = \sum_{i=1}^{m} \log L_i = \sum_{i=1}^{m} \{-n_i \log (1 + e^{\theta_{i2}} + e^{\theta_{i3}}) + y_{i2}\theta_{i2} + y_{i3}\theta_{i3}\}. \qquad (1.9)$$

1.4 Link Function

In order to fit a multinomial logit model, we assume that the probabilities p_{ij} are related to the explanatory variables through the logit link: in other words for the $(i, j)th$ cell in Table 1.2 corresponding to case i and response level j we have one or more explanatory variables, which we denote by $x_{ik}^{(j)}$ $k = 1, 2, \ldots, K$. Thus there are K different explanatory variables for each level of response variable and for each case. The normal convention (Payne et al. 1993) is to denote the linear predictor for cell (i, j) by

$$\eta_{ij} = \sum_{k=1}^{K} x_{ik}^{(j)} \beta_k^{(j)} . \tag{1.10}$$

The logit link makes the canonical link assumption that $\theta_{ij} = \eta_{ij}$; that is the log - odds are linearly related to the explanatory variables (factors). There may be some other forms of relationships between θ_{ij} and η_{ij}, as are explained more in chapter 4.

Therefore the logit link can be formulated using equation (1.8) as

$$\log \left(\frac{p_{ij}}{1 - p_{i2} - p_{i3} \ldots - p_{iJ}} \right) = \sum_{k=1}^{K} x_{ik}^{(j)} \beta_k^{(j)} . \tag{1.11}$$

Note that this is the only link formulation possible when using the Poisson trick approach but our approach allows other link functions such as are defined in chapter 4 and can be easily extended further. To summarise, the logit link assumes for cell (i, j)

$$\theta_{ij} = \eta_i^{(j)} = \sum_{k=1}^{K} x_{ik}^{(j)} \beta_k^{(j)} , \tag{1.12}$$

or, using matrix notation,

$$\eta^{(j)} = \theta^{(j)} = X^{(j)} \beta^{(j)} , \quad j = 1, 2, \ldots, J . \tag{1.13}$$

It is usual to assume that $x_{ik}^{(j)} = x_{ik}$, $j = 1, 2, \ldots, J$; that is; $X^{(j)} = X$. This assumption is, however, is not necessary in our approach for fitting multinomial logit model. Although our approach can be used more generally, and further details are given in the following chapters, we concentrate in this chapter only on the multinomial logit link, for which $\eta_i^{(j)} = \theta_{ij}$, *i.e.* $\eta^{(j)} = \theta^{(j)}$ for $j = 1, 2, \ldots, J$ and $\beta^{(j)} = \beta$.

One of our main objectives is to estimate the parameter β. Our approach is to 'stack' the data; *i.e.* to set $y^T = (y_1^T, y_2^T, \ldots, y_J^T)$ to be the row - vector of length mJ and $\beta^T = (\beta^{(1)T}, \beta^{(2)T}, \ldots, \beta^{(J)T})$ to be a row - vector of length KJ. We define the mJ column vector $\mu = E(Y)$ and the mJ row - vector $\eta^T = (\eta^{(1)T}, \eta^{(2)T}, \ldots, \eta^{(J)T})$.

Similarly we define $\theta^T = (\theta^{(1)T}, \theta^{(2)T}, \ldots, \theta^{(J)T})$ as an mJ row - vector. In usual way, we shall let β_i, η_i and θ_i be the *ith* element of β, η and θ respectively.

The overall design or model matrix is defined as

$$D = \begin{pmatrix} X^{(1)} & 0 & \ldots\ldots\ldots & 0 \\ 0 & X^{(2)} & \ldots\ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots & \ldots\ldots \\ 0 & 0 & \ldots\ldots\ldots & X^{(J)} \end{pmatrix} . \qquad (1.14)$$

For logit link we have that $\eta_{ij} = \theta_{ij}$.

Therefore $$\eta_{i2} = \theta_{i2} = \log\left(\frac{p_{i2}}{p_{i1}}\right) = \log\left(\frac{\mu_{i2}}{\mu_{i1}}\right)$$

$$= \log\left(\frac{\mu_{i2}}{n_i - \mu_{i2} - \mu_{i3}}\right). \qquad (1.15)$$

Similarly, $$\eta_{i3} = \theta_{i3} = \log\left(\frac{\mu_{i3}}{n_i - \mu_{i2} - \mu_{i3}}\right).$$

Thus the equation (1.9) can be rewritten in terms of the μ_{ij},

$$l_i = \log L_i = -n_i \log (1 + \frac{\mu_{i2}}{n_i - \mu_{i2} - \mu_{i3}} + \frac{\mu_{i3}}{n_i - \mu_{i2} - \mu_{i3}})$$

$$+ y_{i2} \log(\frac{\mu_{i2}}{n_i - \mu_{i2} - \mu_{i3}}) + y_{i3} \log (\frac{\mu_{i3}}{n_i - \mu_{i2} - \mu_{i3}})$$

$$= y_{i2} \log \mu_{i2} + y_{i3} \log \mu_{i3} + (n_i - y_{i2} - y_{i3}) \log (n_i - \mu_{i2} - \mu_{i3}) - n_i \log n_i$$

$$= y_{i2} \log \mu_{i2} + y_{i3} \log \mu_{i3} + (n_i - y_{i2} - y_{i3}) \log (n_i - \mu_{i2} - \mu_{i3}) - \text{constant.} \qquad (1.16)$$

1.5 Deviance

The essential aspect of GLIM's fitting procedure is to minimize a measure of discrepancy (called the scaled deviance in GLIM) between the observed data and the corresponding fitted values. The GLIM package fits a model by choosing an estimate of the β's to be those values which give μ's that minimize the deviance. The actual form of the deviance depends upon which member of the exponential family GLIM has been instructed to use. The GLIM user does not need to know the exact formula for the deviance if using a standard model (but in our case we do not use a standard model), nor how GLIM finds the fitted μ's which minimize the deviance, as GLIM does all the calculations internally. In general this technique is the same as fitting the maximum likelihood estimates of the parameters β_j. Thus, GLIM can be thought of as a program for the maximum likelihood estimation in generalized linear models.

The deviance function is twice the difference between the maximum achievable log - likelihood and that attained under the fitted model. This can be expressed symbolically for three levels of response variable as from equation (1.14), as follows:

$$\log L_{(\mu_i)} - \log L_{(y_i)} = y_{i2} \log \mu_{i2} + y_{i3} \log \mu_{i3} + (n_i - y_{i2} - y_{i3}) \log(n_i - \mu_{i2} - \mu_{i3})$$

$$- \{y_{i2} \log y_{i2} + y_{i3} \log y_{i3} - (n_i - y_{i2} - y_{i3}) \log(n_i - y_{i2} - y_{i3})\}$$

$$= y_{i2} \log \frac{\mu_{i2}}{y_{i2}} + y_{i3} \log \frac{\mu_{i3}}{y_{i3}} - (n_i - y_{i2} - y_{i3}) \log \frac{(n_i - \mu_{i2} - \mu_{i3})}{(n_i - y_{i2} - y_{i3})}.$$

Hence we can have

$$-2[\log L_{(\mu_i)} - \log L_{(y_i)}] = -2[y_{i2} \log \frac{\mu_{i2}}{y_{i2}} + y_{i3} \log \frac{\mu_{i3}}{y_{i3}}$$

$$- (n_i - y_{i2} - y_{i3}) \log \frac{(n_i - \mu_{i2} - \mu_{i3})}{(n_i - y_{i2} - y_{i3})}].$$

Thus

$$\sum_{i=1}^{m} \{-2[\log L_{(\mu_i)} - \log L_{(y_i)}]\} = \sum_{i=1}^{m} \{-2[y_{i2} \log \frac{\mu_{i2}}{y_{i2}} + y_{i3} \log \frac{\mu_{i3}}{y_{i3}}$$

$$- (n_i - y_{i2} - y_{i3}) \log \frac{(n_i - \mu_{i2} - \mu_{i3})}{(n_i - y_{i2} - y_{i3})}]\}. \qquad (1.17)$$

We have formulated this in a GLIM program and this is given in Appendix A for the examples presented in this study (with extension to more than 3 - levels of response variable, as appropriate).

1.6 Hat - matrix

The hat - matrix H in generalized linear models framework is given as

$$H = V^{-\frac{1}{2}} X (X^T V^{-1} X)^{-1} X^T V^{-\frac{1}{2}} ,$$

and is used to find the fitted values $\hat{Y} = HY$. The matrix H provides a measure of leverage acting on Y to produce $\hat{Y}$ but in our approach the leverage coefficient $h_r = x_r^T v_r^{-\frac{1}{2}} (X^T V^{-1} X)^{-1} v_r^{-\frac{1}{2}} x_r$, of an observation y_r may, for example, be extracted in GLIM as **%lv**, which enable the user to 'standardize' the GLIM residuals and can be interpreted as the amount of leverage the value y_r has in producing $\hat{y}_r$ regardless of the actual value (Hoaglin and Welsch 1987). It is a measure of remoteness of the *rth* observation from the remaining $n-1$ observations in the space of design matrix of the stacked data.

It is easy to see the following in our approach from the hat - matrix:

1• $\sum_{r=1}^{n} \sum_{q=1}^{n} h_{rq} = \sum_{r=1}^{n} h_r = p$, where p is the number of parameters

2• $\sum_{r=1}^{n} h_{rq} = 1$

3• $0 \leq h_r \leq 1$ for all r

4• $h_r h_q \geq h_{rq}^2$ for all r and q

5• $-0.5 \leq h_{rq} \leq 0.5$

6• $h_r \geq \frac{1}{n}$ [Wetherill (1986)]

The literature on investigating the leverages gives some bounds on the values of h_r

and as the values of $h_r \geq \frac{2p}{n}$ are taken by Belsley et al. (1980) to indicate observations with sufficiently high leverage to require some further investigations. Velleman and Welsch (1981) consider h_r to be large when $h_r \geq \frac{3p}{n}$. For a definition when the leverage coefficient h_r is large, Huber (1981) suggested to break the range of possible values $(0 \leq h_r \leq 1)$ into three intervals: values $h_r < 0.2$ appear to be safe, values $0.2 \leq h_r \leq 0.5$ are risky and values $h_r > 0.5$ should be avoided. Some further cut - off points for the leverage coefficient h_r are also suggested in the literature [see Hadi (1992)].

We find the hat - matrix in our modified approach for multinomial models by using the equation (2.31) in chapter 2 to define H as,

$$H = N(N^T N)^{-1} N^T,$$

where N is a new 'design matrix' and is defined as

$$N = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}_{2m \times 2m} \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix}_{2m \times 4}$$

$$= \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} \\ 0 & A_{22}X^{(3)} \end{pmatrix}_{2m \times 4}. \qquad (1.18)$$

The diagonal matrices A_{ij} are defined later in section (2.3) and the hat - matrix H has a very special structure in our thesis for the multinomial data.

1.7 Residuals

Since in regression analysis the random ε_i are unobservable, they are estimated by the least squares residuals, which are actually the difference between observed and estimated responses when the least squares method is used to fit the model and in our approach the *rth* residual for observation y_r is

$$r_r = y_r - \hat{y}_r, \qquad \text{for} \qquad r = 1,2,\cdots,n$$

$$= y_r - x_r^T \hat{\beta} \ . \tag{1.19}$$

In matrix notation, the residual vector is defined

$$r = y - X\hat{\beta} \ . \tag{1.20}$$

It is of historical interest to note that in 1809, in a text of astronomy, Gauss introduced the concept of the errors in a Normal distribution with zero mean and constant variance and in Theoria Combinations in 1823, he abandoned the Normal distribution replacing it by a weaker assumption of constant variance. The extension of this weaker assumption was given by Wedderburn (1974) to generalized linear models. For generalized linear models we can extend the equation (1.18) to all the distributions that may replace the Normal. In this section, we use the theoretical form, involving $\hat{\mu}$ rather then μ and define the generalized form of residual or Pearson residual as,

$$r_{pr} = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}} \ . \tag{1.21}$$

This is just a raw residual scaled by the estimated standard deviation of y. The Pearson residuals have been used in model diagnostic procedures. The GLIM provides Pearson residuals by default, although other approaches to residuals are becoming increasing widely used. We introduce some of these concepts for future use or in some examples for model checking.

1.7.1 Modified Pearson residuals (the residual of GLIM)

The simplest approach (output by default with GLIM) is the Modified Pearson residual r_r^P. For each unit r, we have:

$$r_r^P = \frac{y_r - \hat{\mu}_r}{\sqrt{\frac{k}{w_r} V(\hat{\mu}_r)}} . \tag{1.22}$$

where k is the scale parameter and the modification is the factor k/w_r, which ensure that the denominator is a reasonable estimate of $\text{var}(y_i)$. The $\hat{\mu}_r$ are the GLIM fitted values (stored as %fv). The r_r^P are available in GLIM as **%rs**.

1.7.2 Standardized residuals

The modified Pearson residual suffers from the obvious disadvantage that it does not take into account that the $\hat{\mu}_r$ are merely estimates of μ_r and hence are correlated with the responses y_r. The estimated variance should ideally take into account this correlation. It is therefore desirable to adjust the modified Pearson residual by dividing it by a factor $\sqrt{(1-h_r}$ which compensates for the correlation between y_r and μ_r. Thus we define the standardized Pearson residual as,

$$r_r^{PS} = \frac{r_r^P}{\sqrt{(1-h_r)}} = \frac{y_r - \hat{\mu}_r}{\sqrt{\frac{k}{w_r} V(\hat{\mu}_r)(1-h_r)}} . \tag{1.23}$$

The calculation of this requires knowledge of the h_r . These quantities are in fact the diagonal entries in the hat matrix. The h_r may be extracted in GLIM as **%lv**, which enables the user to 'Standardize' the GLIM residuals.

1.7.3 Studentized residuals

In residual plots, we are generally interested in the pattern rather then the size of the plots so it is usually not necessary to scale the residuals by an estimate of the unknown scale parameter ϕ. However following McCullagh and Nelder (1989) scaled residuals can be defined as Studentized residuals. For the standardized (modified) Pearson residuals, the Studentized version is:

$$r_r^{PS\prime} = \frac{r_r^{PS}}{\sqrt{\hat{\phi}}} = \frac{r_r^{P}}{\sqrt{\hat{\phi}(1-h_r}}$$

$$= \frac{y_r - \hat{\mu}_r}{\sqrt{\hat{\phi}\frac{k}{w_r}V(\hat{\mu}_r)(1-h_r)}} . \qquad (1.24)$$

1.7.4 Deviance residuals

The deviance plays a central role in inferential aspects of generalized linear modelling. A deviance residual r_r^D can be defined as

$$r_r^D = sign(y_r - \hat{\mu}_r)\sqrt{d_r} . \qquad (1.25)$$

where *sign* ( ) indicates that the deviance residual is taken as positive if $y_r - \hat{\mu}_r > 0$, and negative if $y_r - \hat{\mu}_r < 0$. Here the d_i are the deviance increments and the values of the d_r may be extracted in GLIM as **%di**.

1.7.5 Standardized, Studentized deviance residuals

As with Pearson residuals, it is better to standardized the deviance residuals defining by

$$r_r^{DS} = \frac{r_r^D}{\sqrt{(1 - h_r)}}$$

$$= \frac{sign\ (y_r - \hat{\mu}_r)\sqrt{d_r}}{\sqrt{(1 - h_r}} . \qquad (1.26)$$

It may again be convenient to divide these standardized residuals by an estimate of ϕ to give the Studentized version

$$r_r^{DS\,\prime} = \frac{r_r^D}{\sqrt{\hat{\phi}\,(1 - h_r)}} . \qquad (1.27)$$

Some other residuals are also suggested, see William (1987) and Atkinson (1985).

1.8 Case *i*

In this research we are dealing with the distribution to model a multi - level response variable y_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, J$, as are described in section (1.3) for m observed values (cases) of a J - level response variable. The cell (i, j) denotes the response variable level j of case i, *i.e.* case i has observations $y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ}$, and see Table 1.2. For data on individual set each case i is one individual so $y_{ij} = 0$ or 1, $\sum_{j-1}^{J} y_{ij} = 1$ where as for the aggregated data, each case i consists of n_i individuals, $0 \leq y_{ij} \leq n_i$, $\sum_{j-1}^{J} y_{ij} = n_i$.

Our method of fitting the multinomial models in this research naturally gives the Cook's distance and the Hat - matrix for the observation y_r of stacked data for the J level of response variable. We shall see in the following chapters that this has a little diagnostic value. Instead, the likelihood influence measure of Cook (1986) for a case i is calculated and the details are suggested in chapter 7.

1.9 Cook's distance (1977) as an influence measure

The leverage coefficient h_r given in section (1.6) alone cannot tell us if our linear predictor is being affected strongly by an observation y_r. Another measure in statistics is a Cook's distance (1977) and is commonly used to estimate the influence of the data points when using the standard regression analysis. In our approach for fitting a multinomial logit model for a univariate response variable, Cook's distance (1977) can be used as a measure of influence for an observation y_r and is given as

$$C_r = \frac{h_r}{p(1 - h_r)} (r_r^{PS\prime})^2 . \qquad (1.28)$$

The quantity C_r is a measures of the change in the vector of all n predicted values when observation y_r is not used in estimating β and can be extracted using the GLIM code **%cd**. We may use the alternate definition based upon the deviance residuals, namely

$$C_r^D = \frac{h_r}{p(1-h_r)} (r_r^{DS'})^2 . \qquad (1.29)$$

A Convenient search for influential cells in a traditional GLM is carried out by looking for larger values of Cook's distance (1977), but no clear rules can be given for what constitutes a large value of C_r. A simple index plot of the statistics against case number can be useful in determining the largest value. The observations y_r with large values of C_r can then be weighted out of the analysis and a test carried out of the change in the scaled deviance. The significant changes in the scaled deviance lead us to consider the status of the observation in question.

However, we will see that, when considering influence, it is not appropriate to consider merely the observation y_r, rather we conclude that we must consider the case i (consisting of observations $y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ}$) and as the Cook's distance (1977) approximate the likelihood influence measure of Cook (1986). But the likelihood involves $y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ}$ (or $y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ-1}, n_i$), so if we wish to find the likelihood measure we must delete the whole case i.

We will thus decide that we need a measure of the influence of the case i and Cook's distances (1977) for $y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ}$ might provide some information on the influence of the case i. However, with adequate computing power, likelihood diagnostics using single case i, deletions can be determined using the full likelihood influence measure of Cook (1986) and can be seen in chapter 7.

1.10 Influential observations

Influential observations are those observations that, individually or collectively, excessively influence the fitted regression equation as compared to other observations in the data set. It is therefore important to be able to locate such observations or in this research to locate case *i* and assess their impact on the model. More specifically, Belsley, Kuh, and Welsch (1980) define an influential observation "one which, either individually or together with several other observations, has a demonstrably large impact on the calculated values of various estimates - - - than is the case for most of the other observations''.

An observation, however, may not have the same influence on all the regression results. The question "Influence on what?" is, therefore, an important one. For example, an observation may have influence on $\hat{\beta}$, the estimated variance of $\hat{\beta}$, the predicted values, and/or the goodness - of - fit statistics. In this study we discuss how observation may influence the regression parameter $\hat{\beta}$.

We consider the available tools such as leverage, Cook's distance and how these might provide some information on influence points. In our approach for a multi - level response variable we need to investigate the influence effect of the case *i* instead of the influence of an observation y_{ij}.

1.11 Assessment of influence by deletion

John F. W Herschel (1830) mentioned in *A Preliminary Discourse On the Study of Natural Philosophy* " To arrive inductively at law of this kind, where one quantity depends on or varies with another, all that is required is a series of careful and exact measures in every different state of the datum and quaesitum. Here, however the mathematical form of the law being of the highest importance, the greatest attention must be given to the extreme cases as well as to all those points where the one quantity changes rapidly with a small change of the other''.

To detect the influential observations it is necessary to study what sort of effect they produce on the estimation of regression parameter $\hat{\beta}$. For that we need to change the usual structure of the analysis by deleting observations or by perturbing them.

1.12 Ideas for the deletion of individual observations/cases

Influential points may occur because of a variety of reasons and decisions how to deal with them must be made according to context. In our situation, we will need to distinguish between what we call 'observations' and what we call 'cases'. Some influential observations (in our situation cases) are not necessarily undesirable and can often, in fact, provide more important information than other data. Improperly recorded cases, whether caused by measurement errors during the original experiment or by simple transcription errors later in the analysis, should be corrected if possible, or otherwise be deleted from the data set. However, if, say, deletion of a case from the data set considerably changes the value of an estimated parameter, then relevant inference concerning that parameter may be in doubt.

The method and the theory presented in this research for fitting the multinomial models calculates a 'correct' Cook's distance (1977) and 'hat - matrix' for y_{ij}. However, the use of deletion statistics for a single observation y_{ij} will not provide total information on the effect of deleting a case $(y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ})$.The individual Cook's distances for $y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iJ}$, etc might be used to provide some indication, however, of the potential likelihood displacement.

The theory presented in the next section gives the method of finding the likelihood influence measure of Cook (1986) using the deletion of a single case. The method for finding the likelihood influence measure in our fitting of multinomial model is delete (or in other word to weight out) the case *i*. The details of deleting the case *i* can be found in macros and the framework of data entry as in example 1 of section (7.3.1) for deleting the case $i=1$, (which consists of two 'observations', as we have a trinomial response). We used the equation (7.1) to find the likelihood measure for the full data then we deleted out the case $i=1$, which is in our setup is 1st and 5th observation in the macro given and

finally the equation (1.29) gives us the likelihood influence measure of Cook (1986). We will repeat this process for each $i = 1, 2, \ldots, 5$ and the Table 7.3 is formed for these influence measures of the data in Table 7.1

1.13 Likelihood influence measure of Cook (1986)

The likelihood Cook's distance (1986) idea is based on the likelihood approach for the hypothesised model and for the perturbed model (deletion of an observation y_{ij}) , but in multinomial data the perturbation is a case i, $i = 1, 2, \ldots, m$. The log - likelihood of the perturbed model with deletion and the unperturbed model without deletion are used to measure the influence of the perturbation or deletion.

For clarity we consider here a univariate response variable (rather then multivariate multinomial case) and suppose we have the response values $y_1, y_2, \cdots, y_m$ with each y_i having the probability density function $f_i(y_i; \beta)$, $i = 1, 2, \cdots, m$ where β is a $p \times 1$ vector of unknown parameters. If $l_i(\beta) = \log f_i(y_i; \beta)$ then the log - likelihood is given by

$$l(\beta) = \sum_{i=1}^{m} l_i(\beta) . \qquad (1.30)$$

We now introduce the perturbation of model via deleting the *ith* observation and denoting log - likelihood corresponding to the perturbed model by $l(\beta_{(i)})$ and assume there exists a null perturbation with log - likelihood $l(\beta)$. The maximum likelihood estimators of β and $\beta_{(i)}$ found from $l(\beta)$ and $l(\beta_{(i)})$ are respectively $\hat{\beta}$ and $\hat{\beta}_{(i)}$.

The likelihood influence measure

$$LD(\hat{\beta}_{(i)}) = 2[l(\hat{\beta}) - l(\hat{\beta}_{(i)})] , \qquad (1.31)$$

is simply twice the difference of the maximised log - likelihood of the perturbed model and null perturbed model and is always nonnegative. It can be used to assess the influence of the perturbation (deletion of *ith* observation).

The likelihood influence measure $LD(\hat{\beta}_{(i)})$ may also be interpreted in terms of an asymptotic confidence region in a similar way to the interpretation of Cook's distance (1977). One can consider the general asymptotic confidence region result

$$\{\beta : 2[l(\hat{\beta}) - l(\beta)] \leq \chi^2(\alpha ; p)\}$$

where $\chi^2_{(\alpha;p)}$ is the upper α point of the chi - squared distribution with p degrees of freedom, and p is the dimension of β . Then

$$LD(\hat{\beta}_{(i)}) = 2[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$$

can therefore be calibrated by comparison to the $\chi^2_{(p)}$ distribution. The effect of removing the *ith* case of the data on $\hat{\beta}$ is to consider (now $\hat{\beta}_{(i)}$) moving to the boundary of the $100(1-\alpha_i)\%$ confidence ellipsoidal region where α_i is such that

$$2[l(\hat{\beta}) - l(\hat{\beta}_{(i)})] = \chi^2(\alpha_i ; p)\} \ . \qquad (1.32)$$

Thus a very small level of significance α_i implies very great influence since $\hat{\beta}_{(i)}$ is moved a long way from $\hat{\beta}$.

The so - called Cook's distance given in equation (1.28) is a one step approximation to a likelihood influence measure $LD(\hat{\beta}_{(i)})$, for the case of a univariate response variable. However, consideration of a multivariate response variable (e.g. the multinomial with J - level case deletion) involves the deletion of $J-1$ observations. In this case, the simple single observation Cook's distance (1977) is not an approximate influence measure.

CHAPTER 2

Fitting a Multinomial Logit Model

2.1 Introduction

In this chapter we will give a formulation and the derivation of the iterative re - weighted least squares (IRLS) algorithm for fitting a multinomial logit model given in equation (1.5) by using Cholesky's decomposition approach applied to a matrix which has diagonal sub - matrices. We initially consider a special case for a three level of response variable with one explanatory variable and the further extensions on these results are given in the following chapters. For clarity and simplification we denote $J = 3$ level of response variable as follows,

Response variables			
Y_1	Y_2	Y_3	Totals
y_{11}	y_{12}	y_{13}	n_1
y_{21}	y_{22}	y_{23}	n_2
.....			.....
.....			.....
y_{m1}	y_{m2}	y_{m3}	n_m

Table 2.1: 3 - Level of response variable

In Table 2.1, for example, y_{i2} is the *ith* observation for response level 2 and n_i is a total of the various observations for case i with the assumption that the n_i, $i = 1, 2, \ldots, m$ are fixed. We assume that y_{i1}, y_{i2}, y_{i3} follow a multinomial distribution for each i. We also assume that corresponding to each y_{ij} there is a probability p_{ij} that will be estimated and the sum of probabilities over J is equal to 1 *i.e.* $\sum_{j=1}^{3} p_{ij} = 1$ for all $i = 1, 2, \ldots, m$. Thus in this case there are $3 - 1 = 2$ distinct probabilities to be estimated for each i; more detail on probabilities can be found in section (1.3). The section (2.2) formulates the basis for parameter estimation using the IRLS algorithm with the derivation in section (2.3). The section (2.4) gives a brief description for fitting more than one explanatory variable and section (2.5) is an application of these results using two different types of data. The results found in section (2.5) are compared with estimates obtained using the Poisson trick approach of Francis et al. (1992). The remarks on some difficulties that arises in the application to a particular type of data are given in section (2.6) with a summary in section (2.7).

2.2 Estimation of parameters using IRLS

As explained in the GLIM manual, the iterative re - weighted least squares (IRLS) algorithm can be used to fit more general regression models than just within the framework of generalized linear models. In this section, we will give a derivation of the IRLS algorithm for maximum likelihood estimation in fitting the multinomial logit model, a more general class. The theory here provides an explanation for the success of the method as we have a probability model for the distribution of responses and the explanatory variables, they can be decomposed into a random component and a systematic component as follows,

The random component: The log - likelihood function l for y depends on the explanatory variable x and the unknown parameters β only through the values of a finite dimensional vector of predictors η. Thus l is a specified function of y and η.

The systematic component: The predictor vector η is a prescribed, deterministic, function of the explanatory variables x and the unknown parameters β.

When modelling the data as a multinomial model, we have in general the freedom to choose the predictors. We write

$$\eta = X\beta . \tag{2.1}$$

In our approach here for a three level of response variable in equation (1.11),

$$\eta = \begin{pmatrix} \eta^{(2)} \\ \eta^{(3)} \end{pmatrix} = \begin{pmatrix} X^{(2)}\beta^{(2)} \\ X^{(3)}\beta^{(3)} \end{pmatrix}. \tag{2.2}$$

As the data y (response variable) and x are observed, the model - fitting by maximum likelihood consists of maximising a composite log - likelihood function $l(\eta(\beta))$ (we are suppressing **y** and **x** from the notation for clarity), and it is a numerical calculation that IRLS algorithm is well - adapted to.

It is required to solve the maximum - likelihood equations

$$\frac{\partial l}{\partial \beta} = 0 . \tag{2.3}$$

But in this framework,

$$\frac{\partial l}{\partial \beta} = \boldsymbol{D}^T \frac{\partial l}{\partial \eta}, \tag{2.4}$$

where $\boldsymbol{D}$ is the matrix of derivatives with

$$D_{ij} = \frac{\partial \eta_i}{\partial \beta_j} . \tag{2.5}$$

These equations cannot be solved explicitly, so iteration is needed. The most familiar approach is the Newton - Raphson algorithm:

$$\beta_{new} = \beta + \boldsymbol{H}^{-1} \frac{\partial l}{\partial \beta}, \tag{2.6}$$

where $$\boldsymbol{H} = \left[-\frac{\partial^2 l}{\partial \beta_i \partial \beta_j} \right] = \boldsymbol{D}^T \left[-\frac{\partial^2 l}{\partial \eta_i \partial \eta_j} \right] \boldsymbol{D} + \frac{\partial l}{\partial \eta} \left[-\frac{\partial^2 \eta}{\partial \beta_i \partial \beta_j} \right]. \tag{2.7}$$

Kendall and Stuart (1967) suggest replacing $\boldsymbol{H}$ by $\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D}$,

where $$W_{ij} = E\left(-\frac{\partial^2 l}{\partial \eta_i \partial \eta_j}\right). \tag{2.8}$$

Therefore $$\beta_{new} = \beta + (\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D})^{-1} \frac{\partial l}{\partial \beta}$$

$$= \beta + (\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D})^{-1} \boldsymbol{D}^T \frac{\partial l}{\partial \eta}. \tag{2.9}$$

We can rewrite

$$\beta_{new} = (\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D})^{-1} \boldsymbol{D}^T \boldsymbol{W} \left(\boldsymbol{D}\beta + \boldsymbol{W}^{-1} \frac{\partial l}{\partial \eta}\right), \tag{2.10}$$

or equivalently

$$(\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D}) \beta_{new} = \boldsymbol{D}^T \boldsymbol{W} \left(\boldsymbol{D}\beta + \boldsymbol{W}^{-1} \frac{\partial l}{\partial \eta}\right). \tag{2.11}$$

2.3 Parameter estimates β_{new} for $J=3$

In this section we will illustrate the procedure and the algorithm to find the parameters β_{new} by considering a three level of response variable in equation (2.11) using a standard least squares method. In equation (2.11) we need to find the matrices $\boldsymbol{D}$, W and $\frac{\partial l}{\partial \eta}$, where $\boldsymbol{D}$ is a matrix of derivatives and from equations (2.5);

$$\boldsymbol{D} = \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix}_{2m \times 4}. \tag{2.12}$$

Here $X^{(k)}$ is a matrix of explanatory variables for the response variable k. It is here assumed, although it is not necessary, that $X^{(2)} = X^{(3)}$.

The matrix W can be written $(W_{ij}) = E(-\frac{\partial^2 l}{\partial \eta_i \partial \eta_j}) = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}_{2m \times 2m}$, (2.13)

where using the Cholesky decomposition of symmetric positive definite matrices, we can rewrite and more details can be found in Appendix B

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

$$= \begin{pmatrix} A_{11}^2 & A_{11}A_{12} \\ A_{21}A_{11} & A_{21}A_{12} + A_{22}^2 \end{pmatrix} = \begin{pmatrix} A_{11}^2 & A_{11}A_{12} \\ A_{21}A_{11} & A_{12}^2 + A_{22}^2 \end{pmatrix}. \tag{2.14}$$

The W_{ij} and A_{ij} are diagonal matrices and we can write,

$$A_{11}^2 = W_{11} \Rightarrow A_{11} = sqrt\ (W_{11}),$$

$$A_{11}A_{12} = W_{12} \Rightarrow A_{12} = W_{12}A_{11}^{-1}, \tag{2.15}$$

$$A_{12}^2 + A_{22}^2 = W_{22} \Rightarrow A_{22} = sqrt(W_{22} - A_{12}^2).$$

The notation ***sqrt(W)*** for some diagonal matrix $W = (W_{ij})$, here means the diagonal matrix with diagonal entries $\sqrt{W_{ii}}$.

The equation (1.8) for three level of response is

$$l_i = \log L_i = y_{i2}\theta_{i2} + y_{i3}\theta_{i3} - n_i \log(1 + e^{\theta_{i2}} + e^{\theta_{i3}}) .$$

Or equivalently

$$l_i = \log L_i = y_{i2}\eta_i^{(2)} + y_{i3}\eta_i^{(3)} - n_i \log(1 + e^{\theta_{i2}} + e^{\theta_{i3}}) , \qquad (2.16)$$

where $\eta_i^{(j)}$ is a logit link for the observation for case i and response level j.

Now to find W_{ij} in equation (2.14), we need the following derivatives in equation (2.16),

$$\frac{\partial l_i}{\partial \eta_i^{(2)}} = y_{i2} - \frac{n_i e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}$$

and

$$\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_i^{(2)}} = -\left\{ \frac{n_i e^{\theta_{i2}}(1 + e^{\theta_{i2}} + e^{\theta_{i3}}) - n_i e^{\theta_{i2}} . e^{\theta_{i2}}}{(1 + e^{\theta_{i2}} + e^{\theta_{i3}})^2} \right\}$$

$$= -n_i \left\{ \frac{e^{\theta_{i2}}}{(1 + e^{\theta_{i2}} + e^{\theta_{i3}})} - [\frac{e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}]^2 \right\}$$

$$= -n_i e^{\theta_{i2}} p_{i1} + n_i (e^{\theta_{i2}} . p_{i1})^2$$

$$= -n_i p_{i2} + n_i p_{i2}^2 = -n_i p_{i2}(1 - p_{i2}) . \qquad (2.17)$$

Thus

$$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_i^{(2)}}) = n_i p_{i2}(1 - p_{i2})$$

$$= \mu_{i2}(1 - \frac{\mu_{i2}}{n_i}) . \qquad (2.18)$$

Similarly $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(3)} \partial \eta_i^{(3)}}) = \mu_{i3}(1 - \frac{\mu_{i3}}{n_i}).$$

Or for $i \neq j$ in equation (2.16) we have,

$$\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_j^{(2)}} = \frac{\partial}{\partial \eta_j^{(2)}}\{y_{i2} - \frac{n_i e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}\}$$

$$= 0 - \frac{\partial}{\partial \eta_j^{(2)}}\{\frac{n_i e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}\} = 0.$$

Thus we have $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_j^{(2)}}) = 0.$$

Similarly $$\frac{\partial l_i}{\partial \eta_i^{(3)} \partial \eta_j^{(3)}} = \frac{\partial}{\partial \eta_j^{(3)}}\{y_{i3} - \frac{n_i e^{\theta_{i3}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}\} = 0$$

and $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(3)} \partial \eta_j^{(3)}}) = 0.$$

Now to find the 2nd derivative of l_i with respect to different η_i in equation (2.16),

$$\frac{\partial l_i}{\partial \eta_i^{(3)} \partial \eta_i^{(2)}} = \frac{\partial}{\partial \eta_i^{(3)}}\{y_{i2} - \frac{n_i e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}\}$$

$$= -n_i e^{\theta_{i2}}\{-e^{\theta_{i3}}(1 + e^{\theta_{i2}} + e^{\theta_{i3}})^{-2}\}$$

$$= n_i e^{\theta_{i2}} e^{\theta_{i3}} p_{i1}^2 = n_i p_{i1} e^{\theta_{i2}} p_{i1} e^{\theta_{i3}}.$$

Hence we have $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(3)} \partial \eta_i^{(2)}}) = -\mu_{i2}\frac{\mu_{i3}}{n_i} = -\frac{\mu_{i2}\mu_{i3}}{n_i}. \qquad (2.19)$$

We can write these derivatives for clarity if $i = 1$ and $j = 2$,

$$\frac{\partial l_1}{\partial \eta_1^{(2)}} = y_{i2} - \frac{n_1 e^{\theta_{12}}}{1 + e^{\theta_{12}} + e^{\theta_{13}}} = y_{i2} - \mu_{i2},$$

and

$$\frac{\partial^2 l_1}{\partial \eta_1^{(2)} \partial \eta_1^{(2)}} = -\left\{ \frac{n_1 e^{\theta_{12}} (1 + e^{\theta_{12}} + e^{\theta_{13}}) - n_{1i} e^{\theta_{12}} . e^{\theta_{13}}}{(1 + e^{\theta_{12}} + e^{\theta_{13}})^2} \right\},$$

so

$$E\left(-\frac{\partial^2 l_1}{\partial \eta_1^{(3)} \partial \eta_1^{(3)}}\right) = \mu_{12}\left(1 - \frac{\mu_{12}}{n_1}\right).$$

Also

$$\frac{\partial l_1}{\partial \eta_1^{(2)} \partial \eta_2^{(2)}} = \frac{\partial}{\partial \eta_2^{(2)}}\left\{ y_{12} - \frac{n_1 e^{\theta_{12}}}{1 + e^{\theta_{12}} + e^{\theta_{13}}} \right\} = 0,$$

and

$$E\left(-\frac{\partial^2 l_1}{\partial \eta_1^{(2)} \partial \eta_2^{(2)}}\right) = 0.$$

Also

$$\frac{\partial l_1}{\partial \eta_1^{(3)} \partial \eta_1^{(2)}} = \frac{\partial}{\partial \eta_1^{(3)}}\left\{ y_{12} - \frac{n_1 e^{\theta_{12}}}{1 + e^{\theta_{12}} + e^{\theta_{13}}} \right\},$$

and

$$E\left(-\frac{\partial^2 l_i}{\partial \eta_1^{(3)} \partial \eta_2^{(2)}}\right) = -\frac{\mu_{12}\,\mu_{13}}{n_1}.$$

Similarly we can find derivatives for different values of i and j.

If for clarity we define $W_{jk}(i)$ as the *ith* diagonal element of the diagonal matrix W_{jk}, (see equation 2.18) then, we have

$$W_{11}(i) = \mu_{i2}(1 - \frac{\mu_{i2}}{n_i})$$

$$W_{22}(i) = \mu_{i3}(1 - \frac{\mu_{i3}}{n_i}) \qquad (2.20)$$

$$W_{12}(i) = -\frac{\mu_{i2}\mu_{i3}}{n_i}.$$

Since W is a symmetric matrix, $W_{12}(i) = W_{21}(i)$. (2.21)

In GLIM code we can define these equations as follows; further details are given in the Appendix A:

$$W_{11}(i) = \%fv(i) * (1 - \%fv(i)/n(i)),$$

$$W_{22}(i) = \%fv(n+i) * (1 - \%fv(n+i)/n(i)),$$

$$W_{12}(i) = -\%fv(i) * \%fv(n+i)/n(i).$$

For the observations $i = m$ and $J = 3$ level of response variable, the matrix W is summarised in Appendix B and the matrix D is given in equation (2.12).

We here define the matrix N as

$$\left(D^T W D\right) = \left(D^T \begin{pmatrix} A_{11} & A_{21} \\ 0 & A_{22} \end{pmatrix}^T \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} D \right) = N^T N,$$

where $$N = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} D.$$

Then the equation (2.11) further can then be rewritten as,

$$(N^T N)\beta_{new} = D^T \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \left(D\beta + W^{-1} \frac{\partial l}{\partial \eta} \right) \tag{2.22}$$

$$= D^T \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} D\beta + \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} W^{-1} \frac{\partial l}{\partial \eta} \right)$$

$$= N^T \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} D\beta + \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} W^{-1} \frac{\partial l}{\partial \eta} \right)$$

$$= N^T \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} W^{-1} \frac{\partial l}{\partial \eta} \right)$$

$$= N^T \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix} \frac{\partial l}{\partial \eta} \right), \tag{2.23}$$

where $$\begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} W^{-1}$$

$$= \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \left(\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \right)^{-1}$$

or $$\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix} = I_{2n} \quad . \tag{2.24}$$

From equation (2.24) we have the equivalences,

$$A_{11}B_{11} = I \qquad \Rightarrow \qquad B_{11} = A_{11}^{-1} .$$

$$A_{22}B_{22} = I \qquad \Rightarrow \qquad B_{22} = A_{22}^{-1} .$$

$$A_{21}B_{11} + A_{22}B_{21} = 0 \qquad \Rightarrow \qquad B_{21} = -(A_{21}B_{11})A_{22}^{-1}$$

$$= -(A_{21}A_{11}^{-1})A_{22}^{-1} . \qquad (2.25)$$

We are only left now to find $\frac{\partial l}{\partial \eta}$ in equation (2. 23) to find β_{new}, but in our case in equation (2.2)

$$\eta = \begin{pmatrix} \eta_i^{(2)} \\ \eta_i^{(3)} \end{pmatrix} \quad \text{for} \quad i = 1, 2, \ldots, m .$$

We are assuming canonical links, and we have,

$$\frac{\partial l_i}{\partial \eta_i^{(2)}} = = y_{i2} - n_i p_{i1} e^{\theta_{i2}} = y_{i2} - \mu_{i2} .$$

Thus $\frac{\partial l}{\partial \eta^{(2)}} = Y_2 - \mu_2$, where $Y_j^T = (y_{1j}\ y_{2j} \ldots y_{nj})^T$ for $j = 2,3$.

Similarly $\frac{\partial l}{\partial \eta^{(3)}} = Y_3 - \mu_3 .$

Thus we can write $\frac{\partial l}{\partial \eta} = \begin{pmatrix} Y_2 - \mu_2 \\ \\ Y_3 - \mu_3 \end{pmatrix}.$ (2.26)

The final form of equation (2.23) becomes,

$$N^T N\beta_{new} = N^T\left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix}\begin{pmatrix} \beta \\ \gamma \end{pmatrix}+\begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} Y_2 - \mu_2 \\ Y_3 - \mu_3 \end{pmatrix}\right)$$

$$= N^T\left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} X^{(2)}\beta \\ X^{(3)}\gamma \end{pmatrix}+\begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} Y_2 - \mu_2 \\ Y_3 - \mu_3 \end{pmatrix}\right) \qquad (2.27)$$

$$= N^T\left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} \eta^{(2)} \\ \eta^{(3)} \end{pmatrix}+\begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} Y_2 - \mu_2 \\ Y_3 - \mu_3 \end{pmatrix}\right)$$

$$= N^T\left(\begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} \\ A_{22}\eta^{(3)} \end{pmatrix}+\begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2)+B_{22}(Y_3 - \mu_3) \end{pmatrix}\right). \qquad (2.28)$$

But in GLIM code we have $(\%\, lp) = \begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} \\ A_{22}\eta^{(3)} \end{pmatrix}$

So, equivalently, equation (2.28) becomes

$$N^T N\beta_{new} = N^T\left((\%lp)+\begin{pmatrix} B_{11}(Y_2 - \mu_2 \\ B_{21}(Y_2 - \mu_2)+B_{22}(Y_3 - \mu_3) \end{pmatrix}\right). \qquad (2.29)$$

The equation (2.29) is of the form which gives a least square solution and can be rewritten in the simplest form as follows,

$$\beta_{new} = (N^T N)^{-1} N^T Y. \qquad (2.30)$$

In equation (2.30) we have the new design matrix

$$N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} \\ 0 & A_{22}X^{(3)} \end{pmatrix}_{2m\times 2m}, \tag{2.31}$$

and the **y** - variable

$$\begin{pmatrix} (\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \end{pmatrix} \end{pmatrix}_{2m\times 1}. \tag{2.32}$$

Equivalently, the **y** - variable is given by

$$\begin{pmatrix} (\%lp) + \begin{pmatrix} A_{11}^{-1}(Y_2 - \mu_2) \\ \\ -(A_{12}A_{22}^{-1})A_{11}^{-1}(Y_2 - \mu_2) + A_{22}^{-1}(Y_3 - \mu_3) \end{pmatrix} \end{pmatrix}, \tag{2.33}$$

where $$(\%lp) = \begin{pmatrix} lp_2 \\ \\ lp_3 \end{pmatrix} = \begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} \\ \\ A_{22}\eta^{(3)} \end{pmatrix}. \tag{2.34}$$

Then, equivalently $\eta^{(3)} = A_{22}^{-1}lp_3$,

$$\eta^{(2)} = A_{11}^{-1}(lp_2 - A_{12}\eta^{(3)}). \tag{2.35}$$

The procedure defined here is very simple to find the matrices D, W and N using GLIM software. The y - variable in equation (2.32) depends on $\%lp$ and can be calculated from equation (2.34). The GLIM macros to find these matrices for different data used in this study are given in Appendix A.

The equation (2.30) is basically a standard least squares equation and can be used to fit any multinomial model. The fitting is demonstrated in section (2.5) with the Poisson trick approach of Francis, et al. (1992).

The hat - matrix and Cook's distance or some other statistics are easy to extract from our approach to search for the influential observations or cases (although some extra care is needed to interpret these statistics for multinomial data).

2.4 Cholesky Decomposition

The Cholesky decomposition of the symmetric positive definite weight matrix W in equation (2.11) can be formulated using some other available methods in literature instead of standard Andre - Louis Cholesky decomposition equation (2.14). A simpler reasonable approach can be square - root - free Cholesky decomposition of the variance - covariance matrix of the multinomial distribution given by Tanabe and Sagae (1992). It requires much fewer arithmetic operations than does the general Cholesky algorithm and is not affected by an ill - conditioned matrix. This approach is useful when elements of the probability vector are different orders of magnitude. An explicit formula of the Moore - Penrose inverse for the general ill - condition variance - covariance matrix is also given for the multinomial distribution.

The Cholesky decomposition general equation (2.14) used in our approach requires only diagonal sub - matrices and makes the required matrix inverse much simpler. Our multinomial model is basically a some - sort of least squares method that involves at each step these diagonal sub - matrices to find the equation (2.30) for estimating β_{new}. We are using GLIM algorithms and it makes much easier to manipulate these matrices for

fitting the multinomial model. The square - root - free Cholesky decomposition of the variance - covariance matrix given by Tanabe and Sagae (1992) is much simpler than general decomposition equation (2.14) but it will not give our least square equation ($\beta_{new} = (N^T N)^{-1} N^T Y$) for fitting the model.

2.5 Alternative approach to find W_{ij}

We can also use the standard result of Fisher Scoring to find the variance - covariance matrix $E(-\frac{\partial^2 l}{\partial \eta_i \partial \eta_j})$ equation (2.14) and the log - likelihood equation (2.16) as

$$l_i = \sum_{j=2}^{3} y_{ij}\theta_{ij} - n_i \log(1 + \sum_{j=2}^{3} e^{\theta_{ij}}), \quad i = 1, 2, \ldots, m. \tag{2.36}$$

From equation (2.36) we get

- $$\frac{\partial l_i}{\partial \theta_{ij}} = y_{ij} - \frac{n_i e^{\theta_{ij}}}{1 + \sum_{j=2}^{3} e^{\theta_{ij}}} = y_{ij} - n_i p_{ij}$$

- $$\frac{\partial l_i}{\partial \theta_{ik}} = y_{ik} - n_i p_{ik}.$$

We are using here the logit link or the canonical link $\theta_{ij} = \eta_i^{(j)} = \sum_{k=1}^{K} x_{ik}^{(j)} \beta_k^{(j)}$ and more details can be found in section (1.4). The variance - covariance's diagonal matrices can be obtained as follows

- $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(j)} \partial \eta_i^{(j)}}) = E(\frac{\partial l_i}{\partial \theta_{ij}} \times \frac{\partial l_i}{\partial \theta_{ij}}) = E(y_{ij} - n_i p_{ij})^2$$

$$= Var\ (\ y_{ij}\ ) = n_i\ p_{ij}\ (1 - p_{ij}\ ) = \mu_{ij}\ (1 - \frac{\mu_{ij}}{n_i})$$

- $E(-\frac{\partial^2 l_i}{\partial \eta_i^{(j)} \partial \eta_i^{(k)}}) = E\,(\frac{\partial l_i}{\partial \theta_{ij}} \times \frac{\partial l_i}{\partial \theta_{ij}}) = E\,\{(y_{ij} - n_i\ p_{ij}\ )(\ y_{ik} - n_i\ p_{ik}\ )\}.$

$$= Cov\,(\ y_{ij}\ ,\ y_{ik}\ ) = -\ n_i\ p_{ij}\ p_{ik} = -\ \frac{\mu_{ij}\ \mu_{ik}}{n_i}$$

The above results for $J > 3$ level of response variable can be extended easily and more details other than the logit link function or the Box - Cox link function are given in chapter 4 with full derivation in Appendix D.

2.6 More than one explanatory variable

We consider here in this section a case for a $J = 3$ level of response variable with two explanatory variables x_1 and x_2. The x_1 and x_2 in general can be two different explanatory variables for each level of response variable or in general for some data in Table (1.2) for 'm' can be different for each level of a response variable. The theory and the results given in the previous section are still applicable.

We consider here for illustration the artificial data in Table 1.1, with Plans to enter HE as a 3-level of response variable Y with levels 'Definitely Yes', 'Possible', and 'Definitely No'. We have Age$=x_1$ and Gender $= x_2$ as two explanatory variables. The Age$=x_1$ and Gender $=x_2$ are considered as two explanatory variables for each response level in sections (2.6.1). This data are also considered for further analysis in section (2.7.2) using the Poisson trick approach in section (2.6.2) and the theory is demonstrated when there are different explanatory variables for each response level.

2.6.1 Estimates β_{new} for two explanatory variables

Let Age$=x_1$ and Gender$=x_2$ be two continuous explanatory variables for the 3 - level variable 'Plans to enter HE' as response variable Y. Then assuming $X^{(2)}=X^{(3)}=X$, the link function η and the matrix D in the previous section (2.3) become

$$\eta=\begin{pmatrix}\eta^{(2)}\\ \eta^{(3)}\end{pmatrix}=\begin{pmatrix}X & \beta^{(2)}\\ X & \beta^{(3)}\end{pmatrix}, \quad \text{where} \quad X=\begin{pmatrix}1 & x_1 & x_2\end{pmatrix}_{m\times 3}, \tag{2.36}$$

$$D_{ij}=\frac{\partial \eta_i}{\partial \beta_j}=\begin{pmatrix}1 & x_1 & x_2 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & x_1 & x_2\end{pmatrix}_{2m\times 6}. \tag{2.37}$$

Now to recall equation (2.29):

$$N^T N\beta_{new}=N^T\left((\%lp)+\begin{pmatrix}B_{11}(Y_2-\mu_2)\\ B_{21}(Y_2-\mu_2)+B_{22}(Y_3-\mu_3)\end{pmatrix}\right), \tag{2.38}$$

with a new design matrix

$$N=\begin{pmatrix}A_{11} & A_{11}x_1 & A_{11}x_2 & A_{12} & A_{12}x_1 & A_{12}x_2\\ 0 & 0 & 0 & A_{22} & A_{22}x_1 & A_{22}x_2\end{pmatrix}_{2m\times 6},$$

and the **y** - variable is as in equation (2.32),

$$\left((\%lp)+\begin{pmatrix}B_{11}(Y_2-\mu_2)\\ B_{21}(Y_2-\mu_2)+B_{22}(Y_3-\mu_3)\end{pmatrix}\right).$$

2.6.2 Estimates β_{new} for different explanatory variables

For a general case if we consider in section (2.4.1) that $X^{(2)}$ and $X^{(3)}$ are two different explanatory variables then we have

$$\eta = \begin{pmatrix} X^{(2)}\beta^{(2)} \\ X^{(3)}\beta^{(3)} \end{pmatrix} \qquad \text{where} \qquad X^{(2)} = \begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix}_{m\times 3} \quad ,$$

$$X^{(3)} = \begin{pmatrix} 1 & x_1^{'} & x_2^{''} \end{pmatrix}_{m\times 3} \quad \text{and} \qquad \boldsymbol{D}_{ij} = \begin{pmatrix} 1 & x_1 & x_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_1^{'} & x_2^{''} \end{pmatrix}_{2m\times 6}$$

with a design matrix

$$N = \begin{pmatrix} A_{11} & A_{11}x_1 & A_{11}x_2 & A_{12} & A_{12}x_1^{'} & A_{12}x_2^{''} \\ 0 & 0 & 0 & A_{22} & A_{22}x_1^{'} & A_{22}x_2^{''} \end{pmatrix}_{2m\times 6} .$$

The **y** - variable will stay the same as in section (2.6.1).

It is not that difficult from section (2.6.1) or (2.6.2) to fit a least square model of $\boldsymbol{Y}$ versus $\boldsymbol{N}$. Effectively, we fit a GLIM model of main effects without the constant term. This will then be equivalent to a Poisson trick approach of Francis et al. (1992) for two explanatory variables x_1 and x_2, The above approach is more flexible to fit any required model. (For example, it is not easy to fit the same model via the Poisson trick for the case where x_1 , x_2 are continuous).

2.7 Examples

In this section we will use two different sets of data from a social survey to apply the theory and the methods developed in this chapter for fitting a multinomial logit link model. The statistics found in our approach for fitting a multinomial logit link model are compared with those obtained by the Poisson trick approach of Francis et al. (1992) and some remarks are given for fitting the multinomial logit link model when lots of 0's and 1's exists for the individual response level data

2.7.1: Data set for $J = 3$ response variable

As an application of our method for fitting a multinomial logit model, we will use as an illustrative a set of data from the US 1984 General Social Survey, as provided by Green, M., of the Centre for Applied Statistics, Lancaster University, U.K. 1473 respondents were asked a question (Health) regarding their state of health: the response categories were 1=excellent, 2=good, 3=fair and 4=poor. To illustrate our methodology, we ignore the ordered nature of the response, and fit a multinomial logit model to the data using age of the respondents as an explanatory variable. In this first illustrative example we will consider only three response categories 1=excellent, 2=good and $3 \geq$ fair; age of the respondents is the explanatory variable. We will continue this data set in the following chapters and the data are attached in a compact disk at the last page. More information about the data can be obtained from the compact disk.

We have a three level of response variable and we consider 'age' as a continuous explanatory variable. The data (responses) are arranged in three columns of observations. Here m=1456, since 17 observations are missing. The fourth column is the explanatory variable age and fifth column is the total of responses, y_{i1}, y_{i2}, y_{i3}. The response 1=excellent will, in effect, not be needed in the analysis, as we know that these values are given in the total. In the notation of GLIM, the 'standard length' of our data for our approach defined will be $1456+1456=2912$, as we stack the data and are shown previously in section (2.3).

The design matrix in fitting the multinomial logit model is,

$$N = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix} = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} \\ 0 & A_{22}X^{(3)} \end{pmatrix}_{2912\times 4}.$$

Here A_{11}, A_{12}, A_{22} are diagonal matrices of dimension 1456×1456 and in our notation $X^{(2)} = X^{(3)}$ are matrices of dimension 1456×2, where the 1st columns of $X^{(2)}$ and $X^{(3)}$ matrices are all 1's and the 2nd columns are the explanatory variable, age.

The **y** - variable is given as equation (2.32)

$$\begin{pmatrix} (\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \end{pmatrix} \end{pmatrix}_{2912\times 1}.$$

The multinomial logit model with main effects is fitted in GLIM using macros given inn the Appendix A with starting value defined in the macro startup. A reliable appropriate starting value in the macro startup is needed to obtain the correct degrees of freedom in the model fitting; otherwise the multinomial logit model fitting gives the incorrect degrees of freedom.

We can use the GLIM code $use startup $use loop$ repeatedly in our macro until convergence occurs. In this macro the model with main effects fitted using $fit *d1 + d2 + dp1 + dp2 – 1* $ (*i.e.* without the constant term) and the fitted values are equivalent to those obtained using the Poisson trick approach of Francis et al. (1992). We repeat $use loop$ once more after convergence in order to obtain the correct standard errors.

We have in the macro

d1	*dp1*	*d2*	*dp2*
$\begin{pmatrix} A_{11} \\ 0 \end{pmatrix}_{2912\times 1}$	$\begin{pmatrix} A_{11}Age \\ 0 \end{pmatrix}$	$\begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix}$	$\begin{pmatrix} A_{12}Age \\ A_{22}Age \end{pmatrix}$

and after fitting the multinomial logit model with main effects without the constant term, we get the these statistics:

scaled deviance = 2957.5 at cycle 9

residual df = 2908

	estimate	s.e.	parameter
1	0.1826	0.1630	*d1*
2	-2.056	0.2147	*d2*
3	0.006864	0.003675	*dp1*
4	0.03814	0.004299	*dp2*

Table 2.2: Parameter estimates and standard errors in logit model

scale parameter 1.000

The above statistics in Table 2.2 are exactly the same as are obtained using the Poisson trick approach of Francis et al. (1992) as shown in Table 2.3 (with some minor rounding errors in standard errors).

The fitted model is

$$\eta = 0.1826d1 - 2.056d2 + 0.006864dp1 + 0.03814dp2, \qquad (2.39)$$

and

$$\eta_1 = 0.1826d1 + 0.006864dp,$$

$$\eta_2 = -2.056d2 + 0.03814dp2.$$

We can interpret from above that a unit increase in $d1$ gives an increase in η_1 and similarly a unit decrease in $d2$ gives an increase in η_2 respectively. We can illustrate that the fitted multinomial logit model with main effects gives a good fit. In fact, we will be able to check that there is no clear indication of any serious violation of basic assumptions.

We can easily extract from the fit, Leverages, Cooks distance, fitted values and Pearson residuals and the interpretations of these statistics may be quite difficult.

2.7.2: Poisson trick approach of Francis et al. (1992)

The Poisson trick approach of Francis et al. (1992) for the above data set is given as follows for the comparison and verification of the results found in our multinomial logit model approach. The macro is given in Appendix A and more information about the macro can be found on the compact disk.

```
$fit +group*age(case)$di e $                                    (2.40)

scaled deviance = 2957.5 (change = -96.69) at cycle 4

residual df = 2908 (change = -2 )
```

The parameter estimates and standard errors for this Poisson trick fit are as below. We keep here the, same scale parameter as 1.00 but we can have any scale parameter in our approach

	estimate	s.e.	parameter
1	0.1826	0.1629	GROUP(2)
2	-2.056	0.2135	GROUP(3)
3	0.006864	0.003672	GROUP(2).AGE
4	0.03814	0.004281	GROUP(3).AGE

Table 2.3: Equivalent model of equation (2.39)

scale parameter 1.000

It is obvious from above that our approach to the multinomial logit model and the results using the Poisson trick approach of Francis et al. (1992) gives the same results with exactly the same scaled deviance 2957.5 and degrees of freedom 2908.

Our multinomial logit model approach can be preferred to the Poisson trick approach as we have the choice to consider equal or unequal response levels at each level J in our analysis. We can extract the available statistics from equation (2.39) and that is not a case when fitting via the method of directive (2.40).

2.7.3: Data set with two explanatory variables

Here we consider the artificial illustrative data from Table 1.1, regarding a social survey of young people's intentions to enter higher education (HE). We here consider a 3 - level response, 'definitely yes$_1$', 'possibly$_2$', 'definitely no$_3$', with explanatory variables Sex and Age. This can be arranged as a multinomial data with Sex and Age as categorical variables.

Explanatory Variables		Responses			Total
Sex	Age	$level_1$	$level_2$	$level_3$	
1	1	6	9	5	20
1	2	5	4	1	10
2	1	1	3	11	15
2	2	6	9	6	21

Table 2.4: Rearranged illustrative data of Table 1.1

Here the expanatory variables are labelled as Age 1, '< 21', Age 2, '21+' , and Sex 1, 'Male' , Sex 2 , 'Female' , and the macro for fitting the multinomial logit model for the above data is given in Appendix A. The multinomial logit model for main effects without the constant term is as follows:

$$\eta = \beta_1 d1 + \beta_2 d2 + \beta_3 dp1 + \beta_4 dp2 + \beta_5 dp3 + \beta_6 dp4 \qquad (2.41)$$

In the macro we have,

$$Y = \begin{pmatrix} py1 + A_{11}\eta^{(2)} + A_{12}\eta^{(3)} \\ \\ py2 + A_{22}\eta^{(3)} \end{pmatrix} \qquad \text{and}$$

d1 *d2* *dp1* *dp2* *dp3* *dp4*

$$\begin{pmatrix} A_{11} \\ 0 \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{11} \\ A_{22} \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{11}Sex \\ 0 \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{12}Sex \\ A_{22}Sex \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{11}Age \\ 0 \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{12}Age \\ A_{22}Age \end{pmatrix}_{8\times1}.$$

After fitting the above model we get the output

scaled deviance = 0.40353 at cycle 5

residual df = 2

	estimate	s.e.	parameter
1	0.4079	1.179	*d1*
2	-0.2430	1.308	*d2*
3	0.1077	0.6893	*dp1*
4	2.273	0.8125	*dp2*
5	-0.6071	0.6893	*dp3*
6	-2.103	0.8032	*dp4*

Table 2.5: Parameter estimates of model (2.41)

scale parameter 1.000

Here we have

$$\eta = 0.4079d1 - 0.2430d2 + 0.6071dp1 + 2.273dp2 - 0.6071dp3 - 2.103dp4$$

$$\eta_1 = 0.4079d1 + 0.6071dp1 + 2.273dp2$$

$$\eta_2 = -0.2430d2 - 0.6071dp3 - 2.103dp4.$$

We now compare the above statistics with those obtained by the Poisson trick approach with Sex and Age as explanatory variables. The macro is given in Appendix A.

```
$fit+group*sex(case)+group*Age(case)+Sex(case)$
```
(2.42)

```
scaled deviance = 0.40353 (change = -14.64)at cycle3

residual df = 2 (change = -4)
```

	estimate	s.e.	parameter
1	0.4079	1.179	GROUP(2)
2	-0.2430	1.308	GROUP(3)
3	0.6071	0.6892	GROUP(2).SEX
4	2.273	0.8123	GROUP(3).SEX
5	-0.6071	0.6892	GROUP(2).AGE
6	-2.103	0.8030	GROUP(3).AGE

Table 2.6: Equivalent model of equation (2.42)

scale parameter 1.000

The above Table 2.6 and Table 2.5 are almost exactly the same but obtained with two different methods. The method given in this study and the Poisson trick approach of Francis et al. (1992) give exactly the same - scaled deviance and parameter estimates if the data entered in the macros are in the same form; otherwise the parameter estimates change as explained with more detail in section (5.3).

2.8 Important aspects

The following important aspects are needed in the process of fitting a multinomial logit model equation (2.30). The macros are given in the Appendix A, where Y and N are some special matrices and are found in section (2.3). It is easy to see that we can fit any multinomial models.

a) We need to have a good knowledge about the initial start - up values in the macro. If they are not near to be the correct values, we may not get convergence or the correct degree of freedom

b) We will need to repeat the GLIM code $use startup $use loop$ until convergence is achieved.

c) We have to set a reasonable number of cycles in the macro loop. For example, we can set the program to a maximum of 500 cycles using the GLIM code $cycle 500 2 1.0 e-5 $. The higher number of cycles can be important for obtaining convergence.

d) If many response levels values have zeros then we need some extra care about the statistics found in our model because we may get the scaled deviance equal to zero. We need not to be worried if this is the case, instead we keep repeating code $use startup $use loop$ and look for convergence with the correct degree of freedom for the fitted model.

e) We can obtain the scaled deviance, degree of freedom and parameter estimates as are obtained in the Poisson trick approach without improving the weight matrix at each cycle. After observing convergence, we can use the code $use loop$ to get the correct updated weight matrix, and hence the correct standard errors.

f) If we have many 0's or 1's in the levels of the response variable, we need to make 0's either to the nearest 0's or increase the cycle levels. We may change the convergence criteria for easy convergence in fitting the model.

g) To generalize our approach, we need only some minor changes in our macros given in Appendix A.

2.9 Summary

In the approach presented in this research we need only to calculate the equation (2.11) for fitting a multinomial logit model. We have also given a procedure to find the matrices D, W and $\frac{\partial l}{\partial \eta}$ those are needed in the equation for a $J = 3$ level of response variable.

The following matrices are obtained for fitting a multinomial logit model and are denoted as follows,

a) $\Sigma = (N^T N)^{-1}$ where N is defined in equation (2.31).

b) $$H = N(N^T N)^{-1} N^T$$

$$= AD\ (D^T A^T AD\ )^{-1} D^T A^T$$

$$= AD\ (D^T WD\ )^{-1} D^T A^T,$$

where A is defined in section (2.3) and the **y** - variable is given in equation (2.32). We fit a multinomial logit model as sort of a least square fit of Y versus N with any number of explanatory variables. Our fitting procedure fits the explanatory variables without the constant term. The results in our approach are equivalent to the Poisson trick approach of Francis et al. (1992). Our fitting of a multinomial logit model is reasonably straightforward and a satisfying aspect of the algorithm is that the calculations only require a program that can handle ordinary least square so could be handled by a range of standard statistical software. Our approach uses a simple form of the Cholesky decomposition applied to a matrix which consists of diagonal sub - matrices. The formulation involves matrix inversion using array calculation. This approach is quite general can allow us to use any link function.

We can extract the correct standard statistics but the interpretation of these statistics may be quite difficult. The hat - matrix and Cook's distances are easy to extract for influence purposes in each cell. This can be contrasted with the 'Poisson Trick' approach for the multinomial logit which produces the inappropriate leverages and Cook's distances. More detail on influence measures is given in chapter 7.

CHAPTER 3

More than 3-levels of the response variable

3.1 Introduction

This chapter is an extension of chapter 2 for fitting a multinomial logit model with more than 3 - level of response variable. It is not that hard to extend the theory and the results for any number of explanatory variables for any levels of response variable. First we consider the case with $J=4$ levels of response variable, with one and more than one explanatory variables. The extension for $J=k$ levels of response variable, where k is any number of levels of response variable is given in section (3.4). An application with two explanatory variables and a 4-level response variable is given in section (3.5). The scaled deviance and parameter estimates are also compared to those obtained using the Poisson trick approach in section (3.5).

As defined in the previous chapter that the response value y_{ij} is the *ith* observation for the response level j and n_i is a total for the *ith* case with the assumption that n_i are fixed. We also assume that y_{i1}, y_{i2}, y_{i3}, y_{i4} follows a multinomial distribution for each case i. We further assume that corresponding to each y_{ij} there is a probability p_{ij}, which we will estimate.

3.2 For $J = 4$ levels of response variable

In this section we will illustrate the form of β_{new} in equation (2.11) for the case where there is a 4 - level of the response variable. In this case,

$$D_{ij} = \begin{pmatrix} X^{(2)} & 0 & 0 \\ 0 & X^{(3)} & 0 \\ 0 & 0 & X^{(4)} \end{pmatrix}_{3m\times 6} . \qquad (3.1)$$

Here we assume that $X^{(2)} = X^{(3)} = X^{(4)}$ is a common design matrix of explanatory variables for the different level of response variable,

and $$W = \begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix}_{3m\times 3m}$$, where $$W_{ij} = E(-\frac{\partial^2 l}{\partial\eta_i \partial\eta_j}) \qquad (3.2)$$

Using the Cholesky decomposition of symmetric positive definite matrices, as is given section (2.3) and equation (2.14),

$$\begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix} = \begin{pmatrix} A_{11}^2 & A_{11}A_{12} & A_{11}A_{13} \\ A_{11}A_{21} & A_{12}^2 + A_{22}^2 & A_{21}A_{31} + A_{22}A_{32} \\ A_{11}A_{31} & A_{21}A_{31} + A_{22}A_{32} & A_{13}^2 + A_{23}^2 + A_{33}^2 \end{pmatrix}, \qquad (3.3)$$

where W_{ij} and A_{ij} are diagonal matrices and $A_{ij} = A_{ji}$; also $W_{ij} = W_{ji}$.

Then, as before, $A_{11} = sqrt\ (W_{11})$, $A_{12} = W_{12} A_{11}^{-1}$,

$$A_{13} = W_{13} A_{11}^{-1}, \quad A_{22} = sqrt(W_{22} - A_{12}^2),$$

$$A_{23} = (W_{23} - A_{21} A_{31}) A_{22}^{-1},$$

and $$A_{33} = sqrt(W_{33} - A_{13}^2 - A_{23}^2).$$

The equation (1.8) here become

$$l_i = \log L_i = y_{i2}\eta_i^{(2)} + y_{i3}\eta_i^{(3)} + y_{i4}\eta_i^{(4)} - n_i \log(1 + e^{\theta_{i2}} + e^{\theta_{i3}} + e^{\theta_{i4}}), \qquad (3.4)$$

where $\eta_i^{(k)}$ is the logit link for observation i and of response level k.

Now we again further need to find $W_{ij} = E(-\frac{\partial^2 l}{\partial \eta_i \partial \eta_j})$ in equation (2.16) and from equation (3.4) we have that

$$\frac{\partial l_i}{\partial \eta_i^{(2)}} = y_{i2} - \mu_{12} \ , \quad \text{as before}$$

and $$\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_i^{(2)}} = -n_i \, p_{i2}(1 - p_{i2}) \ , \quad \text{as before} \ . \qquad (3.5)$$

Thus $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_i^{(2)}}) = = \mu_{i2}(1 - \frac{\mu_{i2}}{n_i}) , \quad \text{as before} \ . \qquad (3.6)$$

Similarly $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(3)} \partial \eta_i^{(3)}}) = \mu_{i3}(1 - \frac{\mu_{i3}}{n_i}) ,$$

and $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(4)} \partial \eta_i^{(4)}}) = \mu_{i4}(1 - \frac{\mu_{i4}}{n_i}) .$$

For $i \neq j$ in equation (3.4), $E(-\frac{\partial^2 l_i}{\partial \eta_i^{(2)} \partial \eta_j^{(2)}}) = 0$, as before

Similarly $$E(-\frac{\partial^2 l_i}{\partial \eta_i^{(3)} \partial \eta_j^{(3)}}) = E(-\frac{\partial^2 l_i}{\partial \eta_i^{(4)} \partial \eta_j^{(4)}}) = 0 .$$

Likewise $$\frac{\partial l_i}{\partial \eta_i^{(3)} \partial \eta_i^{(2)}} = n_i \, p_{i1} e^{\theta_{i2}} p_{i1} e^{\theta_{i3}} .$$

Thus $$E\left(-\frac{\partial^2 l_i}{\partial \eta_i^{(3)} \partial \eta_i^{(2)}}\right) = -\frac{\mu_{i2}\mu_{i3}}{n_i}. \qquad (3.7)$$

We denote $$\left(\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D}\right) = \left(\boldsymbol{D}^T \begin{pmatrix} A_{11} & A_{21} & A_{31} \\ 0 & A_{22} & A_{32} \\ 0 & 0 & A_{33} \end{pmatrix}^T \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix} \boldsymbol{D} \right) = \boldsymbol{N}^T \boldsymbol{N},$$

where $$\boldsymbol{N} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}_{3m \times 3m} \boldsymbol{D}.$$

Similarly to equation (2.25) we have,

$$B_{11} = A_{11}^{-1}, \qquad B_{22} = A_{22}^{-1}, \qquad B_{33} = A_{33}^{-1},$$

$$B_{21} = -(A_{21}B_{11})A_{22}^{-1} = -(A_{21}A_{11}^{-1})A_{22}^{-1},$$

$$B_{31} = -(A_{31}B_{11} + A_{32}B_{21})A_{33}^{-1} = -(A_{31}A_{11}^{-1} + A_{32}(-A_{21}A_{11}^{-1})A_{22}^{-1})A_{33}^{-1},$$

$$B_{32} = -(A_{32}B_{22})A_{33}^{-1} = -(A_{32}A_{22}^{-1})A_{33}^{-1}. \qquad (3.8)$$

Now we need $\frac{\partial l}{\partial \eta}$, where $\eta = \begin{pmatrix} \eta_i^{(2)} \\ \eta_i^{(3)} \\ \eta_i^{(4)} \end{pmatrix}$ for $i = 1, 2, \ldots, m$

and for the logit link $\eta_i^{(2)} = \theta_{i2}$, $\eta_i^{(3)} = \theta_{i3}$, $\eta_i^{(4)} = \theta_{i4}$.

As before (e.g. equation 2.26),

$$\frac{\partial l}{\partial \eta} = \begin{pmatrix} Y_2 - \mu_2 \\ Y_3 - \mu_3 \\ Y_4 - \mu_4 \end{pmatrix}_{3m \times 1}. \qquad (3.9)$$

Here $(N^T N)\beta_{new} =$

$$N^T\left(\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}\begin{pmatrix} \eta^{(2)} \\ \eta^{(3)} \\ \eta^{(4)} \end{pmatrix} + \begin{pmatrix} B_{11} & 0 & 0 \\ B_{21} & B_{22} & 0 \\ B_{31} & B_{32} & B_{33} \end{pmatrix}\begin{pmatrix} Y_2 - \mu_2 \\ Y_3 - \mu_3 \\ Y_4 - \mu_4 \end{pmatrix}\right)$$

$$= N^T\left(\begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} + A_{13}\eta^{(4)} \\ A_{22}\eta^{(3)} + A_{23}\eta^{(4)} \\ A_{33}\eta^{(4)} \end{pmatrix} + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \\ B_{31}(Y_2 - \mu_2) + B_{32}(Y_3 - \mu_2) + B_{33}(Y_4 - \mu_4) \end{pmatrix}\right)$$

But in GLIM code we have

$$(\%lp) = \begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} + A_{13}\eta^{(4)} \\ A_{22}\eta^{(3)} + A_{23}\eta^{(4)} \\ A_{33}\eta^{(4)} \end{pmatrix}_{3m\times 1} . \qquad (3.10)$$

Thus using equation (3.10),

$$N^T N\beta_{new} = N^T_{3n\times 3n}\left((\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \\ B_{31}(Y_2 - \mu_2) + B_{32}(Y_3 - \mu_3) + B_{33}(Y_4 - \mu_4) \end{pmatrix}\right)_{3m\times 1} . \qquad (3.11)$$

The above equation is same as equation (2.30), with the design matrix

$$N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} & A_{13}X^{(4)} \\ 0 & A_{22}X^{(3)} & A_{23}X^{(4)} \\ 0 & & A_{33}X^{(4)} \end{pmatrix}_{3m\times 3m} . \tag{3.12}$$

and the **y** - variable

$$\begin{pmatrix} (\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \\ B_{31}(Y_2 - \mu_2) + B_{32}(Y_3 - \mu_3) + B_{33}(Y_4 - \mu_4) \end{pmatrix} \end{pmatrix}_{3m\times 1} . \tag{3.13}$$

Or equivalently the **y** - variable is

$$\begin{pmatrix} (\%lp) + \begin{pmatrix} A_{11}^{-1}(Y_2 - \mu_2) \\ -(A_{12}A_{22}^{-1})A_{11}^{-1}(Y_2 - \mu_2) + A_{22}^{-1}(Y_3 - \mu_3) \\ -(A_{31}A_{11}^{-1} + A_{32}(-A_{21}A_{11}^{-1})A_{22}^{-1})A_{33}^{-1}(Y_2 - \mu_2) - (A_{32}A_{22}^{-1})A_{33}^{-1}(Y_3 - \mu_{32}) + A_{33}^{-1}(Y_4 - \mu_4) \end{pmatrix} \end{pmatrix}$$

where $$(\%lp) = \begin{pmatrix} lp_2 \\ lp_3 \\ lp_4 \end{pmatrix} = \begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} + A_{13}\eta^{(4)} \\ A_{22}\eta^{(3)} + A_{23}\eta^{(4)} \\ A_{33}\eta^{(4)} \end{pmatrix} . \tag{3.14}$$

In other words

$$\eta^{(4)} = A_{33}^{-1} lp_4 \ ,$$

$$\eta^{(3)} = A_{22}^{-1}(lp_3 - A_{23}\eta^{(3)}) \ = A_{22}^{-1}(lp_3 - A_{23}A_{33}^{-1}lp_4)\,.$$

$$\eta^{(2)} = A_{11}^{-1}(lp_2 - A_{12}\eta^{(3)} + A_{13}\eta^{(4)})$$

$$= A_{11}^{-1}(lp_2 - A_{12}A_{22}^{-1}(lp_3 - A_{23}A_{33}^{-1}lp_4) + A_{13}A_{33}^{-1}lp_4)\,. \qquad (3.15)$$

This derivation can thus be seen to be a simple extension of the section (2.3) and the equation (3.15) is a simple extension of equation (2.35).

3.3 More than one explanatory variable

Let us consider here x_1, x_2 and x_3 that we have three explanatory variables with a $J = 4$ level of response variable. Then, we may consider that

$$\eta = \begin{pmatrix} \eta^{(2)} \\ \eta^{(3)} \\ \eta^{(4)} \end{pmatrix} = \begin{pmatrix} X\ \beta^{(2)} \\ X\ \beta^{(3)} \\ X\ \beta^{(4)} \end{pmatrix} . \tag{3.16}$$

In this case, $X = \begin{pmatrix} 1 & x_1 & x_2 & x_3 \end{pmatrix}_{m\times 4}$,

where x_1, x_2 and x_3 are column vectors of dimension $n \times 1$,

and $$D = \begin{pmatrix} X & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1 & x_2 & x_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_1 & x_2 & x_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_1 & x_2 & x_3 \end{pmatrix}.$$

As in equation (3.12) the new design matrix N becomes

$$N = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix} \begin{pmatrix} X & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{pmatrix} \tag{3.17}$$

$$= \begin{pmatrix} A_{11} & A_{11}x_1 & A_{11}x_2 & A_{11}x_3 & A_{12} & A_{12}x_1 & A_{12}x_2 & A_{12}x_3 & A_{13} & A_{13}x_1 & A_{13}x_2 & A_{13}x_3 \\ 0 & 0 & 0 & 0 & A_{22} & A_{22}x_1 & A_{22}x_2 & A_{22}x_3 & A_{23} & A_{23}x_1 & A_{23}x_2 & A_{23}x_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{33} & A_{33}x_1 & A_{33}x_2 & A_{33}x_3 \end{pmatrix}.$$

and the **y** - variable,

$$\left((\%lp)+\begin{pmatrix} A_{11}^{-1}(Y_2-\mu_2) \\ -(A_{12}A_{22}^{-1})A_{11}^{-1}(Y_2-\mu_2)+A_{22}^{-1}(Y_3-\mu_3) \\ -(A_{31}A_{11}^{-1}+A_{32}(-A_{21}A_{11}^{-1})A_{22}^{-1})A_{33}^{-1}(Y_2-\mu_2)-(A_{32}A_{22}^{-1})A_{33}^{-1}(Y_3-\mu_{32})+A_{33}^{-1}(Y_4-\mu_4) \end{pmatrix}\right)_{3m}$$

From above we can fit the model of Y versus N. As before the equivalent fit to that model can be obtained by the Poisson trick approach, with some minor changes in the macros given in Appendix A.

3.4 For $J = k$ levels of response variable

If we have k - levels of response variable then, from section (3.2),

$$\boldsymbol{D} = \begin{pmatrix} X^{(2)} & 0 & ..... & 0 \\ 0 & X^{(3)} & ..... & 0 \\ ..... & ..... & ..... & ... \\ 0 & ..... & 0 & X^{(k)} \end{pmatrix}, \qquad (3.18)$$

where $X^{(k)}$ is a design matrix for response level k.

Using the Cholesky decomposition of symmetric positive definite matrices, we have

$$W=\begin{pmatrix} W_{11} & W_{12} & ..... & W_{1k-1} \\ W_{21} & W_{22} & ..... & W_{2k-1} \\ ..... & ..... & ..... & ..... \\ W_{k-11} & ..... & ..... & W_{k-1k-1} \end{pmatrix}=\begin{pmatrix} A_{11} & 0 & ..... & 0 \\ A_{21} & A_{22} & ..... & 0 \\ ..... & ..... & ..... & ... \\ A_{k-11} & A_{k-12} & ..... & A_{k-1k-1} \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} & ..... & A_{1k-1} \\ 0 & A_{22} & ..... & A_{2k-1} \\ ..... & ..... & ..... & ..... \\ 0 & 0 & ..... & A_{k-1k-1} \end{pmatrix}.$$

Since W_{ij} and A_{ij} are diagonal matrices and $A_{ij}=A_{ji}$, $W_{ij}=W_{ji}$, W_{ij} can be found and

are given in section (3.2).

The equation (1.7) for $J=k$ level of response variable is,

$$l_i = y_{i2}\eta_i^{(2)} + y_{i3}\eta_i^{(3)} + \ldots + y_{ik}\eta_i^{(k)} - n_i \log(1+e^{\theta_{i2}} + e^{\theta_{i3}} + \ldots + e^{\theta_{ik}}). \qquad (3.19)$$

Now taking derivatives we get, as before,

$$\frac{\partial^2 l_i}{\partial\eta_i^{(2)}\partial\eta_i^{(2)}} = -\left\{\frac{n_i e^{\theta_{i2}}(1+e^{\theta_{i2}} + e^{\theta_{i3}} + \ldots + e^{\theta_{ik}}) - n_i e^{\theta_{i2}} e^{\theta_{i2}}}{(1+e^{\theta_{i2}} + e^{\theta_{i3}} + \ldots + e^{\theta_{ik}})^2}\right\} = -n_i\, p_{i2}(1 - p_{i2}) .$$

Therefore $$E(-\frac{\partial^2 l_i}{\partial\eta_i^{(2)}\partial\eta_i^{(2)}}) = \mu_{i2}(1 - \frac{\mu_{i2}}{n_i}) . \qquad (3.20)$$

The same applies to all such derivatives; *i.e.,*

$$E(-\frac{\partial^2 l_i}{\partial\eta_i^{(r)}\partial\eta_i^{(r)}}) = \mu_{ir}(1 - \frac{\mu_{ir}}{n_i}) , \quad \text{for} \quad r = 2, 3, \ldots, k .$$

Also for $i \neq j$ $$E(-\frac{\partial^2 l_i}{\partial\eta_i^{(r)}\partial\eta_j^{(r)}}) = 0 , \quad \text{for} \quad r = 2, 3, \ldots, k .$$

We have $$E(-\frac{\partial^2 l_i}{\partial\eta_i^{(3)}\partial\eta_i^{(2)}}) = -\frac{\mu_{i2}\mu_{i3}}{n_i} ,$$

similarly for all pairs of denominators.

The basic equation (2.31) is $(\boldsymbol{D}^T \boldsymbol{W} \boldsymbol{D}) = \boldsymbol{N}^T \boldsymbol{N}$,

where $$\boldsymbol{N} = \begin{pmatrix} A_{11} & A_{12} & \ldots\ldots & A_{1k-1} \\ 0 & A_{22} & \ldots\ldots & A_{2k-1} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & \ldots\ldots & A_{k-1k-1} \end{pmatrix} \boldsymbol{D} .$$

The equation (3.11) becomes

$$\left(N^T N\right)\beta_{new} = N^T \begin{pmatrix} A_{11} & A_{12} & \ldots\ldots & A_{1k-1} \\ 0 & A_{22} & \ldots\ldots & A_{2k-1} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & \ldots\ldots & A_{k-1k-1} \end{pmatrix} \begin{pmatrix} X^{(2)} & 0 & \ldots\ldots & 0 \\ 0 & X^{(3)} & \ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots \\ 0 & \ldots\ldots & 0 & X^{(k)} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \\ .. \\ \delta \end{pmatrix}$$

$$+N^T \begin{pmatrix} B_{11} & 0 & \ldots\ldots & 0 \\ B_{21} & B_{22} & \ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots \\ B_{k-11} & B_{k-12} & \ldots\ldots & B_{k-1k-1} \end{pmatrix} \frac{\partial l}{\partial \eta} \quad , \qquad (3.21)$$

where $$\begin{pmatrix} A_{11} & 0 & \ldots\ldots & 0 \\ A_{21} & A_{22} & \ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots \\ A_{k-11} & A_{k-12} & \ldots\ldots & A_{k-1k-1} \end{pmatrix} \begin{pmatrix} B_{11} & 0 & \ldots\ldots & 0 \\ B_{21} & B_{22} & \ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots \\ B_{k-11} & B_{k-12} & \ldots\ldots & B_{k-1k-1} \end{pmatrix} = I_{(k-1)m} \quad . \qquad (3.22)$$

The B_{ij} can be found in equation (3.8).

Now we need only $\frac{\partial l}{\partial \eta}$ where $\eta = \begin{pmatrix} \eta_i^{(2)} \\ \eta_i^{(3)} \\ \ldots \\ \eta_i^{(k)} \end{pmatrix}$ for $i = 1, 2, \ldots, m$.

For the logit link

$$\frac{\partial l_i}{\partial \eta_i^{(2)}} = y_{i2} - \frac{n_i e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}} + \ldots + e^{\theta_{ik}}} = y_{i2} - \mu_{i2} \, .$$

Similarly $\frac{\partial l}{\partial \eta^{(r)}} = Y_r - \mu_r$, for $r = 2, 3, \ldots, k$.

Thus, as before $$\frac{\partial l}{\partial \eta} = \begin{pmatrix} Y_2 - \mu_2 \\ Y_3 - \mu_3 \\ \cdots\cdots \\ Y_k - \mu_k \end{pmatrix}.$$

The equation (3.21) can be written after some manipulations, as previously,

$$\left(N^T N\right)\beta_{new} =$$

$$= N^T \left((\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ B_{k-11}(Y_2 - \mu_2) + B_{k-12}(Y_3 - \mu_3) + \cdots\cdots\cdots\cdots + B_{k-1k-1}(Y_k - \mu_k) \end{pmatrix} \right).$$

The above equation is exactly as equation (3.11) with new design matrix

$$N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} & \cdots\cdots & A_{1k-1}X^{(k)} \\ 0 & A_{22}X^{(3)} & \cdots\cdots & A_{2k-1}X^{(k)} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & \cdots\cdots & A_{k-1k-1}X^{(k)} \end{pmatrix}_{m(k-1)\times 2k}.$$

and the y - variable is

$$\left((\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2) \\ B_{21}(Y_2 - \mu_2) + B_{22}(Y_3 - \mu_3) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ B_{k-11}(Y_2 - \mu_2) + B_{k-12}(Y_3 - \mu_3) + \ldots\ldots\ldots\ldots + B_{k-1k-1}(Y_k - \mu_k) \end{pmatrix} \right)_{m(k-1)\times 1}.$$

Here

$$(\%lp) = \begin{pmatrix} lp_2 \\ \ldots\ldots \\ lp_{k-1} \\ lp_k \end{pmatrix} = \begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} + \ldots\ldots\ldots\ldots\ldots\ldots + A_{1k-1}\eta^{(k)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_{k-2k-2}\eta^{(k-1)} + A_{k-2k-1}\eta^{(k)} \\ A_{k-1k-1}\eta^{(k)} \end{pmatrix}_{m(k-1)\times 1},$$

so $$\eta^{(k)} = A_{k-1k-1}^{-1} lp_k .$$

$$\eta^{(k-1)} = A_{k-2k-2}^{-1}(lp_{k-1} - A_{k-2k-1}\eta^{(k)})$$

$$= A_{k-2k-2}^{-1}(lp_{k-1} - A_{23}A_{k-1k-1}^{-1} lp_k).$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\eta^{(2)} = A_{11}^{-1}(lp_2 - A_{12}\eta^{(3)} + \ . \ . \ . \ + A_{1k-1}\eta^{(k)}).$$

3.5 Example

We continue section (2.5.1) here for the response categories 1=excellent, 2=good, 3=fair and 4=poor. In this case we use all 4 - level of response variable. Ignoring the ordered nature of the response, we fit a multinomial logit model to the data using age of the respondents as an explanatory variable. The data (responses) are arranged in four columns of observations. Here m=1456, since 17 observations are missing. The fifth column will be the explanatory variable and the sixth column is the total of responses. The $y_{i1}, y_{i2}, y_{i3}, y_{i4}$ follow a multinomial distribution and response level 1=excellent is not needed in our analysis. The standard length of our data will be 4368 because the first response 1=excellent is known when we know the total $y_{i.}$ and the responses 2=good, 3=fair, and the 4=poor.

The design matrix in fitting the multinomial logit model is,

$$N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} & A_{13}X^{(4)} \\ 0 & A_{22}X^{(3)} & A_{23}X^{(4)} \\ 0 & & A_{33}X^{(4)} \end{pmatrix}_{4368\times 6}.$$

Here A_{ij} are diagonal matrices of dimension 1456×1456 and in our notation $X^{(j)}$ are the matrices of dimension 1456×2, where the 1st column of these matrices is all 1's and the 2nd column is the explanatory variable age and the y - variable is given in equation (3.11),

$$\begin{pmatrix} (\%lp)+\begin{pmatrix} B_{11}(Y_2-\mu_2) \\ B_{21}(Y_2-\mu_2)+B_{22}(Y_3-\mu_3) \\ B_{31}(Y_2-\mu_2)+B_{32}(Y_3-\mu_3)+B_{33}(Y_4-\mu_4) \end{pmatrix} \end{pmatrix}_{3m\times 1}.$$

The multinomial logit model is fitted using macros given in Appendix A. We repeatedly use the GLIM code $use startup $use loop$ until convergence is achieved but in some cases the convergence is not easy to reach. In this case the best approach is not to update the weight matrix in each cycle. In this macro the model of main effects $fit *d1 + d2 + d3 + dp1 + dp2 + dp3* - 1$ (without the constant term) gives a fit equivalent to use of the Poisson trick approach. Both of these models are fitted here as a cross check of our results.

Here we have,

d1	*d2*	*d3*	*dp1*	*dp2*	*dp3*
$\begin{pmatrix} A_{11} \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} A_{12} \\ A_{22} \\ 0 \end{pmatrix}$	$\begin{pmatrix} A_{13} \\ A_{23} \\ A_{33} \end{pmatrix}$	$\begin{pmatrix} A_{11}Age \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} A_{12}Age \\ A_{22}age \\ 0 \end{pmatrix}$	$\begin{pmatrix} A_{13}Age \\ A_{23}Age \\ A_{33}Age \end{pmatrix}$

and after fitting the model, we get the following statistics

scaled deviance = 3264.4 at cycle 12

residual df = 4362

	estimate	s.e.	parameter
1	0.1801	0.1636	*d1*
2	-1.904	0.2240	*d2*
3	-5.487	0.4832	*d3*
4	0.007	0.0037	*dp1*
5	0.031	0.0046	*dp2*
6	0.061	0.0080	*dp3*

Table 3.1: Parameter estimates and standard errors in logit model

scale parameter 1.000

The Table 3.1 gives the same statistics as in Table 3.2 below, with some minor rounding errors in the estimates and the standard errors. The standard errors obtained by each method will be obviously different if the weight matrix is not updated at each cycle.

The Poisson trick approach equivalent to our approach (with results in Table 3.1) gives the following output:

```
$fit +group*age(case)$dis e$
```

scaled deviance = 3264.2 (change = -123.4) at cycle 6

residual df = 4362 (change = -3)

	estimate	s.e.	parameter
1	0.1801	0.1636	GROUP(2)
2	-1.904	0.2238	GROUP(3)
3	-5.485	0.4943	GROUP(4)
4	0.007	0.0037	GROUP(2).AGE
5	0.031	0.0046	GROUP(3).AGE
6	0.071	0.0080	GROUP(4).AGE

Table 3.2: Equivalent model of equation (3.14)

scale parameter 1.000

CHAPTER 4

Multinomial Model with Own Links

4.1 Introduction

In this chapter we will give some more general formulation and theory for the fitting of a multinomial model when we have a user defined link function. In previous chapters, we concentrated on the logit link function only but here we will broaden that concept with so called 'own link functions'. The own link functions used in this chapter are given in section (4.2) and can be used to form an equation equivalent to (1.7), which used for the logit link when determining the weight matrix W_{ij} .

An alternative approach to find the weight matrix W_{ij} is given in section (4.4) by using a simple equation (4.19) and it can also be used in general to find W_{ij} . The extension of section (4.3) is given in section (4.5) with the general form of W_{ij} . A practical example is given in section (4.6) with an interval estimate for a single parameter of a link function. We also give interval estimates and some contour plots of the deviances for different parameter values in our own link function.

The confidence limits for parameters in our own link functions can be found from a plot of the likelihood, as in section (4.6.2), and are illustrated for one and more than one parameter. It is explained there in examples of section (4.6) that for particular values of

the parameter in our own link function, that coincides with logit link function given in previous chapters.

4.2 Our own link function

In this section we consider link functions that can be applied easily within the framework of fitting a multinomial logit model explained in previous chapters. Using the notation in section (1.4), for $\eta_{ij}=\theta_{ij}$ and equation (1.13), we have for the logit link that

$$\eta_{i2}= \theta_{i2} = \log\left(\frac{p_{i2}}{p_{i1}}\right) \quad \text{and} \quad \eta_{i3} = \theta_{i3} = \log\left(\frac{p_{i3}}{p_{i1}}\right).$$

Equivalently, we have

$$\theta_{i2} = \log\left(\frac{p_{i2}}{p_{i1}}\right) \Rightarrow e^{\theta_{i2}} = \frac{p_{i2}}{p_{i1}} \Rightarrow p_{i2} = p_{i1}e^{\theta_{i2}}. \tag{4.1}$$

$$\theta_{i3} = \log\left(\frac{p_{i3}}{p_{i1}}\right) \Rightarrow e^{\theta_{i3}} = \frac{p_{i3}}{p_{i1}} \Rightarrow p_{i3} = p_{i1}e^{\theta_{i3}}.$$

We shall now keep these definitions for the log - odds θ_{ij} but remove the logit link assumption that $\eta_{ij} = \theta_{ij}$.

4.2.1 A convenient single parameter own link function

We define here own link functions η_{i2} and η_{i3} in a similar way to equation (1.13), namely

$$\eta_{i2} = \frac{\left(\frac{p_{i2}}{p_{i1}}\right)^{a} - 1}{a} \Rightarrow \frac{p_{i2}}{p_{i1}} = (1 + a\eta_{i2})^{\frac{1}{a}}. \tag{4.2}$$

Thus $$e^{\theta_{i2}} = (1 + a\eta_{i2})^{\frac{1}{a}} \Rightarrow \eta_{i2} = \frac{e^{a\theta_{i2}} - 1}{a}. \tag{4.3}$$

It can be shown using L' Hospital's rule that

$$\eta_{i2} = \frac{e^{a\theta_{i2}} - 1}{a} \quad \rightarrow \quad \left. \frac{\theta_{i2} e^{a\theta_{i2}}}{1} \right|_{a=0} = \theta_{i2} \quad \text{as} \quad a \rightarrow 0 \qquad (4.3a)$$

Thus $\eta_{i2} = \dfrac{e^{a\theta_{i2}} - 1}{a}$ can be thought of as a general form of logit link function.

Similarly $\eta_{i3} = \dfrac{(\frac{p_{i3}}{p_{i1}})^{a} - 1}{a} \quad \Rightarrow \quad \eta_{i3} = \dfrac{e^{a\theta_{i3}} - 1}{a}$. (4.4)

Using equations (4.1) and (4.2), we can rewrite the link functions in terms of the log - odds as follows

$$\theta_{i2} = \frac{1}{a}\log(1 + a\eta_{i2}), \qquad \theta_{i3} = \frac{1}{a}\log(1 + a\eta_{i3}) .$$

We now derive the log-likelihood in terms of η_{i2} and η_{i3}.

From equation (4.2), $p_{i2} = p_{i1}(1 + a\eta_{i2})^{\frac{1}{a}}$, (4.5)

and from equation (4.4) $p_{i3} = p_{i1}(1 + a\eta_{i3})^{\frac{1}{a}}$.

Therefore $p_{i3} = p_{i2} \dfrac{(1 + a\eta_{i3})^{\frac{1}{a}}}{(1 + a\eta_{i2})^{\frac{1}{a}}}$.

Since $p_{i1} = 1 - p_{i2} - p_{i3}$,

we can write $$p_{i1} = 1 - p_{i2} - p_{i2} \frac{(1 + a\eta_{i3})^{\frac{1}{a}}}{(1 + a\eta_{i2})^{\frac{1}{a}}}.$$

Using equation (4.5) $$p_{i2} = (1 + a\eta_{i2})^{\frac{1}{a}} \left(1 - p_{i2} - p_{i2} \frac{(1 + a\eta_{i3})^{\frac{1}{a}}}{(1 + a\eta_{i2})^{\frac{1}{a}}}\right),$$

or $$p_{i2}[1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}] = (1 + a\eta_{i2})^{\frac{1}{a}},$$

i.e. we have $$p_{i2} = \frac{(1 + a\eta_{i2})^{\frac{1}{a}}}{[1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}]}. \qquad (4.6)$$

Similarly $$p_{i3} = \frac{(1 + a\eta_{i3})^{\frac{1}{a}}}{[1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}]},$$

and $$p_{i1} = \frac{1}{[1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}]}.$$

Thus the equation (1.5) for the log - likelihood form can be rewritten as

$$\log L_i = n_i \log (p_{i1}) + y_{i2} \log (\frac{p_{i2}}{p_{i1}}) + y_{i3} \log (\frac{p_{i3}}{p_{i1}})$$

$$= -n_i \log(1 + e^{\theta_{i2}} + e^{\theta_{i3}}) + y_{i2}\theta_{i2} + y_{i3}\theta_{i3}.$$

$$= -n_i \log[1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}] + y_{i2}\{\frac{1}{a}\log(1 + a\eta_{i2})\}$$

$$+ \; y_{i3}\{\frac{1}{a}\log(\;1 + a\,\eta_{i3})\}. \qquad (4.7)$$

The above equation is a log - likelihood of multinomial model as is equation (1.7) in terms of our 'own link function'. This is an important equation in our study to find the parameters β_{new} and is used in section (4.3).

4.2.2 A more general own link function

We can generalize our own link functions for η_{i2} and η_{i3} given in the previous section by utilising different parameters 'a' and 'b' for η_{i2} and η_{i3} respectively. Using different parameters, we assume that

$$\eta_{i2} = \frac{(\frac{p_{i2}}{p_{i1}})^{a} - 1}{a} \quad \Rightarrow \quad \frac{p_{i2}}{p_{i1}} = (a\,\eta_{i2} + 1)^{\frac{1}{a}},$$

and $$\eta_{i3} = \frac{(\frac{p_{i3}}{p_{i1}})^{b} - 1}{b} \quad \Rightarrow \quad \frac{p_{i3}}{p_{i1}} = (b\,\eta_{i3} + 1)^{\frac{1}{b}}. \qquad (4.8)$$

Then the equation (4.7) becomes,

$$\log\; L_i = n_i \log\;(p_{i1}) + y_{i2} \log\;(\frac{p_{i2}}{p_{i1}}) + y_{i3} \log\;(\frac{p_{i3}}{p_{i1}})$$

$$= -n_i \log[\,1 + (1 + a\,\eta_{i2})^{\frac{1}{a}} + (1 + b\,\eta_{i3})^{\frac{1}{b}}\,] + y_{i2}\{\frac{1}{a}\log(\,1 + a\,\eta_{i2})\}$$

$$+ \; y_{i3}\{\frac{1}{b}\log(\;1 + b\,\eta_{i3})\}\;. \qquad (4.9)$$

4.3 Parameter estimates β_{new} for $J=3$

We will illustrate here how to estimate β_{new} using our single parameter own link function for a 3 - level of response variable. An alternative derivation of this is also given in section (4.4). Section (4.4) is much more succinct and the reader may omit section (4.3) for ease of reading.

To find the W_{ij} we need the derivatives,

$$\frac{\partial l_i}{\partial \eta_{i2}} = y_{i2}\left(\frac{1}{1+a\eta_{i2}}\right) - \frac{n_i (1+a\eta_{i2})^{\frac{1}{a}}}{[1+(1+a\eta_{12})^{\frac{1}{a}} + (1+a\eta_{i3})^{\frac{1}{a}}][1+a\eta_{i2}]}$$

$$= \frac{1}{(1+a\eta_{i2})}\left\{ y_{i2} - \frac{n_i (1+a\eta_{i2})^{\frac{1}{a}}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}} + (1+a\eta_{i3})^{\frac{1}{a}}\}} \right\}.$$

Thus $$\frac{\partial^2 l_i}{\partial^2 \eta_{i2}} = \frac{-a}{(1+a\eta_{i2})^2}\left\{ y_{i2} - \frac{n_i (1+a\eta_{i2})^{\frac{1}{a}}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}} + (1+a\eta_{i3})^{\frac{1}{a}}\}} \right\} +$$

$$\frac{-n_i}{(1+a\eta_{i2})}\left\{ \frac{(1+a\eta_{i2})^{\frac{1}{a}-1}\{1+(1+a\eta_{i2})^{\frac{1}{a}} + (1+a\eta_{i3})^{\frac{1}{a}}\} - (1+a\eta_{i2})^{\frac{1}{a}}(1+a\eta_{i2})^{\frac{1}{a}-1}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}} + (1+a\eta_{i3})^{\frac{1}{a}}\}^2} \right\}$$

$$= \frac{-a}{(1+a\eta_{i2})^2}\left\{ y_{i2} - \frac{n_i e^{\theta_{i2}}}{\{1+e^{\theta_{i2}} + e^{\theta_{i3}}\}} \right\} +$$

$$\frac{-n_i}{(1+a\eta_{i2})^2}\left\{\frac{n_i e^{\theta_{i2}}\{1+e^{\theta_{i2}}+e^{\theta_{i3}}\}-n_i e^{\theta_{i2}}e^{\theta_{i2}}}{\{1+e^{\theta_{i2}}+e^{\theta_{i3}}\}^2}\right\},$$

$$=\frac{-a}{(1+a\eta_{i2})^2}\{y_{i2}-n_i p_{i2}\}+\frac{1}{(1+a\eta_{i2})^2}\{-n_i e^{\theta_{i2}}p_{i1}+n_i(e^{\theta_{i2}}p_{i1})^2\}, \quad (4.10)$$

$$=\frac{-a}{(1+a\eta_{i2})^2}\{y_{i2}-n_i e^{\theta_{i2}}p_{i1}\}+\frac{n_i p_{i2}(1-p_{i2})}{(1+a\eta_{i2})^2}. \quad (4.11)$$

Thus $$E(-\frac{\partial^2 l_i}{\partial^2\eta_{i2}})=0+\frac{n_i p_{i2}(1-p_{i2})}{(1+a\eta_{i2})^2}$$

$$=\mu_{i2}(1-\frac{\mu_{i2}}{n_i})\frac{1}{(1+a\eta_{i2})^2}. \quad (4.12)$$

Similarly $$E(-\frac{\partial^2 l_i}{\partial^2\eta_{i3}})=\mu_{i3}(1-\frac{\mu_{i3}}{n_i})\frac{1}{(1+a\eta_{i3})^2}.$$

Now, for $i \neq j$ in equation (4.7),

$$\frac{\partial^2 l_i}{\partial\eta_{i2}\partial\eta_{j2}}=\frac{\partial}{\partial\eta_{j2}}\{\frac{1}{(1+a\eta_{i2})}\left\{y_{i2}-\frac{n_i(1+a\eta_{i2})^{\frac{1}{a}}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}}+(1+a\eta_{i3})^{\frac{1}{a}}\}}\right\}\}$$

Since for the own link function $E\left(-\frac{\partial^2 l_i}{\partial\eta_{i2}\partial\eta_{j2}}\right) = E\left(-\frac{\partial^2 l_i}{\partial\eta_{i3}\partial\eta_{j3}}\right) = 0$, as for the logit link function, the own link function does not make much difference in the calculation of the weight matrix.

Now
$$\frac{\partial^2 l_i}{\partial\eta_{i3}\partial\eta_{i2}} = \frac{\partial}{\partial\eta_{i3}}\left\{\frac{1}{(1+a\eta_{i2})}\left\{y_{i2} - \frac{n_i(1+a\eta_{i2})^{\frac{1}{a}}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}}+(1+a\eta_{i3})^{\frac{1}{a}}\}}\right\}\right\}$$

$$= 0 + \frac{\partial}{\partial\eta_{i3}}\left\{\frac{-1}{(1+a\eta_{i2})}\left\{\frac{n_i(1+a\eta_{i2})^{\frac{1}{a}}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}}+(1+a\eta_{i3})^{\frac{1}{a}}\}}\right\}\right\}$$

$$= \left\{\frac{1}{(1+a\eta_{i2})}\left\{\frac{n_i(1+a\eta_{i2})^{\frac{1}{a}}(1+a\eta_{i3})^{\frac{1}{a}-1}}{\{1+(1+a\eta_{i2})^{\frac{1}{a}}+(1+a\eta_{i3})^{\frac{1}{a}}\}^2}\right\}\right\}$$

$$= \frac{1}{(1+a\eta_{i2})(1+a\eta_{i3})}\, n_i\, p_{i1}e^{\theta_{i2}}\, p_{i1}e^{\theta_{i3}}\,.$$

Thus
$$E\left(-\frac{\partial^2 l_i}{\partial\eta_{i3}\partial\eta_{i2}}\right) = \frac{1}{(1+a\eta_{i2})(1+a\eta_{i3})}\left\{-\mu_{i2}\,\frac{\mu_{i3}}{n_i}\right\}$$

$$= \frac{1}{(1+a\eta_{i2})(1+a\eta_{i3})}\left\{-\frac{\mu_{i2}\mu_{i3}}{n_i}\right\}.$$

We require $\frac{\partial l}{\partial \eta}$ in equation (4.7) where we have $\eta = \begin{pmatrix} \eta_{i2} \\ \eta_{i3} \end{pmatrix}$ for $i = 1, 2, 3, \ldots, m$

and $$\eta_{i2} = \frac{e^{a\theta_{i2}} - 1}{a}, \quad \theta_{i2} = \frac{1}{a}\log(1 + a\eta_{i2}),$$

$$\eta_{i3} = \frac{e^{a\theta_{i3}} - 1}{a}, \quad \theta_{i3} = \frac{1}{a}\log(1 + a\eta_{i3}).$$

Here $$\frac{\partial l_i}{\partial \eta_{i2}} = y_{i2}\left(\frac{1}{1 + a\eta_{i2}}\right) - \frac{n_i (1 + a\eta_{i2})^{\frac{1}{a}}}{[1 + (1 + a\eta_{12})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}][1 + a\eta_{i2}]}$$

$$= \frac{1}{(1 + a\eta_{i2})}\left\{ y_{i2} - \frac{n_i (1 + a\eta_{i2})^{\frac{1}{a}}}{\{1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + a\eta_{i3})^{\frac{1}{a}}\}} \right\}$$

$$= \{ y_{i2} - n_i p_{i2} \} \frac{1}{1 + a\eta_{i2}}$$

$$= \{ y_{i2} - \mu_{i2} \} \frac{1}{1 + a\eta_{i2}}.$$

Thus $$\frac{\partial l}{\partial \eta_2} = \{Y_2 - \mu_2\} \frac{1}{1 + a\eta_2}, \quad \text{where } Y_j^T = (y_{1j}\ y_{2j} \cdots y_{nj})^T \quad j = 2, 3$$

Similarly $$\frac{\partial l}{\partial \eta_3} = \{Y_3 - \mu_3\} \frac{1}{1 + a\eta_3}.$$

Or
$$\frac{\partial l}{\partial \eta} = \begin{pmatrix} \{Y_2 - \mu_2\}\dfrac{1}{1+a\eta_2} \\ \{Y_3 - \mu_3\}\dfrac{1}{1+a\eta_3} \end{pmatrix}. \qquad (4.13)$$

The equation (2.23) can be rewritten,

$$N^T N\beta_{new} = N^T \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} \{Y_2 - \mu_2\}\dfrac{1}{1+a\eta_2} \\ \{Y_3 - \mu_3\}\dfrac{1}{1+a\eta_3} \end{pmatrix} \right)$$

$$= N^T \left(\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} \eta_2 \\ \eta_3 \end{pmatrix} + \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} \{Y_2 - \mu_2\}\dfrac{1}{1+a\eta_2} \\ \{Y_3 - \mu_3\}\dfrac{1}{1+a\eta_3} \end{pmatrix} \right)$$

$$= N^T \left(\begin{pmatrix} A_{11}\eta_2 + A_{12}\eta_3 \\ A_{22}\eta_3 \end{pmatrix} + \begin{pmatrix} B_{11}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} \\ B_{21}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + B_{22}(Y_3 - \mu_3)\dfrac{1}{1+a\eta_3} \end{pmatrix} \right). \qquad (4.14)$$

In GLIM code $(\%\ lp) = \begin{pmatrix} A_{11}\eta_2 + A_{12}\eta_3 \\ A_{22}\eta_3 \end{pmatrix}$.

Thus equation (4.14) is equivalent to

$$N^T N\beta_{new} = N^T \left((\%lp) + \begin{pmatrix} B_{11}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} \\ B_{21}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + B_{22}(Y_3 - \mu_3)\dfrac{1}{1+a\eta_3} \end{pmatrix} \right),$$

where $$N = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix} = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} \\ 0 & A_{22}X^{(3)} \end{pmatrix},$$

and the y - variable is

$$\left((\%lp) + \begin{pmatrix} A_{11}^{-1}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} \\ -(A_{12}A_{22}^{-1})A_{11}^{-1}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + A_{22}^{-1}(Y_3 - \mu_3)\dfrac{1}{1+a\eta_3} \end{pmatrix} \right), \qquad (4.15)$$

where $$(\%lp) = \begin{pmatrix} lp_2 \\ lp_3 \end{pmatrix} = \begin{pmatrix} A_{11}\eta_2 + A_{12}\eta_3 \\ A_{22}\eta_3 \end{pmatrix}. \qquad (4.16)$$

Or we have $\eta_3 = A_{22}^{-1} lp_3$,

$$\eta_2 = A_{11}^{-1}(lp_2 - A_{12}\eta_3). \tag{4.17}$$

The way we have calculated the weight matrix here to find β_{new} may not used for some other own link functions because it is not easy in every case to write the equation (4.7). If this is the case then an alternative formulation to find the weight matrix W_{ij} is given in the next section.

4.4 An alternative derivation of W_{ij}

The parameter estimates β_{new} calculated in the previous section depends on the log - likelihood approach of section (2.3) for finding the weight matrix W_{ij}. The derivatives to find W_{ij} can be found easily using our own link function and log - odds θ_{ij} without the assumption $\eta_{ij} = \theta_{ij}$ as follows:

$$\frac{\partial l_i}{\partial \eta_{i2}} = \frac{\partial l_i}{\partial \theta_{i2}} \times \frac{\partial \theta_{i2}}{\partial \eta_{i2}}, \tag{4.18}$$

where $\frac{\partial l_i}{\partial \theta_{i2}}$ are found from the equation (4.7) without the assumption $\eta_{ij} = \theta_{ij}$

and also where we can write $\frac{\partial \theta_{i2}}{\partial \eta_{i2}} = \frac{1}{\frac{\partial \eta_{i2}}{\partial \theta_{i2}}}$.

Thus from equation (4.3) $\frac{\partial \eta_{i2}}{\partial \theta_{i2}} = \frac{ae^{a\theta_{i2}}}{a} = e^{a\theta_{i2}} = (1 + a\eta_{i2})$ and the equation (4.18) can be rewritten as

$$\frac{\partial l_i}{\partial \eta_{i2}} = \frac{\partial l_i}{\partial \theta_{i2}} \times \frac{1}{1 + a\eta_{i2}} = \frac{\partial l_i}{\partial \theta_{i2}} \times e^{-a\theta_{i2}},$$

and $$\frac{\partial^2 l_i}{\partial^2 \eta_{i2}} = \frac{\partial}{\partial \theta_{i2}} \{ \frac{\partial l_i}{\partial \theta_{i2}} \times e^{-a\theta_{i2}} \} \frac{\partial \theta_{i2}}{\partial \eta_{i2}}.$$

The full derivation with all the derivatives of the weight matrix W_{ij} is given in the Appendix D. This approach is recommended to find W_{ij} in any general case, without the need to use the log - likelihood approach of equation section (2.3).

4.5 More than one parameter in our own link function

We can estimate the parameters β_{new} using our own link function with more than one parameter. For illustration we use Green's data as in the example of chapter 3. We fit a multinomial model using age of the respondent as an explanatory variable and using our own link function with more than one parameter. The data are arranged as before with responses in the first four columns and with the fifth column containing the sole explanatory variable, age. The sixth column is the total of responses.

We assume $y_{i1}, y_{i2}, y_{i3}, y_{i4}$ follows a multinomial distribution with 'own link functions' with different parameters as follows:

$$\eta_{i2} = \frac{(\frac{p_{i2}}{p_{i1}})^{a} - 1}{a} = (1 + a\eta_{i2})^{\frac{1}{a}} = \frac{e^{a\theta_{i2}} - 1}{a}$$

$$\eta_{i3} = \frac{(\frac{p_{i2}}{p_{i1}})^{b} - 1}{b} = (1 + a\eta_{i3})^{\frac{1}{b}} = \frac{e^{b\theta_{i3}} - 1}{b}$$

$$\eta_{i4} = \frac{(\frac{p_{i4}}{p_{i1}})^{c} - 1}{c} = (1 + a\eta_{i4})^{\frac{1}{c}} = \frac{e^{c\theta_{i4}} - 1}{c}.$$

The log - likelihood equation (4.7) can be rewritten as

$$\log L_i = n_i \log (p_{i1}) + y_{i2} \log (\frac{p_{i2}}{p_{i1}}) + y_{i3} \log (\frac{p_{i3}}{p_{i1}}) + y_{i4} \log (\frac{p_{i4}}{p_{i1}})$$

$$= -n_i \log[\, 1 + (1 + a\eta_{i2})^{\frac{1}{a}} + (1 + b\eta_{i3})^{\frac{1}{b}} + (1 + c\eta_{i2})^{\frac{1}{c}}\,]$$

$$+ y_{i2}\{\frac{1}{a}\log(1+a\eta_{i2})\} + y_{i3}\{\frac{1}{b}\log(1+b\eta_{i3})\} + y_{i4}\{\frac{1}{c}\log(1+c\eta_{i4})\}. \qquad (4.19)$$

The equation (4.14) becomes

$$N^T N\beta_{new} = N^T\left(\begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} + A_{13}\eta^{(4)} \\ A_{22}\eta^{(3)} + A_{23}\eta^{(4)} \\ A_{33}\eta^{(4)} \end{pmatrix} + \begin{pmatrix} B_{11}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} \\ B_{21}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + B_{22}(Y_3 - \mu_3)\dfrac{1}{1+b\eta_3} \\ B_{31}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + B_{32}(Y_3 - \mu_2)\dfrac{1}{1+b\eta_3} + B_{33}(Y_4 - \mu_4)\dfrac{1}{1+c\eta_4} \end{pmatrix}\right)$$

with the design matrix

$$N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} & A_{13}X^{(4)} \\ 0 & A_{22}X^{(3)} & A_{23}X^{(4)} \\ 0 & & A_{33}X^{(4)} \end{pmatrix}_{3m\times 3m}$$

and the **y** - variable

$$\left(\begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} + A_{13}\eta^{(4)} \\ A_{22}\eta^{(3)} + A_{23}\eta^{(4)} \\ A_{33}\eta^{(4)} \end{pmatrix} + \begin{pmatrix} B_{11}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} \\ B_{21}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + B_{22}(Y_3 - \mu_3)\dfrac{1}{1+b\eta_3} \\ B_{31}(Y_2 - \mu_2)\dfrac{1}{1+a\eta_2} + B_{32}(Y_3 - \mu_2)\dfrac{1}{1+b\eta_3} + B_{33}(Y_4 - \mu_4)\dfrac{1}{1+c\eta_4} \end{pmatrix}\right).$$

The above design matrix and the y - variable for such link functions with different parameters can be generalized further without any major problem.

4.6 Examples

We continue here to use Green's data, as given in example of chapter 3 when applying a Box - Cox own link function with a single parameter 'a' (with a 4 - levels of response variable and age as a explanatory variable). These data are here further used for more general Box - Cox own link functions, and confidence limits for parameters in these link function are illustrated. The confidence limits indicate that a logit link function is acceptable for this particular data set.

4.6.1 Single parameter own link function

As we have a data for four levels of response variable and with one explanatory variable, the own link functions in this case will be taken as

$$\eta_{i2} = \frac{e^{a\theta_{i2}} - 1}{a}, \quad \eta_{i3} = \frac{e^{a\theta_{i3}} - 1}{a} \quad \text{and} \quad \eta_{i4} = \frac{e^{a\theta_{i4}} - 1}{a}.$$

The design matrix for fitting the multinomial model is

$$N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} & A_{13}X^{(4)} \\ 0 & A_{22}X^{(3)} & A_{23}X^{(4)} \\ 0 & 0 & A_{33}X^{(4)} \end{pmatrix}_{3n\times 6}.$$

Here as in section (3.5), A_{11}, A_{12}, A_{13}, A_{22}, A_{23}, A_{33} are diagonal matrices of dimension 1456 × 1456 and $X^{(2)} = X^{(3)} = X^{(4)}$ are the matrices of dimension 1456×2, where the 1st column of matrices is all 1's and the 2nd column is a explanatory variable 'age'. The y - variable is given as in equation (4.15) and the multinomial logit model is fitted using the macros given in Appendix A.

We fit the model for fixed values of the link parameter 'a' using a looping macro. We first fit the model with a = 0.0001, which should be close to the logit link fit.

The GLIM linear predictor of 'main effects' only (without the GLIM constant term) is

$$\eta = \beta_1 d1 + \beta_2 d2 + \beta_3 d3 + \beta_4 dp1 + \beta_5 dp2 + \beta_6 dp3 . \qquad (4.20)$$

We have in equation (4.20)

$$\underline{d1} \quad \underline{d2} \quad \underline{d3} \quad \underline{dp1} \quad \underline{dp2} \quad \underline{dp3}$$

$$\begin{pmatrix} A_{11} \\ 0 \\ 0 \end{pmatrix}_{4368\times1} \begin{pmatrix} A_{12} \\ A_{22} \\ 0 \end{pmatrix} \begin{pmatrix} A_{13} \\ A_{23} \\ A_{33} \end{pmatrix} \begin{pmatrix} A_{11}X^{(2)} \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} A_{12}X^{(3)} \\ A_{22}X^{(3)} \\ 0 \end{pmatrix} \begin{pmatrix} A_{13}X^{(4)} \\ A_{23}X^{(4)} \\ A_{33}X^{(4)} \end{pmatrix}_{4368\times1} .$$

We get the statistics

scaled deviance = 3264.2 at cycle 9

residual df = 4362

	estimate	s.e.	parameter
1	0.1801	0.1636	*d1*
2	-1.904	0.2240	*d2*
3	-5.486	0.4834	*d3*
4	0.0069	0.0037	*dp1*
5	0.0306	0.0046	*dp2*
6	0.0705	0.0079	*dp3*

Table 4.1: Parameter estimates of model (4.21)

scale parameter 1.000

We can write the linear predictor as

$$\eta = 0.1801d1 - 1.904\,d2 - 5.484d3 + 0.0069dp1 + 0.0306dp2 + 0.0705dp3 \qquad (4.21)$$

It may be observed that the fit here is almost identical to the logit link fit in chapter 3, as expected.

4.6.2 Interval estimate for single parameter 'a' in our own link function

Now we investigate if the logit link function seems acceptable in this example. The model is fitted for different values of 'a' (see section 4.6.1) and the overall maximum likelihood (minimum scaled deviance) can be found by inspection. An interval estimate of 'a' can then be found. Here we give the scaled deviance for different values of parameter 'a'.

a	Deviance
-0.04	No Convergence
-0.02	3264.9
-0.01	3264.7
-0.0001	3264.4
0.0001	3264.2
0.05	3263.4
0.1	3262.8
0.2	3262.4
0.3	3263.3
0.4	3265.4
0.5	3268.9
0.55	No Convergence

Table 4.2: Scaled deviance for different values of parameter 'a'.

The above Table 4.2 gives the deviance for different values of parameter 'a'. For $-0.02 \leq a \leq 0.5$, convergence is achieved. There may be the case that for $a \geq 0.50000009$ or some other values of 'a' the convergence can be achieved but it will be a very tiny increase or decrease in the scaled deviance outside limits $-0.02 \leq a \leq 0.5$. The minimum scaled deviance (maximum likelihood) estimate of parameter 'a' can be found for plotting the deviance (or likelihood) as a function of 'a', as illustrated in the following figure.

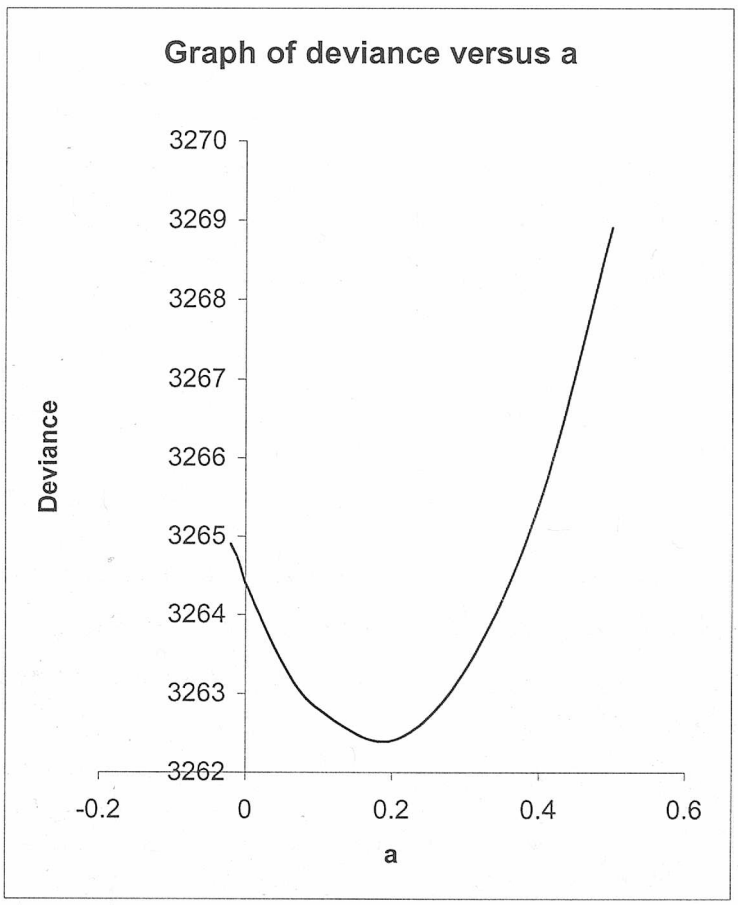

Fig. 4.1: Graph of Deviance versus 'a' in model (4.21)

The maximum likelihood estimate $\hat{a}$ can be obtained by visual inspection of this graph. Using a Chi-square approximation to the distribution of the scaled deviance, a 95 percent interval estimate for the parameter 'a' is then given by set of values of parameter 'a' such that, Scaled deviance(a) – Scaled deviance $(\hat{a})$ < 3.84 $= \chi^2_{0.05(1)}$ (GLIM manual page 290) . For cases where the scaled deviance equals the deviance (as for the multinomial), this interval estimate is $\{ a \mid D(a) < D(\hat{a}) + 3.84 \}$. Thus in this example we have,

$$\{ a \mid D(a) < 3262.4 + 3.84 \},$$

or $\{ a \mid D(a) < 3266.24 \}$ *i.e.* approximately (-0.02 , 0.4).

It appears that $a = 0$ lies within the 95% confidence interval, hence the logit link seems acceptable within the Box - Cox link family with just one extra parameter 'a'. We generalize this in the case of more than one parameter in the following section.

4.6.3 Different parameters own link function

We continue further here the data used above for different parameters 'a', 'b' and 'c' for the Box - Cox own link functions, with four levels of response variable and with 'age' of the respondents as a covariate. Thus we have

$$\eta_{i2} = \frac{e^{a\theta_{i2}} - 1}{a}, \qquad \eta_{i3} = \frac{e^{b\theta_{i3}} - 1}{b},$$

and $$\eta_{i4} = \frac{e^{c\theta_{i4}} - 1}{c}.$$

The design matrix and the y - variable are obtained with some minor changes from the single parameter own link function case. The multinomial model is first fitted using the macros given in Appendix A.

We fit the GLIM linear predictor of 'main effects' only without the constant term as follows:

$$\eta = \beta_1 d1 + \beta_2 d2 + \beta_3 d3 + \beta_4 dp1 + \beta_5 dp2 + \beta_6 dp3 . \qquad (4.22)$$

We mutually set `a = b = c = 0.0001,` and get the statistics as,

scaled deviance = `3264.2 at cycle 9`

residual df = `4362`

	estimate	s.e.	parameter
1	0.1801	0.1636	*d1*
2	-1.904	0.2240	*d2*
3	-5.486	0.4834	*d3*
4	0.0069	0.0037	*dp1*
5	0.0306	0.0046	*dp2*
6	0.0705	0.0079	*dp3*

Table 4.3: Parameter estimates of model (4.22)

scale parameter `1.000`

Equivalently, we can write

$$\eta = 0.1801d1 - 1.904d2 - 5.486d3 + 0.0069\,dp1 + 0.0360dp2 + 0.0705dp3 \qquad (4.23)$$

This is almost the same fit as equation (4.21) because the parameter values in the Box - Cox links give links that are almost equivalent to the logit link. This model can have different parameter estimates with different scaled deviances if the own link function parameter values are changed. These scaled deviances for various 'a' , 'b' and 'c' values is calculated and is given in Table 4.4.

4.6.4 Interval estimate for different parameters in own link function

We now check to see if the logit link function seems acceptable for the model (4.22) by fitting different parameter values of 'a', 'b' and 'c'. The overall minimum scaled deviance can be found from the following table.

a	b / c	-0.02	-0.01	0.0001	0.1	0.2	0.4	0.5
-0.02	-0.02	3264.9	3265.4	3265.5	3269.8	3274.8	3284.3	3288.6
	-0.01	3265.2	3265.4	3265.6	3269.9	3274.8	3284.3	3288.6
	0.0001	3265.5	3265.6	3265.8	3270.0	3270.0	3284.3	3288.6
	0.1	3266.8	3267.2	3267.3	3275.4	3275.4	3284.5	3288.7
	0.2	3268.3	3268.6	3268.7	3275.9	3275.9	3284.7	3288.8
	0.4	3270.8	3270.0	3271.1	3277.0	3277.0	3285.9	3289.1
	0.5	3271.9	3271.8	3272.3	3277.8	3277.8	3285.1	3294.7
-0.01	-0.02	3264.3	3264.5	3264.8	3269.1	3274.1	3283.8	3288.2
	-0.01	3264.5	3264.6	3264.9	3269.2	3274.2	3283.8	3288.2
	0.0001	3264.7	3264.8	3265.1	3269.3	3274.2	3283.8	3288.2
	0.1	3266.2	3266.4	3266.6	3270.2	3274.7	3284.0	3288.3
	0.2	3267.6	3268.0	3268.0	3271.2	3275.2	3284.2	3288.4
	0.4	3270.2	3270.6	3270.6	3273.1	3276.3	3284.5	3288.6
	0.5	3271.4	3271.8	3271.8	3273.9	3276.8	3284.5	3288.9
0.0001	-0.02	3263.5	3263.8	3264.0	3268.4	3273.4	3283.2	3287.7
	-0.01	3263.7	3263.9	3264.2	3268.5	3273.5	3283.3	3287.7
	0.0001	3263.8	3264.1	3264.4	3268.5	3273.6	3283.3	3287.7
	0.1	3265.5	3265.7	3265.9	3269.5	3274.1	3283.4	3287.8
	0.2	3267.0	3267.3	3267.3	3270.1	3274.6	3283.6	3287.9
	0.4	3269.9	3269.9	3269.9	3272.1	3275.8	3284.0	3288.2
	0.5	3272.1	3271.2	3271.1	3273.0	3276.1	3284.2	3288.3
0.1	-0.02	3257.7	3257.9	3258.1	3261.6	3267.0	3278.0	3283.1
	-0.01	3257.9	3258.1	3258.2	3261.7	3267.0	3278.1	3283.8
	0.0001	3258.1	3258.2	3258.4	3261.8	3267.1	3278.0	3283.1
	0.1	3259.8	3259.8	3260.0	3262.8	3267.5	3278.2	3283.2
	0.2	3261.2	3261.3	3261.5	3263.9	3268.2	3278.4	3283.2
	0.4	3263.9	3264.6	3264.5	3266.0	3269.3	3278.8	3283.5
	0.5	3266.2	3265.9	3265.7	3266.7	3270.1	3279.3	3283.5

Table 4.4: Scaled deviance for different parameter values.

a	b / c	-0.02	-0.01	0.0001	0.1	0.2	0.4	0.5
0.2	-0.02	3255.3	3255.2	3255.2	3257.1	3261.6	3272.8	3278.4
	-0.01	3255.3	3255.3	3255.4	3257.2	3261.5	3272.8	3278.4
	0.0001	3255.5	3255.4	3255.5	3257.3	3261.6	3272.9	3278.4
	0.1	3256.6	3256.6	3256.6	3257.9	3261.9	3272.9	3278.4
	0.2	3258.0	3258.0	3257.8	3258.8	3262.4	3273.0	3278.5
	0.4	3260.2	3260.6	3260.5	3261.0	3263.5	3273.3	3278.4
	0.5	3261.7	3261.8	3261.6	3261.7	3264.3	3273.7	3278.7
0.4	-0.02	3258.4	3258.3	3258.3	3257.9	3259.0	3265.9	3270.7
	-0.01	3258.4	3258.3	3258.3	3257.9	3259.0	3265.8	3270.7
	0.0001	3258.5	3258.6	3258.3	3257.9	3259.0	3265.8	3270.7
	0.1	3259.0	3258.7	3258.7	3258.1	3259.1	3265.7	3270.5
	0.2	3259.4	3259.2	3259.1	3258.4	3259.1	3265.6	3270.4
	0.4	3261.2	3260.1	3260.1	3258.9	3259.7	3265.2	3270.3
	0.5	3260.7	3261.8	3260.6	3259.5	3259.5	3265.2	3270.2
0.5	-0.02	3262.9	3262.9	3262.8	3262.3	3262.4	3266.2	3269.6
	-0.01	3263.0	3262.8	3262.9	3262.2	3262.5	3266.1	3269.6
	0.0001	3263.1	3262.9	3262.9	3262.2	3262.5	3266.2	3269.6
	0.1	3263.3	3263.1	3263.1	3262.3	3262.5	3266.1	3269.5
	0.2	3263.5	3263.3	3263.3	3262.4	3262.5	3266.0	3269.4
	0.4	3263.9	3263.6	3263.7	3262.5	3262.7	3265.8	3269.0
	0.5	3264.1	3263.8	3263.7	3261.7	3262.8	3265.2	3269.2

Table 4.4: Continued for different parameter values.

The minimum scaled deviance (3255.2) occurs at about $a = 0.2$, $b = 0.0001$ and $c = -0.02$. The contour plots for different fixed values of the parameter 'a' given in the above table and for $-0.02 \leq b, c \leq 0.02$ are plotted in Fig. 4.2 and Fig. 4.3.

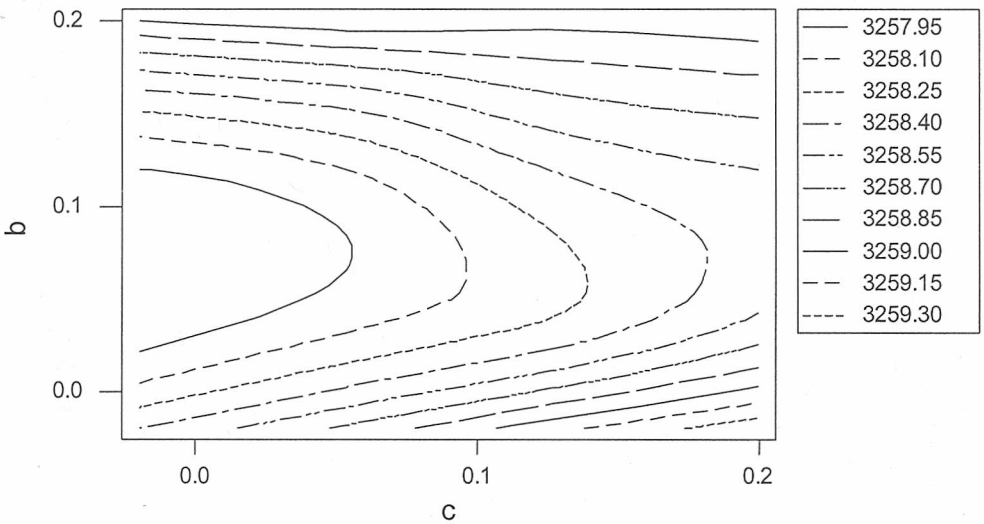

Fig 4.2: Contour plots of scaled deviance for a=0.4

The above contour plot for $a = 0.4$ shows that the minimum for $a = 0.4$ appears to occur at about $b = 0.08$ and $c = -0.02$. This minimum appears to be at the boundary of the region of convergence. Or it appears overall we might reasonably take $a = b = c = 0$, *i.e.* the logit link function in all cases

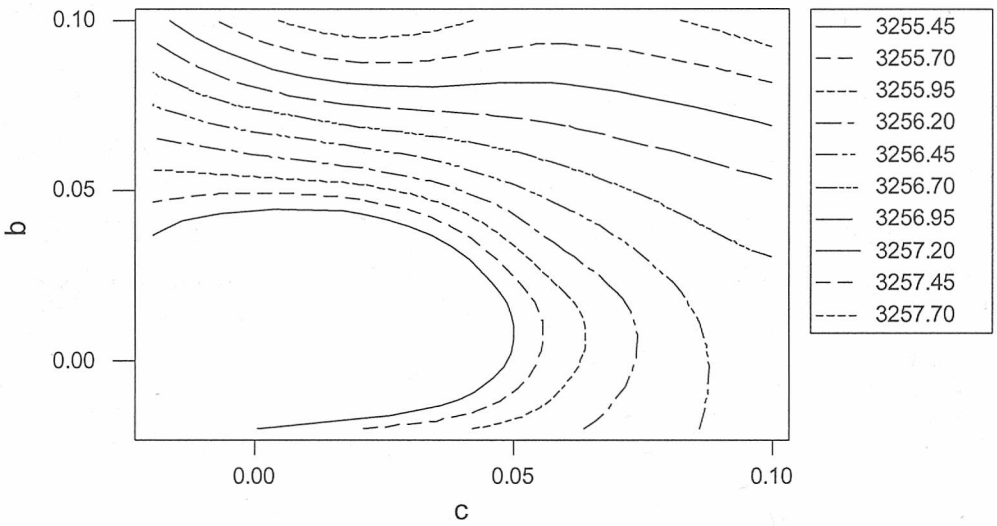

Fig 4.3: Contour plots of scaled deviance for a=0.2

Similarly as Fig 4.2, we see from the above contour plot that the minimum for the parameter $a = 0.2$ appears to occur at about $b = -0.02$ and $c = -0.02$. The minimum appears to be at the boundary of the region of convergence. The contour plots for all other parameter values in Table 4.4 are drawn and no obvious shape or pattern of convergence is found as in Fig.4.2 or Fig. 4.3.

4.7 Testing link function

We explain here the data analytic procedure of Pregibon (1980) to examine the adequacy of the Box - Cox link function used in our approach for fitting the multinomial regression model in section 4.6.1. We attempt to aid the reader by outlining the procedures for the Box - Cox link function specification only.

We fit the multinomial model with the hypothesized link function $\eta_{i2} = g(\theta_{i2})$ when in fact, the correct link function is $\eta_{i2}^{\bullet} = g(\theta_{i2}, a)$ for some unknown parameter a but in general, there could be more than one unknown parameters.

Now using a first - order Taylor series expansion about the hypothesized link function, we have the approximate relationship as

$$\eta_{i2}^{\bullet} = g(\theta_{i2}, a=0) + (a-0)\frac{\partial}{\partial a} g(\theta_{i2}) \Big|_{a=0} + - - - \qquad (4.24)$$

$$= \eta_{i2} + a \frac{\partial}{\partial a} g(\theta_{i2}) \Big|_{a=0} + - - - \quad .$$

For Box - Cox link $a \dfrac{\partial g}{\partial a} \Big|_{a=0} = a \dfrac{\partial}{\partial a} \left(\dfrac{e^{a\theta_{i2}} - 1}{a} \right) \Bigg|_{a=0}$

$$= a \frac{\partial}{\partial a} \left(\frac{1 + a\theta_{i2} + a^2\theta_{i2}^2/2! + a^3\theta_{i2}^3/3! + - - - \quad -1}{a} \right)_{a=0} = \frac{a\,\theta_{i2}^2}{2}$$

Hence the link modification in our case for section 4.6.1 be reformulated easily in terms of original hypothesized link function with the additional factors as follows

$$\eta_{i2}^{\bullet} = \eta_{i2} + \frac{a\,\theta_{i2}^2}{2} + - - -$$

Similarly $\eta_{i3}^{\bullet} = \eta_{i3} + \dfrac{b\,\theta_{i3}^2}{2} + - - - ,$ (4.25)

$$\eta_{i4}^{\bullet} = \eta_{i4} + \frac{c\,\theta_{i4}^{2}}{2} + - - - .$$

Here $\eta_{i2} = \log\theta_{i2}$, $\eta_{i3} = \log\theta_{i3}$ and $\eta_{i4} = \log\theta_{i4}$

To test the adequacy of the link function we fit the multinomial model with the logit link and than fit an 'auxiliary variable' $\frac{\theta_{i2}^2}{2}$ (or $\frac{\theta_{i3}^2}{2}$, $\frac{\theta_{i3}^2}{2}$) to give us an estimate of the unknown parameter a (or b c).

A significant reduction in the resulting deviance will indicate the departure from the original link function or this corresponds to the fact that the wrong link function is a systematic mis - specification of fitting the model.

CHAPTER 5

Improving convergence by Re-parameterisation

5.1 Introduction

In this study for fitting the multinomial logit model with the methodology of chapter 2, the data can be collected and stored for each individual, rather than being grouped. The response variables will then have only yes or no type responses (typically, a row of zeros, with just one 1.0), as for the data attached with the compact disk. For this type of data convergence of the model fitting can take a long time and many iterations by our method. (In some of the examples given in this study, convergence took more then five hundred iterations). To minimize these iterations some changes in the design matrix are given. These changes to the design matrix make the convergence very sharp, as in some cases we only need few iterations to obtain convergence (to the same scaled deviance as we can get from the Poisson trick approach). More details of these changes in the design matrix is given in section (5.2). Our initial changes in the design matrix were derived from the idea of spectral decomposition of the covariance matrix of the parameter estimates obtained from fitting the multinomial logit model. The spectral decomposition of this

covariance matrix yields the principal components of the response variable as orthogonal variables. The multinomial logit model on these orthogonal response variables converges very fast and needs only few iterations. To illustrate this idea, we may note that, in section 5.3 (a), the covariance matrix can be extracted as follows:

$$\Sigma = \begin{pmatrix} 0.0416 & -0.0178 & 0.0155 & -0.0067 \\ -0.0178 & 0.0091 & -0.0067 & 0.0034 \\ 0.0155 & -0.0067 & 0.0404 & -0.0174 \\ -0.0067 & 0.0034 & -0.0174 & 0.0090 \end{pmatrix}.$$

The diagonalized form of Σ uses the spectral decomposition *i.e.* $\Sigma = \Gamma\Lambda\Gamma'$, where

Λ = diagonal matrix of eigenvalues of Σ.

Γ = orthogonal matrix whose columns are standardized eigenvectors.

$$\Rightarrow \Gamma\Gamma' = I \qquad \text{or} \qquad \Gamma' = \Gamma^{-1}.$$

$$\Rightarrow \Gamma'\Sigma\Gamma = \Lambda . \tag{5.1}$$

and from equation (5.1) we say that Γ' diagonalizes Σ. The spectral decomposition of Σ yields the principal components and using the notation in our macros we can get transformed explanatory variables as

$$vd1 = -0.288d1 - 0.653d2 + 0.284dp1 + 0.64dp2$$

$$vd2 = -0.282d1 - 0.641d2 - 0.291dp1 - 0.621dp2$$

$$vd3 = 0.636d1 - 0.279d2 - 0.657dp1 + 0.292dp2$$

$$vd4 = -0.658d1 + 0.292d2 - 0.635dp1 + 0.282dp2.$$

Hence, if we write $d1 - dp1 = u1$, $d2 - dp2 = u3$ and $d1 + dp1 = u2$, $d2 + dp2 = u4$, then we can rewrite the transformed explanatory variables as approximately

vd1= - u1 – 2u3, vd2= - u2 – 2u4

vd3= 2u1 – u3, vd4= - 2u2 + u4

The two alternative sets of variables u1, u2, u3, u4 or vd1, vd2, vd3, vd4 can be fitted instead of the original variables *d1*, *dp1*, *d2*, *dp2*, and both sets improve convergence. Thus, when fitting the multinomial logit model with the above transformed explanatory variables convergence is achieved quickly, only needing a few iterations to get the same scaled deviance as can be obtained from using Poisson trick approach. This transformed variate approach requires at least approximate knowledge of the covariance matrix, which might be obtained from the least squares estimates of the parameters. This might be achieved by using estimates from a restricted number of iterations; perhaps from just one iteration. However, our initial idea has led us to consider a further different approach; in particular, we now consider an approach based upon a simple alternative form of the design matrix.

5.2 Alternative form of design matrix

In chapter 2; we find β_{new} in equation (2.11), where $\boldsymbol{D}$ is a matrix of derivatives which is defined as

$$\boldsymbol{D} = \begin{pmatrix} X^{(2)} & 0 \\ 0 & X^{(3)} \end{pmatrix}_{2m \times 4}$$

and the design matrix is

$$N = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} D = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} \\ 0 & A_{22}X^{(3)} \end{pmatrix}_{2n\times 4}.$$

If we have, for example, only one covariate x and $m = 4$, then we can rewrite D as follows;

(a) $$D = \begin{pmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ 1 & x_3 & 0 & 0 \\ 1 & x_4 & 0 & 0 \\ 0 & 0 & 1 & x_1 \\ 0 & 0 & 1 & x_2 \\ 0 & 0 & 1 & x_3 \\ 0 & 0 & 1 & x_4 \end{pmatrix} \quad \text{with parameters} \quad \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix} \qquad (5.2)$$

and $$N = \begin{pmatrix} A_{11}^{(1)} & 0 & 0 & 0 & A_{12}^{(1)} & 0 & 0 & 0 \\ 0 & A_{11}^{(2)} & 0 & 0 & 0 & A_{12}^{(2)} & 0 & 0 \\ 0 & 0 & A_{11}^{(3)} & 0 & 0 & 0 & A_{12}^{(3)} & 0 \\ 0 & 0 & 0 & A_{11}^{(4)} & 0 & 0 & 0 & A_{12}^{(4)} \\ 0 & 0 & 0 & 0 & A_{22}^{(1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_{22}^{(2)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_{22}^{(3)} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{22}^{(4)} \end{pmatrix} \begin{pmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ 1 & x_3 & 0 & 0 \\ 1 & x_4 & 0 & 0 \\ 0 & 0 & 1 & x_1 \\ 0 & 0 & 1 & x_2 \\ 0 & 0 & 1 & x_3 \\ 0 & 0 & 1 & x_4 \end{pmatrix}$$

or $$N = \begin{pmatrix} A_{11} & A_{11}X & A_{12} & A_{12}X \\ 0 & 0 & A_{22} & A_{22}X \end{pmatrix}_{8\times 4}. \qquad (5.3)$$

In the macros we have given the notation as

$$\underline{d1} \qquad \underline{d2} \qquad \underline{dp1} \qquad \underline{dp2}$$

$$\begin{pmatrix} A_{11} \\ 0 \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{11}X \\ 0 \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix}_{8\times1} \quad \begin{pmatrix} A_{12}X \\ A_{22}X \end{pmatrix}_{8\times1}. \tag{5.4}$$

For a sharper convergence we can consider an alternative parameterisation using an approach similar to that obtained from diagonalizing the covariance matrix. We consider replacing $\boldsymbol{D}_{ij}$ by the following alternative form:

(b) $$\boldsymbol{D} = \begin{pmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ 1 & x_3 & 0 & 0 \\ 1 & x_4 & 0 & 0 \\ 1 & x_1 & 1 & x_1 \\ 1 & x_2 & 1 & x_2 \\ 1 & x_3 & 1 & x_3 \\ 1 & x_4 & 1 & x_4 \end{pmatrix}$$

with parameters $$\begin{pmatrix} \gamma_1 \\ \delta_1 \\ \gamma_2 \\ \delta_2 \end{pmatrix} \tag{5.5}$$

or $$N = \begin{pmatrix} A_{11} + A_{12} & A_{11}X + A_{12}X & A_{12} & A_{12}X \\ A_{22} & A_{22}X & A_{22} & A_{22}X \end{pmatrix}_{8\times 4} . \qquad (5.6)$$

The equivalence in the parameterisations in equation (5.2) and (5.5) is as follows:

$$\alpha_1 = \gamma_1 \qquad \text{or} \qquad \gamma_1 = \alpha_1$$

$$\beta_1 = \delta_1 \qquad\qquad \delta_1 = \beta_1$$

$$\alpha_2 = \gamma_1 + \gamma_2 \qquad\qquad \gamma_2 = \alpha_2 - \alpha_1$$

$$\beta_2 = \delta_1 + \delta_2 \qquad\qquad \delta_2 = \beta_2 - \beta_1 \ .$$

We have given the notation in macros as

d1	*d2*	*dp1*	*dp2*
$\begin{pmatrix} A_{11} + A_{12} \\ A_{22} \end{pmatrix}_{8\times 1}$	$\begin{pmatrix} A_{11}X + A_{12}X \\ A_{22}X \end{pmatrix}_{8\times 1}$	$\begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix}_{8\times 1}$	$\begin{pmatrix} A_{12}X \\ A_{22}X \end{pmatrix}_{8\times 1}$

. (5.7)

Now in the next section we will use the alternative form of the design matrix with the original design matrix in an example of a yes or no type three level response variable. For illustration, we use with one covariate and are able to show that this alternative form of the design matrix gives exactly the same findings as is the original design matrix but with many less iterations required for convergence.

5.3 Example

We consider the data here in this example by Collier et al. (2001) on whether young people intended to enter UK Higher Education (HE). These multinomial data were collected on many variables but for the simplicity and clarity of the ideas presented in this chapter, we consider only 'friends encouragement' as a three level response variable and Age as a explanatory variable. The data is attached in a compact disk. The data consists of 1742 cases

The multinomial logit models are fitted here with the design matrices given in section (5.2) to compare how quickly convergence is achieved when using the original and alternative form of design matrices. The results are also compared with the Poisson trick approach.

(a) We first fit the model $Y = d1 + d2 + dp1 + dp2$ using the original design matrix equation (5.2) and the notation in chapter 2, with macros in the Appendix A.

We need to repeat the GLIM code in the macro $use startup$use loop$ to get the output as follows:

scaled deviance = 1471.4 at cycle 12

residual df = 1738

	estimate	s.e.	parameter
1	0.2262	0.2026	$d1 = \alpha_1$
2	0.3589	0.0951	$d2 = \beta_1$
3	0.3826	0.2034	$dp1 = \alpha_2$
4	0.7774	0.0954	$dp2 = \beta_2$

Table 5.1: Parameter estimates and standard errors for (a)

scale parameter 0.8466

(b) We now fit again the model $Y = d1 + d2 + dp1 + dp2$ but using the alternative form of design matrix equation (5.5), with new parametric notations.

Again, starting up by repeating the GLIM code $use startup$use loop$ until convergence occurred and exactly with same parameter estimates as in (a). In this case, we again need 12 iterations but these results can also be achieved with fewer iterations.

scaled deviance = 1471.4 at cycle 12

residual df = 1738

	estimate	s.e.	parameter
1	0.2262	0.2026	$d1 = \gamma_1$
2	0.3589	0.0951	$d2 = \delta_1$
3	0.1564	0.2265	$dp1 = \gamma_2$
4	0.4184	0.1065	$dp2 = \delta_2$

Table 5.2: Parameter estimates and standard errors for (b)

scale parameter 0.8466

In parametric equivalence for (a) and (b) we can write

(a)		**(b)**
$\alpha_1 = 0.2262$	or	$\gamma_1 = 0.2262$
$\beta_1 = 0.3589$		$\delta_1 = 0.3589$
$\alpha_2 = 0.3826$		$\gamma_2 = 0.3826 - 0.2262$
$\beta_2 = 0.7774$		$\delta_2 = 0.7774 - 0.3589$

(c) For verifications of the parameter estimates and the scaled deviance found in (a) and (b) we fit the same model via Poisson trick approach of Francis et al. (1992) as follows;

```
$c macro for Poisson trick approach of Francis et al. (1992)$
$c level of response variable declared in data file$
$calc vy1=(fr==1):vy2=(fr==2):vy3=(fr==3)$
$calc n1=1$c response variables total$
$num vl$
$calc vl=3*%sl$c 3*standard length$
$ass freq=vy1,vy2,vy3$c freq y-variable$
$ass ag21=ag,ag,ag$c explanatory variable$
$var vl case group$
$calc case =%gl(%sl,1):group =%gl(3,%sl)$
$fact case %sl group 3$
$eliminate case $error p$yvar freq$
$fit+group*ag21(case)$dis e$
```

scaled deviance = 1471.4 (change = -42.80) at cycle 5

residual df = 1738 (change = -2)

	estimate	s.e.	parameter
1	0.2343	0.2830	GROUP(2)=α_1
2	0.4006	0.2593	GROUP(3)=α_2
3	0.3490	0.1579	GROUP(2).AG21=β_1
4	0.7649	0.1444	GROUP(3).AG21=β_2

Table 5.3: Parameter estimates and standard errors for (c)

scale parameter 1.000

From above we see that the parameter estimates and the scaled deviance in (a), (b) and (c) effectively have exactly the same values. Any small differences can be considered as being due to the rounding errors.

5.4 Generalization

We generalize here the alternative design matrix given in section 5.2 (b) for any number of response levels and any number of covariates. This approach will produce exactly the same findings as are given in chapter 3, but with different parameterisations. The equation (3.1) can be rewritten in the form of a new design matrix given in section (5.2) as follows:

$$\boldsymbol{D}_{ij} = \begin{pmatrix} X^{(2)} & 0 & 0 \\ X^{(2)} & X^{(3)} & 0 \\ X^{(2)} & X^{(3)} & X^{(4)} \end{pmatrix} \quad \text{with parameters} \quad \begin{pmatrix} \gamma_1 \\ \delta_1 \\ \gamma_2 \\ \delta_2 \\ \gamma_3 \\ \delta_3 \end{pmatrix} \tag{5.8}$$

The $X^{(k)}$ is a design matrix of explanatory variables for response k. We then have N defined as:

$$N = \begin{pmatrix} A_{11}X^{(2)} + A_{12}X^{(2)} + A_{13}X^{(2)} & A_{12}X^{(3)} + A_{13}X^{(30} & A_{13}X^{(4)} \\ A_{22}X^{(2)} + A_{23}X^{(2)} & A_{22}X^{(3)} + A_{23}X^{(3)} & A_{23}X^{(4)} \\ A_{33}X^{(2)} & A_{33}X^{(3)} & A_{33}X^{(4)} \end{pmatrix}$$

Although it is not necessary we usually consider $X^{(2)} = X^{(3)} = X^{(4)}$. The notational representation in the macros will then be:

d1

$$\begin{pmatrix} A_{11}+A_{12}+A_{13} \\ A_{22}+A_{23} \\ A_{33} \end{pmatrix}_{3n\times 1}$$

dp1

$$\begin{pmatrix} A_{11}X+A_{12}X+A_{13}X \\ A_{22}X+A_{23}X \\ A_{33}X \end{pmatrix}_{3n\times 1}$$

d2

$$\begin{pmatrix} A_{12}+A_{13} \\ A_{22}+A_{23} \\ A_{33} \end{pmatrix}_{3n\times 1}$$

dp2

$$\begin{pmatrix} A_{12}X+A_{13}X \\ A_{22}X+A_{23}X \\ A_{33}X \end{pmatrix}_{3n\times 1}$$

d3

$$\begin{pmatrix} A_{13} \\ A_{23} \\ A_{33} \end{pmatrix}_{3n\times 1}$$

dp3

$$\begin{pmatrix} A_{13}X \\ A_{23}X \\ A_{33}X \end{pmatrix}_{3n\times 1}$$

We can fit the model $Y = d1 + dp1 + d2 + dp2 + d3 + dp3$, which has the same parametric interpretation as is given in section 5.3. It is very easy to see from above that it can be generalized in any number of response levels with any number of covariates. The parametric equivalence in section 5.3 can be extended as follows:

	Original parameterisation		Alternative parameterisation
Four response levels	$\begin{cases} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \alpha_3 \\ \beta 3 \end{cases}$	or	$\begin{cases} \gamma_1 = \alpha_1 \\ \delta = \beta_1 \\ \gamma_2 = \alpha_2 - \alpha 1 \\ \delta_2 = \beta_2 - \beta_1 \\ \gamma_3 = \alpha_3 - \alpha_2 \\ \delta_3 = \beta_3 - \beta_2 \end{cases}$

Thus from above it can be extended further without any problem to any number of levels of response variable.

CHAPTER 6

A More General Example of Fitting A Multinomial Model

6.1 Introduction

In this chapter we will illustrate a more general example to illustrate how the theory and the methods developed in the previous chapters are applicable for any number of response levels, with any number of explanatory variables. For this purpose, dermatology data from Nilsel Ilter, Gazi University, and H. Altay Guvenir, Bilkent University, Ankara, Turkey is used to fit a multinomial regression model with a logit link function, with our own link function with different parameters, and using the improved version of design matrix for the reduction of iterations.

This dermatology data was used in the past by G. Demiroz, H. A. Guvenir and N. Ilter to determine the type of Eryhemato-Squamous Disease. The data contains 34 attributes, 33 of which are continuous and one is nominal.

The differential diagnosis of erythemato - squamous disease is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis and pityriasis rubra pilaris which are to from the six levels our of response variable. Usually a biopsy is necessary for the diagnosis but

unfortunately these diseases share many histopathological features. Another difficulty for the differentiation is that a disease may show the features of another disease at the beginning stage but may have the characteristic at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

Previous studies on this data have considered the selection of a linear combination of the explanatory variables. Thus, here, the six variables parakeratosis, spongiosis, PNL infiltrate, koebner phenomenon, follicular horn plug and focal hypergranulosis are considered as explanatory variables, with levels of a response variable. The Class Code (Disease) for four levels of the response variable can be described as follows:

Class Code	Class	Number of instances
1	psoriasis	112
2	seboreic dermatitis	
+	cronic dermatitis	113
3	lichen planus	72
4	pityriasis rosea	
+	pityriasis rubra pilaris	69
	Total	366

The clinical and histopathological attributes in the data take values 0, 1, 2, and 3 unless otherwise indicated. More information about the levels of response variable and explanatory variables can be found in the attached compact disk.

6.2 Selecting an appropriate model using the Poisson trick

The Poisson trick approach is used here to fit a model for four levels of response variable and with six explanatory variables. The backward selection procedure is adopted to find the significant explanatory variables in the model by deleting one explanatory variable from the main effect model, and comparing the fit of each reduced model. It was found each time that the deleted explanatory variable has a significant effect. Therefore, the main effects model with all six explanatory variables is considered as an appropriate model in this case for the dermatology data.

The following macros were written to fit the main effects model using the Poisson trick approach of Francis et al (1992). It does not converge using the default number of cycles. The number of cycles was increased to obtain convergence. It is noted that the parameter estimates change when the number of cycles are changed to achieve the convergence.

```
$c macros for the Poisson trick model for the dermatology data and all
 other notations can be found in the attached compact disk$
$cal n1=1$
$cal cod1=(cod==1)+2*(cod==2)+2*(cod==5)
              +3*(cod==3)+4*(cod==4)+4*(cod==6)$
$calc  vy1=(cod1==1) : vy2=(cod1==2)$
$calc  vy3=(cod1==3) : vy4=(cod1==4)$
$num vl$
$calc vl=4*%sl$
$ass freq=vy1,vy2,vy3,vy4$
$var vl case group$
$calc case=%gl(%sl,1) :group=%gl(4,%sl)$
$fact case %sl group 4$
$eliminate case$error p$yvar freq$
$c  number of cycle and the convergence criteria$
$cycle 20 2 1.0e-4$
$fit group +group*koe(case)+group*pnl(case)+group*par(case)+
      group*foc(case)+group*spo(case)+group*fol(case)$
```

The above macros are executed and the statistics obtained are given here; these can be compared further in the following sections with the theory and the methods developed in this thesis.

Using number of cycles and convergence criteria as $cycle 20 2 1.0e-4$, we get the parameter estimates and standard errors as follows

scaled deviance =148.57(change = -846.7) at cycle 17

residual df =1077

	estimate	s.e.	parameter
1	5.707	1.401	GROUP(2)
2	-17.56	**47.89**	GROUP(3)
3	3.297	1.468	GROUP(4)
4	-4.006	0.9385	GROUP(2).KOE
5	28.40	**44.35**	GROUP(3).KOE
6	0.4075	0.6033	GROUP(4).KOE
7	-1.944	0.5390	GROUP(2).PNL
8	-10.83	74.86	GROUP(3).PNL
9	-3.091	0.7215	GROUP(4).PNL
10	-3.245	0.7116	GROUP(2).PAR
11	-60.46	**87.11**	GROUP(3).PAR
12	-3.495	0.7732	GROUP(4).PAR
13	-29.31	**190.0**	GROUP(2).FOC
14	149.5	**251.1**	GROUP(3).FOC
15	-24.09	**172.5**	GROUP(4).FOC
16	15.27	**113.7**	GROUP(2).SPO
17	1.455	**116.5**	GROUP(3).SPO
18	15.75	**113.7**	GROUP(4).SPO
19	14.73	**343.6**	GROUP(2).FOL
20	-25.13	**485.7**	GROUP(3).FOL
21	18.76	**343.6**	GROUP(4).FOL

Table 6.1: Poisson trick model of 4 -level of response & six explanatory variable

scale parameter 1.000

The above statistics have some unstable parameter estimates with very high standard errors. This leads us to explore more about the data and to investigate why we are getting some unusual parameter estimates. This may be because there are a great number of observed counts are zeros and ones. This suggests that the Poisson trick method may be less suitable for this type of data when convergence is not achieved on the default number of cycles. We shall observe in the next section that the multinomial regression method works well and produces the parameter estimates with what appears to be acceptable standard errors.

6.3 Using our approach model

We fit here the same multinomial model using our approach as is selected in section (6.2) and gets the parameter estimates with some apparently acceptable standard errors. The macro converges but needs to increase the number of cycles for the convergence criteria. The macros are given in appendix A, leading to the following output:

```
Scaled deviance = 149.68 at cycle 35
residual df = 1077
```

	estimate	s.e.	parameter(Francis)	
1	5.705	1.186	*d1*	(Group2)
2	-4.484	2.456	*d2*	(Group3)
3	3.290	1.264	*d3*	(Group4)
4	-4.003	0.919	*dp1*	(GROUP2.KOE)
5	8.630	0.954	*dp2*	(GROUP3.KOE)
6	0.409	0.571	*dp3*	(GROUP4.KOE)
7	-1.944	0.463	*dp4*	(GROUP2.PNL)
8	-4.682	2.266	*dp5*	(GROUP3.PNL)
9	-3.093	0.661	*dp6*	(GROUP4.PNL)
10	-3.244	0.606	*dp7*	(GROUP2.PAR)
11	-19.24	1.471	*dp8*	(GROUP3.PAR)

12	-3.492	0.673	*dp9* (GROUP4.PAR)
13	-8.302	3.261	*dp10* (GROUP2.FOC)
14	48.64	3.395	*dp11* (GROUP3.FOC)
15	-6.916	4.992	*dp12* (GROUP4.FOC)
16	14.21	1.436	*dp13* (GROUP2.SPO)
17	10.97	1.676	*dp14* (GROUP3.SPO)
18	14.68	1.461	*dp15* (GROUP4.SPO)
19	10.21	4.126	*dp16* (GROUP2.FOL)
20	-28.18	6.465	*dp17* (GROUP3.FOL)
21	14.24	4.157	*dp18* (GROUP4.FOL)

Table 6.2: Four levels of response and six explanatory variables

```
scale parameter 1.000
```

The data entered in the macros for fitting the multinomial logit model in section (6.3) are exactly the same form as was entered when using the Poisson trick approach in section (6.2), in order to keep Table 6.1 and Table 6.2 in the same formats. The parameter estimates in Table 6.1 for GROUP3, GROUP3.KOE, GROUP3.PNL, GROUP3.PAR, GROUP2.FOC, GROUP3.FOC, GROUP4.FOC, GROUP2.SPO, GROUP4.SPO, GROUP2.FOL, GROUP3.FOL and GROUP4.FOL do not appear very sensible; they all have unstable standard errors. The fitting of our multinomial logit model in Table 6.2 gives some sensible looking parameter estimates with reasonable standard errors. The standard errors for parameters in the same group have exactly the same values when the macros are not upgraded at each cycle, but we get approximately the correct scaled deviance in this case (The macro is given in Appendix A at page 154). This is because we may not have much knowledge about the start - up values or how we have chosen the initial values in the macros.

6.4 Variance - Covariance matrix not upgraded at each cycle

The variance - covariance matrix in section (6.3) is of a symmetric form for each level of response variable when it is not upgraded at each cycle. Since the GLIM code for macro startup is $calc %fv=0.25$ with four levels of response variable and responses are either zeros or ones only and we have

$$V(\hat{\beta}_{new(i)}) = (N^T N)^{-1} V(Y_i),$$

$$= (D^T W D)^{-1} V(Y_i). \qquad (6.1)$$

The $(\boldsymbol{D^T W D})$ in the standard form of the design matrix X be interpreted as

$$(\boldsymbol{D^T W D}) = \begin{pmatrix} X & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{pmatrix}^T \begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix} \begin{pmatrix} X & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{pmatrix}. \qquad (6.2)$$

In this case we have %fv=0.25 and the weight matrix in equation (6.2) will be as

$$W_{11}(i) = \mu_{i2}(1 - \frac{\mu_{i2}}{n_i}) = W_{22}(i) = \mu_{i3}(1 - \frac{\mu_{i3}}{n_i}) = W_{33}(i) = \mu_{i4}(1 - \frac{\mu_{i4}}{n_i}),$$

$$W_{12}(i) = -\frac{\mu_{i2}\mu_{i3}}{n_i} = W_{13}(i) = -\frac{\mu_{i2}\mu_{i4}}{n_i} = W_{23}(i) = -\frac{\mu_{i3}\mu_{i4}}{n_i}.$$

If we denote $W_{11}(i) = W_{22}(i) = W_{33}(i) = \lambda I_{m\times m}$ and

$W_{12}(i) = W_{13}(i) = W_{23}(i) = \delta I_{m\times m}$, then we have

$$(\boldsymbol{D^T W D}) = \begin{pmatrix} X^T \lambda I X & X^T \delta I X & X^T \delta I X \\ X^T \delta I X & X^T \lambda I X & X^T \delta I X \\ X^T \delta I X & X^T \delta I X & X^T \lambda I X \end{pmatrix}_{3p\times 3p}. \qquad (6.3)$$

The above symmetric positive definite matrix has the diagonal matrices of dimensions $p \times p$ and the diagonal elements of the inverse of these be involved in finding the standard errors of the parameters.

If we consider the improved form of design matrix X that we will use in section (6.6) then the matrix $(D^T W D)$ will be as

$$\begin{pmatrix} X^T \lambda I X & X^T \lambda I X + X^T \delta I X & X^T \lambda I X + 2X^T \delta I X \\ X^T \lambda I X + X^T \delta I X & 2[X^T \lambda I X + X^T \delta I X] & 2[X^T \lambda I X + 2X^T \delta I X] \\ X^T \lambda I X + 2X^T \delta I X & 2[X^T \lambda I X + 2X^T \delta I X] & 3[X^T \lambda I X + 2X^T \delta I X] \end{pmatrix}$$

To find $V(\hat{\beta}_{new(i)})$ in equation (6.1) we need the y - variable that involves special inverse matrices and *(%lp)*. The σ_i^2 is formulated in the macro using macro devcalc. It make some sense that the parameter estimates for each group of response variable have the same standard errors when observed counts are zeros or ones only and macro startup with GLIM code is $calc %fv=0.25$.

6.5 Our own link model

We fit in this section a multinomial model using our own link function as is given in section (4.2.2) with different values of the link parameters. We fitted the parameter values as a=b=c=0.0001, as this is approximately equivalent to the multinomial logit model. The macros are given in Appendix A and we get the statistics as

Scaled deviance = 149.69 at cycle 35

residual df = 1077

	estimate	s.e.	parameter
1	5.705	1.186	*d*1
2	-4.445	2.455	*d*2
3	3.287	1.263	*d*3
4	-4.011	0.919	*dp*1
5	8.598	0.956	*dp*2
6	-0.411	0.571	*dp*3
7	-1.944	0.463	*dp*4
8	-4.687	2.265	*dp*5
9	-3.095	0.661	*dp*6
10	-3.244	0.606	*dp*7
11	-19.19	1.469	*dp*8
12	-3.492	0.672	*dp*9
13	-8.318	3.259	*dp*10
14	48.50	3.392	*dp*11
15	-6.934	4.988	*dp*12
16	14.50	1.436	*dp*13
17	11.20	1.676	*dp*14
18	14.98	1.460	*dp*15
19	10.61	4.125	*dp*16
20	-27.26	6.455	*dp*17
21	14.65	4.155	*dp*18

Table 6.3: Four levels of response and six explanatory variables

scale parameter 1.000

The above Table 6.3 almost produces the same statistics as in Table 6.2 and that we expect in sections (6.3) and (6.5). This table has the same interpretation as in section (6.3) but the minor difference in parameter estimates can reasonably be attributed to the different parameter values in own link functions.

6.6 Multinomial model with improved design matrix

We use here the improved version of design matrix as is given in section (5.2) in order to reduce the iterations to obtain convergence. We only need in the macros to change the design matrix equation (5.3) for the improved version of design matrix equation (5.6). We need to be careful about entering the data and the notations in macros of the parameters in order to preserve the equivalence of results to those of previous sections. The macros for the improved version of design matrix are given in Appendix A with the following statistics being output:

```
Scaled deviance = 149.68 at cycle 400

residual df = 1077
```

	estimate	s.e.	parameter
1	5.705	1.186	*p*1
2	-10.19	2.185	*p*2
3	7.774	2.176	*p*3
4	-4.003	0.991	*pd*1
5	12.63	1.117	*pd*2
6	-8.222	0.800	*pd*3
7	-1.944	0.463	*pd*4
8	-2.738	2.235	*pd*5
9	1.589	2.212	*pd*6
10	-3.244	0.606	*pd*7
11	-16.00	1.361	*pd*8

12	15.75	1.330	*pd9*
13	-8.302	3.261	*pd10*
14	56.95	3.574	*pd11*
15	-55.56	5.103	*pd12*
16	14.21	1.436	*pd13*
17	-3.232	0.861	*pd14*
18	3.708	0.839	*pd15*
19	10.21	4.126	*pd16*
20	-38.39	5.038	*pd17*
21	42.42	4.965	*pd18*

Table 6.4: Multinomial logit model for improved version of design matrix

scale parameter 1.000

There is no change in the scaled deviance as in section (6.3) with the interperation of parameter estimates and standard errors. The parametric equivalence in the Table 6.4 and Table 6.3 can be seen as,

p1 = *d1*

p2 = *d2-d1*

p3 = *d3-d2*

pd1 = *dp1*

pd2 = *dp2-dp1*

pd3 = *dp3-dp2*

pd4 = *dp4*

pd5 = *dp5-dp4*

pd6 = *dp6-dp5*

pd7 = *dp7*

pd8 = *dp8-dp7*

pd9 = *dp9-dp8*

```
pd10  =   dp10
pd11  =   dp11-dp10
pd12  =   dp12-dp11
pd13  =   dp13
pd14  =   dp14-dp13
pd15  =   dp15-dp14
pd16  =   dp16
pd17  =   dp17-dp16
pd18  =   dp18-dp17
```

The above parametric equivalence can be interpreted in the reverse order for the interpretation of the standard multinomial logit model with or without own link function for different parameters. It is noted here unusual for this data that we need 400 iterations for the improved design matrix or we may need to consider different startup values.

6.7 Comments for fitting a appropriate model

The following considerations are needed for the dermatology data used in fitting a more general multinomial regression model. This may not be the case when the data counts have not so many zeros and ones. These considerations become more important when the convergence does not occur with the default number of cycles and we need to increase the number of cycles or to relax the convergence criteria.

a) The Poisson trick approach of Francis, et al. (1992) can give large s.e.'s so may be not appropriate one for multinomial counts with zeros and ones when convergence is not achieved with default cycles only. It would seem that it can give unstable parameter estimates with very high values of standard errors.

b) The multinomial regression model method given in this thesis works very well for the multinomial counts with zeros and ones but for the convergence criteria some extra care is needed with respect to the number of cycles.

c) It is noted here that the use of improved version of design matrix in the multinomial counts with zeros or ones for the larger data set may not reduce the iterations but this is not the case in real counts for level of response variable.

d) The standard errors of parameter estimates appear as blocks of equal (incorrect) values for our approach to fitting multinomial counts with zeros or ones if the GLIM code for updating the weight matrix at each cycle is omitted. The initial startup values for %fv plays an integral part in determining the 'standard errors' in these blocks. It also depends how we are entering the data in macros for the model fitting.

e) The form of the design matrix in the macros for model fitting does not essentially affect the multinomial regression model theory as given in this study, but you do have to identify the parameter estimates. It does not have any effect on the scaled deviance or the degree of freedom.

CHAPTER 7

Likelihood Influence Measures Using Cook's Distance (1986)

7.1 Introduction

The regression diagnostics for normal linear models are well established in the literature since early 1960s. The principal ideas of assessment of influence may be found in Cook and Weisberg (1982), Belsley et al. (1980) and are surveyed comprehensively in Chatterjee and Hadi (1988), especially in chapter 4 and 5, a general approach for sensitive analysis in the linear regression models, there is detailed discussion about assessing the effects of observations.

An assessment of local influence in regression models using the likelihood displacement influence measures and the ideas of global influence are found in Cook (1986). These ideas of influence measures Cook (1986) are introduce here for the multinomial data using case i, $i = 1, 2, \ldots, m$ deletion method.

The general reference on empirical and theoretical influence functions are found in Hample et al. (1986) and some work by Pregibon (1981) gives notions about the influence function in

linear logistic regression models. Thomas and Cook (1989) gives an assessment of influence on regression coefficient in generalized linear models and also Thomas and Cook (1990) using local influence methods, assesses the effect of perturbations of data on prediction from generalized linear models.

Preisser and Qaqish (1996) proposed the deletion diagnostic for generalised estimating equations. They consider the leverage and residuals to measure the influence of a subset of observations on the $\hat{\beta}$ and on the estimated values on the linear predictor.

Andrews and Pregibon (1978), Atkison (1982) and Geisser (1982) proposes the diagnostics based on observation (case) deletion schemes. Moolgavkar et al. (1984) used cases deletion for non - linear regression diagnostics with application to matched case - control studies, and Storer and Crowley (1985) give diagnostics for the parameter estimate $\hat{\beta}$ using the changes in the maximum likelihood due to deletion in logistic and linear regression models.

The diagnostics for examining the influence on confidence intervals or on confidence regions for regression coefficients is given by Thomas (1990), Pena and Yohai (1993) and Chatterjee and Hadi (1988) analyse the eigenvectors corresponding to the non - zero eigenvalues of the hat - matrix. Ellis and Morgenthaler (1992) used the diagonal elements of the hat - matrix as the diagnostic indicators.

Davis and Snell (1991) give a discussion about residuals and diagnostics and Davis and Tasi (1992) a critique about regression diagnostics, including residuals and influence but these may not have any use in the multinomial data used during this research.

The regression analysis in the context of *masking* and *boosting* is given by Lawrance (1995) and classifies the possible effects on pairs of cases on Cook's distance (1977). Hjort (1992) considers the diagnostics in the parametric survival data studies. Lustbader and Moolgavkar (1985) gives the diagnostics statistic for hypothesis testing by observation change in the score statistic after deletion of observation.

Critchley (1985) develops the influence function for the detection of influential observations in the principal component analysis by applying the perturbation on the symmetric matrices. Critchley and Vitiello (1991) discussed pairs of cases, as well as single cases for use of case weight perturbation.

Many of these diagnostics for influence measures use statistics that measure the effect of deleting a single observation from the data and more details of deletion a single case i, $i = 1, 2, \ldots, m$ or deletion of cell (i, j) can be found in section (1.8). These statistics exploit the exact algebraic relationship between the least squares fit of the linear model to the complete set of m cases, and the fit to the $m-1$ cases remaining after the deletion of a single case i. The maximum likelihood (ML) estimation of generalized linear models (GLMs) requires iterative methods. The maximum likelihood estimates from $m-1$ cases cannot then be obtained as explicit functions of the results of the fit to all m cases.

The approach we used here in this study to fit a multinomial logit model is some sort of least squares fit and it is very easy to extract Cook's distances (1977), leverages and some other appropriate statistics for each cell (i, j) in the multinomial data but these are of very little use for the influence effect or for diagnostic of a case i on the parameter estimate $\hat{\beta}$.

We find here the likelihood displacement influence measure of Cook (1986) given in the equation (1.29) for fitting a multinomial logit model for deleting a single case i, $i = 1, 2, \ldots, m$. These may be interpreted to detect the influence of a case i for the parameter estimate $\hat{\beta}$ only. More detail on likelihood influence measures of Cook (1986) is given in section (1.13) for deleting a single case i (note, for convenience of exposition, we there presented the theory about likelihood displacement in terms of a univariate response considering deletion of a single 'observation').

In section (7.2) we give some notations and the log - likelihood for the multinomial regression models. These can be used to find the likelihood influence measure of Cook (1986) to detect unusual cases in a multinomial data. In section (7.3) we consider two examples for finding the Cook's distance (1977), leverages, Pearson residuals and the likelihood influence measure of Cook (1986). In section (7.4) we discuss the joint and multiple influence measure of Cook (1986). In section (7.5) we reuse the data of example 7.3.1 using the Poisson trick approach of Francis et al. (1992) to find the likelihood influence measures of Cook (1986) for a single case deletion; this will give the same result as section (7.3). In section (7.6) the key findings and general difficulties in the application for a larger data are presented with some appropriate suggestions.

7.2 Notations and the log - likelihood statistic

Repeating for clarity from section (1.3), we have for the multinomial model

$$L_i \propto p_{i1}^{y_{i1}} \cdot p_{i2}^{y_{i2}} \cdot p_{i3}^{y_{i3}}$$

$$l_i = y_{i1} \log p_{i1} + y_{i2} \log p_{i2} + y_{i3} \log p_{i3} + \text{constant}$$

$$= y_{i2} \log \mu_{i2} + y_{i3} \log \mu_{i3} + (n_i - y_{i2} - y_{i3}) \log (n_i - \mu_{i2} - \mu_{i3}) + \text{constant}$$

and for GLIM code the log - likelihood omitting a constant is given in the macros as

$$l_i = \$\text{calc devc} = \%\text{cu(yv*\%log(fv))}\$. \qquad (7.1)$$

We use the bracketed subscripts to denote the deletion case i when fitted the multinomial model of $m-1$ cases or, equivalently, weighting out the case i. The calculation $LD(\hat{\beta}_{(i)})$ of equation (1.29) for the multinomial model is given in section (7.3) exactly for each i and is computationally expensive, requiring $m+1$ fits of the model. Unthinking application of asymptotic likelihood theory suggests the statistic $2[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$ is distributed asymptotically as $\chi^2(\alpha;p)$ as $n \to \infty$ (as mentioned in equation (1.30)) where p are the number of independent parameters.

When the linear predictor $\eta = X\beta$ is tested for the outlier, when the identity of the potential outlier is unknown or the statistic $LD(\hat{\beta}_{(i)})$ is maximum over i, $i = 1, 2, \ldots, m$. The influence of case i can be measured by the change in the value of the likelihood ratio statistics when case i is deleted.

My own experience of such comparisons over a number of data during this research suggests that those cases whose deletion has a substantial effect on the likelihood influence measure of

Cook (1986) or the significant change in the $LD(\hat{\beta}_{(i)})$ over case i, $i = 1, 2, \ldots, m$ leads us to consider the status of the case i in question.

The Cook's distance (1977) in our study measures the influence of a cell (i, j) on all the parameter estimates, but the main aim of fitting a GLM for the multinomial data may be the influence of case i that can be measured by the change in the likelihood ratio statistic when case i is deleted, and thereby identify those cases which most influence the likelihood ratio statistic.

In normal linear regression the influence of an observation y_{ij} on the parameter estimate $\hat{\beta}$ can be measured by Cook's distance (1977), which calibrates $\hat{\beta} - \hat{\beta}_{(ij)}$ by comparison to the confidence contours for β. An equivalent influence measure is the likelihood influence measure of Cook (1986), $LD(\hat{\beta}_{(i)}) = 2p^{-1}[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$ which is approximated by equation (1.29) and is calculated for case i in the examples given in section (7.3).

It may be noted that we may further be interested to assess the influence of an observation of particular interest both before and after the deletion of another observation; the latter is called conditional influence by Lawrence (1995). When the individual influence of an observation of interest is found to be more after the deletion of the other case, it is said to have been *masked* by the other observation and when it is less, this opposite effect is described as *boosting*. This is basically a conditionally influence measure of observation i after the deletion of observation j and it is straightforward to find the *masking* or *boosting* for a case i in the multinomial data from the idea to find Likelihood influence measure of Cook (1986) and we have described briefly in section (7.4).

7.3 Examples

We consider here two sets of data for fitting a multinomial logit model and calculate the likelihood influence measure of Cook (1986) that may lead us to determine the influential effect of a case i on the parameter estimate $\hat{\beta}$. We also extracted the Cook's distance (1977), leverages and Pearson residuals to see any effect of each cell (i, j) on a multinomial data.

7.3.1 Example 1

To create an illustrative example, we split the Quantal assay data taken from the classic Table V of Irwin (1937) into three response levels, to fit a multinomial logit model. These data were chosen for couple of reasons. Firstly they have been reanalysed by Copenhaver and Mielke (1977), Morgan (1985) and Williams (1987) and they are known to contain some features of interest. Secondly, they comprise only five observations or cases, and a variety of single case diagnostics can be completely tabulated and compared without taking much space. These data are intended to serve only as a numerical example. They cannot convincingly demonstrate the particular utility of single case deletions whose accuracy and computational advantages in the theory presented in this study increases as the number of cases increases.

The following artificial example of data is constructed from Table V of Irwin (1937) by the addition of an artificial extra response levels Y_1 , Y_2 and Y_3 of numbers of mice responding at dose X as

Explanatory Variable	Responses			Total
X	Y_1	Y_2	Y_3	
1	0	0	40	40
2	2	0	38	40
3	9	5	26	40
4	13	6	21	40
5	20	10	10	40

Table 7.1: Modified version of Quantal assay data Table V Irwin (1937)

We consider the level of response variable Y_1 , Y_2 and the dose X as an explanatory variable to fit a multinomial logit model. The data are arranged into four columns as in examples of previous chapters, 1st two columns of response levels, 3rd column of explanatory variable dose and the 4th column of the totals of the response levels. The standard length of the data will be 10 and the macro to fit a multinomial logit main effect model is given in the Appendix A. We extracted the standard Cook's distance (1977), leverages and Pearson residuals for the cell (i, j) as in Table 7.1, $i = 1, 2, \ldots, 5$ and $j = 1, 2, 3$. The leverages extracted here follows all the properties given in section (1.6) and the Cook's distance (1977) or the residuals appears appropriate for the cells (i, j). They can be tabulated here as follows

	Cooks distance	Leverages	Pearson Residuals
1	0.0000	0.0000	-0.7620
2	0.0359	0.3950	-0.3647
3	**3.5650**	0.5235	2.4873
4	0.1231	0.3395	-0.7955
5	1.5700	0.8026	-0.5521
6	0.0000	0.0000	-0.2745
7	0.0000	0.0000	-0.9300
8	**23.020**	0.7952	2.2034
9	0.0421	0.3372	-0.4683
10	**1.4750**	0.8070	-0.5217

Table 7.2: Cook's distance (1977), leverages and Pearson residuals

We examine the above Table 7.2 and clearly see that the Cook's distance for response level Y_1 , Y_2 in Table 7.1 for cell $(1, 3)$ and $(2, 3)$ are quite large, $\sum_{i=1}^{n} h_i = p = 4$ and the leverages for cells $(1, 3)$, $(1, 5)$, $(2, 3)$, $(2, 5)$ are large. The Pearson residuals is large when indicate as the Cook's distance is large but this is not enough for us to form any

obvious conclusion that if there is any influential observation for the modified version of the quantal assay data Table 7.1. We may obtain some idea about the validity of values in cells $(1,3)$, $(2,3)$, $(1,5)$ or $(2,5)$ but not for the case $i=3,5$. To know about any influential case i we calculate the likelihood influence measure of Cook (1986) using equation (7.1) both for the full data and after deleting the single case i *i.e.* weighting out the case $i=1,2,...\ 5$ in the data. For example for case $i=1$, we need to delete the cells (1, 1) and (1, 2) and similarly for $i=2$ the cells (2,1) and (2,2) and so on for other values of i . The macros are given in Appendix A.

We executed the macros and using equation (7.1) we found $2p^{-1}l(\hat{\beta}) = 305.0$ then we weighted out each case for $i=1,2,\ldots,5$ respectively. The likelihood influence measure of Cook (1986), $2\ p^{-1}[\,l(\hat{\beta}) - l(\hat{\beta}_{(i)}]$, can be tabulated as follows:

i	$2p^{-1}[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$
1	0.10
2	0.20
3	1.20
4	0.00
5	0.20

Table 7.3: Likelihood influence measure of Cook (1986) after weighting out observations $i=1,2,\ldots,5$

Looking for large values of likelihood influence measure of Cook (1986) the search for influential cases can be carried out; unfortunately, no clear rules can be given for what constitutes a large of the likelihood distance. Nevertheless, provided that the sample size is not too large, a substantial change in the likelihood influence measure leads us to consider the status of the case in questions. From the above Table 7.3 for case $i = 3$ the likelihood

distance is relatively large which leads us to conclude that this case may have influential effect on the estimation of parameter estimate $\hat{\beta}$. It can be observed from Table 7.2 that the Cook's distance for the response levels in the cells $(1,3)$, $(2,3)$ are quite large that also suggests that the case $i = 3$ is an influential case in Table 7.1.

7.3.2 Example 2

We reuse the artificial data based on Collier, et al. (2003) from a survey of young people as given in Table 2.5, for three response levels and for 'Age' as the explanatory factor.

Explanatory Variables		Response Variables			Total
i	Age	Y_1	Y_2	Y_3	
1	1	6	9	5	20
2	2	5	4	1	10
3	1	1	3	11	15
4	2	6	9	6	21

Table 7.4: Data of Table 2.5 with only 'Age' as explanatory factor.

The above data is arranged as usual to fit a multinomial logit model using the approach presented in this study. The standard length for this data in this format is 8. The macros for fitting multinomial logit model with 'age' as explanatory factor are given in Appendix A. The Cook's distance (1977), leverages and the Pearson residual are extracted for the cells are as follows

	Cooks distance	Leverages	Pearson Residuals
1	**25.957387**	0.8226	1.9925
2	0.512769	0.3774	1.4516
3	0.318787	0.2019	-2.0056
4	**2.542876**	0.6368	-1.4516
5	**3.353517**	0.5628	2.1347
6	0.002492	0.1792	-0.1935
7	**1.381258**	0.4127	-2.1489
8	0.201966	0.8066	0.1935

Table 7.5: Cook's distance (1977), leverages and Pearson residuals

Examining the above Table 7.5 and the Cook's distances for the response level Y_1 , Y_2 in Table 7.4 for cells $(1,4)$ and $(2,1)$ are large, but for cell $(1,1)$ is very large. A similar pattern exists for the leverages and Pearson residual but we cannot form any conclusion as to whether there is any influential case i. Therefore, we wrote the same macro as in the previous example to find the influence measure of Cook (1986). We executed that macros and using equation (7.1) we find $2p^{-1}l(\hat{\beta}) = 58.87$ then we weighted out each case $i = 1, 2, \ldots, 4$ respectively for each i as in previous example. The likelihood influence measure of Cook (1986) , $2p^{-1}[l(\hat{\beta}) - l(\hat{\beta}_{(i)}]$, can be tabulated as follows

i	$2p^{-1}[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$
1	3.29
2	0.22
3	1.77
4	1.27

Table 7.6: Likelihood influence measure of Cook (1986)

Examining the above Table 7.6 we see that the likelihood measure of Cook (1986) for case $i = 1, 3$ and 4 is large but for case $i = 1$ is quite large. This may lead us to investigate further if case 1 may have some undue influential effect on the parameter estimate $\hat{\beta}$. We may note that Cook's distances in Table 7.5 indicates case 1 might be influential (with Cook's distances of 26.0 and 3.4)

In these examples that Cook's distance is here large in those cells for the response level where we also have quite large likelihood influence measure for the case. Thus both the likelihood measure of Cook (1986) and Cook's distance (1977) for the cells or units leads us to the same direction of finding the influential for the case *i*.

7.4 The joint and multiple Influence measure of Cook (1986)

In many practical data analytic problems, considerations of cases one at a time will provide the analyst with most of the information needed concerning the inference of cases on the fitted model. However, it can happen that a group of cases will be influential, but this influence can go undetected when cases are examined individually. To find the effect of multiple observations is important from the theoretical as well as the practical point of view, there may exist situations in which observations are jointly but not individually influential, or the other way around.

The generalization of likelihood influence measure of Cook (1986) is straightforward for joint and multiple cases problem and we give here for the deletion of pairs of cases and or more generally for a deletion of set of I cases.

The joint influence measure based on the deletion of pairs of cases, can be defined as

$$LD(\hat{\beta}_{(i,j)}) = 2[l(\hat{\beta}) - l(\hat{\beta}_{(i,j)})] \tag{7.2}$$

and for example 7.3.1 we need the joint likelihood measures,

$$2[l(\hat{\beta}) - l(\hat{\beta}_{(1,2)}], \; 2[l(\hat{\beta}) - l(\hat{\beta}_{(1,3)}] \; , \; 2[l(\hat{\beta}) - l(\hat{\beta}_{(1,4)}] \; , \; 2[l(\hat{\beta}) - l(\hat{\beta}_{(1,5)}],$$

$$2[l(\hat{\beta}) - l(\hat{\beta}_{(2,3)}] \; , \; 2[l(\hat{\beta}) - l(\hat{\beta}_{(2,4)}] \; , \; 2[l(\hat{\beta}) - l(\hat{\beta}_{(2,5)}],$$

$$2[l(\hat{\beta}) - l(\hat{\beta}_{(3,4)}] \; , \; 2[l(\hat{\beta}) - l(\hat{\beta}_{(3,5)}],$$

$$2[l(\hat{\beta}) - l(\hat{\beta}_{(4,5)}]$$

to detect if there are any joint influence observations or generally we define the influence measure for a deletion of set of I cases as

$$LD(\hat{\beta}_{(I)}) = 2[l(\hat{\beta}) - l(\hat{\beta}_{(I)})] \; . \tag{7.3}$$

This topic could be a subject of the future study.

7.5 $LD(\hat{\beta}_{(i)})$ using the Poisson trick approach

For completeness of exposition, we find in this section the likelihood influence measure of Cook (1986) for a single case deletion when the Poisson trick approach is used for example 1 in section (7.3.1).

Using the Poisson trick approach, we weighted out the case i, $i=1,2,\ldots,5$ and fitted the model. We find the maximum likelihood $l(\hat{\beta}_{(i)})$ for the weighted out model (macros to find $l(\hat{\beta})$ and $l(\hat{\beta}_{(i)})$ are given in Appendix A). The likelihood influence measure of Cook (1986) $2\ p^{-1}[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$ can be tabulated as follows

i	$2p^{-1}[l(\hat{\beta}) - l(\hat{\beta}_{(i)})]$
1	0.10
2	0.20
3	1.20
4	0.00
5	0.20

Table 7.7: Likelihood influence measure of Cook (1986).

The above Table 7.7 and Table 7.3 gives exactly the same statistics but calculated using two different methods. Our approach is more flexible than used in Poisson trick approach of Francis et al. (1992) for calculating likelihood influence measure of Cook (1986) and is easy to use. Our approach mainly requires the reliable start - up or initial values.

7.6 One - Step deletion diagnostics for $\hat{\beta}$

Preisser J. S. and Qaqish B. F. (1996) have introduced one - step deletion diagnostics for generalised estimating equations for the case of independent or clustered observations. In theorem 1 of their paper, they gave a general computational formula for $DEBTA_{m}$, the approximate for $\hat{\beta} - \hat{\beta}_{(m)}$, where $\hat{\beta}$ is the fully iterated p - dimensional generalised estimating equations estimator of the marginal mean regression parameter and $\hat{\beta}_{(m)}$ is the fully iterated estimator after deletion of an arbitrary subset, denoted by m, of the observations.

Preisser J. S. and Qaqish B. F. (1996) also gave a special case that corresponds to the deletion, and influence of a single observation, as the weight matrix W becomes a diagonal matrix, so the result reduces to the one - step approximation for generalised linear models with all cluster sizes equal to 1 and for example, the influence of the *ith* observation defined by $\hat{\beta} - \hat{\beta}_{(i)}$, is approximated as

$$\hat{\beta} - \hat{\beta}_{(i)} \cong (X^T W X)^{-1} X_i^T W_i^{1/2} (1 - h_i)^{-1/2} r_i^{PS}, \qquad (7.4)$$

where W_i is a **scalar**. In our approach, we stack the data by the response level and we do not have a diagonal weight matrix W and in our block - diagonal weight matrix the *jth* row corresponds to the *jth* level of the response variable, so a modified formulation of equation (7.4) would be required using our approach.

Preisser et al. (2008) discuss how and the results provided in their paper (1996) can be extended to a broad class of regression models and their estimating equations that employ iterated least squares as a fitting algorithm and the proof of the theorem do not require the block - diagonal structure of the weight matrix and the derivation is based solely on the iteratively reweighted least square algorithm. The computational algorithm required to find the corresponding deletion measure to equation (7.4) involves iterative application of the Sherman - Morrison - Woodbury formula for the inversion of the matrices with special structure that occur in formulae for deletion diagnostics, and the use of Cholesky's decomposition are illustrated.

The changes of the influence effect in the individual $\hat{\beta}_j$'s in equation (7.4) for the effect of deletion of one or more observations are measured on the estimated values of the linear predictor and some plots are also proposed to see the influence effect but , for modest data, our approach of likelihood influence measure is simpler as a means to find the influence effect on the regression parameter $\hat{\beta}$.

This chapter could be extended further by implementing Preisser J. S. and Qaqish B. F. (1996) approach to multinomial logit regression method. This implementation could be particularly effective when using a large data sets.

7.7 Key Remarks

Some important features are noted for fitting a full or a single case deletion models for finding the likelihood influence measures of Cook (1986). These are pointed out here for further multinomial data analysis with or without lots of zeros response entries.

a) If we have lots of 0's or 1's data entries in response levels then our approach for fitting a multinomial model may not converge on default convergence criteria. In this case, we need some careful considerations either for initial startup values or the 0's can be changed to the nearest 0's.

b) If in our approach convergence does not occurs then still we get all the statistics in the output exactly the same as Poisson trick model except the standard errors of the parameter estimates.

c) If the convergence occurs in our approach, the GLIM code $use loop$ used at the end of the macro only gives us the correct standard errors of parameter estimates

d) The macros given here to find the likelihood influence measure of Cook (1986) may be time consuming for a larger number of observations, especially for the joint and multiple influence measures.

e) The equivalence of parameter estimates between our approach and Poisson trick model method can be obtained either by the GLIM code 'groups' in macros of Poisson trick or using the respective response levels in our approach.

f) Our approach of finding the likelihood measures is flexible but requires a good knowledge for initial start-up values

g) Our approach can be preferred on Poisson trick approach due to the flexibility of fitting the model and for extracting some 'correct' statistics

7.8 Innovations

Some important achievements are discovered during our new direct iteratively re-weighted least squares (IRLS) approach of fitting the multinomial logit regression models and they can be noted as follows:-

a) Our approach is quite simple with all the calculations in array format and uses an algorithm that can be handled via the ordinary least squares approach.

b) We identify the block diagonal nature of the weight matrix for the multinomial regression and indicate how Cholesky reduction can provide efficient parameter estimation.

c) We discuss Cook's distance and the so called hat - matrix for the cells $(i\ ,\ j)$ of a multinomial response variable and conclude that these have only a limited use for multinomial data.

d) Our approach can use link functions other than the logit link.

e) We discuss how interval estimates and confidence limits for the parameters in the Box - Cox link functions can be found from plots of the deviance.

f) We show how to achieve faster convergence in some cases by using re - parameterisation of design matrix.

g) We discuss how the diagnostic approach of cell or observation deletion for multinomial data is not appropriate and suggest that case deletion can be appropriate, based on the likelihood influence measures of Cook (1986).

h) We discuss considerations that are needed in the process of fitting multinomial regression models via our approach.

7.9 Use of some other statistical package

In our approach for fitting the multinomial logit model using glim, we basically fit the model equation (2.30) where N and **y** - variable are defined by equations (2.31) and (2.32) respectively are of the form which gives a least square solution.

The glim procedure is defined to find $N = \begin{pmatrix} A_{11}X^{(2)} & A_{12}X^{(3)} \\ 0 & A_{22}X^{(3)} \end{pmatrix}$

and the **y** - variable $\begin{pmatrix} (\%lp) + \begin{pmatrix} A_{11}^{-1}(Y_2 - \mu_2) \\ -(A_{12}A_{22}^{-1})A_{11}^{-1}(Y_2 - \mu_2) + A_{22}^{-1}(Y_3 - \mu_3) \end{pmatrix} \end{pmatrix}$, where the

number of parameters not intrinsically aliased in the current model are defined in the glim procedure are given by

$(\%lp) = \begin{pmatrix} lp_2 \\ lp_3 \end{pmatrix} = \begin{pmatrix} A_{11}\eta^{(2)} + A_{12}\eta^{(3)} \\ A_{22}\eta^{(3)} \end{pmatrix}$ and the A_{ij} are the diagonal matrices obtained from

Cholesky decomposition of symmetric positive definite matrices defined in equation (2.15).

Using some different statistical package other than glim such as 'R', C^{++} and S^{+}, those can give the Cholesky decomposition equation (2.15) to find the diagonal matrices A_{ij} but still the equation (2.28) will be complicated to find the **y** - variable and the task to find the linear predictors equation (2.35) in our format seems impossible without knowing $(\%lp)$, which is easy to extract using the glim codes.

If we know how to extract $(\%lp)$ in any other statistical package then the macros of finding the linear predictors can be applied otherwise it will be more complicated to model the equation (2.30)to implemented our approach to fitting the multinomial logit regression.

CHAPTER 8

Conclusions and Remarks on Future Work

This study discusses a general technique for fitting a multinomial logit model. We have noted that the proposed method may not converge if we have lots of 0's in the response variable, as in general is the case in our examples in this study. In such a case, it seems that, if we do not update the weight matrix at each iteration, convergence can still be obtained using the GLIM code $use startup$ $use loop$. The correct fitted values and parameter estimates are then obtained but correct standard errors are not obtained. However, one more iteration of the form $use loop$ is needed to upgrade the weight matrix, which then leads to the correct standard errors. If convergence is not achieved within reasonable number of iterations, we may revise the form of the design matrix as is discussed in chapter 5.

The theory and methods given in this thesis for *grouped* multinomial data, as in the example on page 46, appear to work well when updating the weight matrix at each iteration. This approach, with the weight matrix for updated at each iteration, will provide the correct standard errors.

We may need to utilise some more knowledge about the data. For grouped data, we may use the observed proportions as start - up values; for non - grouped data, it is not sensible to use the observed proportions as start - up values. Hence random numbers are used by us as the start - up values and seem to work effectively. The theory and methods presented in this thesis can be extended to any number of levels of response variable, with any number of explanatory variables, without any major problem.

An interesting aspect of the method of this thesis is that the calculations only require a program that can only handle ordinary least squares, and so can be handled in a range of standard statistical software. This approach uses a simple form of Cholesky's decomposition applied to the weight matrix that here consists of diagonal sub - matrices. The formulation gives the required matrix inversion straightforwardly using only array calculations, so can be effected in standard statistical software without explicit matrix inversion.

The fitted method allows us any arbitrary link functions. This is therefore more general than the standard Poisson trick approach which uses only logit link function (although we too have concentrated in this research mainly on the logit link function). We can fit any appropriate linear predictor with or without interaction terms. Our results for the logit link functions can always be checked with the Poisson trick approach.

This research can be further extended to incorporate the testing of link functions, using the methods given in chapter 4 for other appropriate link functions. It would not be difficult to form the weight matrix in the macros for fitting a multinomial logit model with the use of the Appendix D for our own link function and equation (2.35).

The Poisson trick approach for the multinomial logit link function necessarily produces inappropriate leverages and Cook's distance. Our approach in this thesis produce the 'correct' leverages and Cook's distance for each cell; *i.e.* each level of the response variable for the multinomial data. However, we note that the single cell influence statistics are not appropriate and case deletion is needed to find the influence of case i on the parameter estimate $\hat{\beta}$.

We may extend this research by further investigating the hat - matrix, Cook's distance for each cell in the multinomial data as a cell influence on the parameter estimate $\hat{\beta}$ and some link functions different from the Box - Cox link functions could be considered.

In a practical situation, we may suspect the presence of outliers, or under or over - dispersion with respect to the multinomial assumption. The fitting of the multinomial logit model for contaminated data may become inefficient and can be biased. Because the estimates can be dragged towards the outlier and the variance estimate can be inflated, the result is that the outlier can be masked. In such a case we may further develop the theory given in this thesis for fitting a multinomial regression models in the context of robust estimation and outlier detection for over - dispersed multinomial models for grouped data, as for example discussed by Walter and Jasjeet (2004). They found a robust estimator - the hyperbolic tangent estimator - for over - dispersed multinomial regression models for grouped data. This provided accurate estimates and reliable inferences even when the specified model is not good for as much as half of the data. The examples that they considered were the votes cast for the President in the 2000 election in Florida, and the Polish parliament election 1993. As for the vote counts, the alternatives are the candidates or political parties that are competing for a particular office and the multinomial model is relevant when each voter cast one vote. Their model does not examine each individual separately but instead analyzes aggregates that correspond to the unit of observation. For vote counts the aggregates are usually legally defined voting districts, counties or provinces. Observations in this model measure the number of individuals in each unit who choose each alternative. In such a multinomial grouped data seriously ill - fitted counts were identified as outliers. The analysis of the example shows that with contaminated data, estimation fails with non - robust maximum likelihood estimator and the additive logistic model.

In general, the multinomial model treats the number of individuals in each observational unit as fixed and estimation focuses on how the proportion expected to choose each alternative depends on regressors. Each expected proportion corresponds to the probability of making each choice is usually modelled by the multinomial model. Usually in such analyses these probabilities are defined as logistic function of linear combination of regressors, as chosen in this thesis as our default option.

In practice the multinomial model may be inadequate, as in the example of Walter and Jaseet (2004) for vote counts. It has been widely recognised that aggregate vote data usually exhibit greater variability than the basic multinomial model can account for. In the basic multinomial model, the mean and the variance are determined by the same parameters. A common theme in several recently proposed models is to introduce additional parameters to allow the variance to be greater then the basic multinomial model would allow. Katz and King (1999), Jackson (2002) and Tomz, Tucker and Wittenberg (2002) allow not only the variance of each vote but also the covariance between votes for different candidates to differ from what the basic multinomial model specifies.

We may note that the recent analyses of count data in political science, such as Bratton and Ray (2002), Canes - Wrone et al. (2002), Hahn and Kenny (2002) , McDonagh (2002) and Monroe and Rose (2002), reduce the counts to percentages or proportions and hence ignore the over - dispersion. But ignoring the over - dispersion may result in incorrect statistical inferences. This could be the basis of our future work.

The work at a later stage could also be used the likelihood influence measure method given in chapter 7 for such a contaminated multinomial data to identify the part of the data for which the model is good and to separate those observations from the others. *i.e.* isolate the observations that are outliers relative to the specified model and not let them to distort the analysis. That would lead us further into the problem of the *masking* or *boosting* influence measure (Lawrence (1995)) of observation i after the deletion of observation j, and for these reasons it may not work out to try to identify the outliers one point at a time. It would seem desirable to have a method that locates all the outliers at once and further our ideas of deletion diagnostics can be extended using the one - step deletion diagnostic for $\hat{\beta}$ of Preisser J. S. and Qaqish B. F. (1996).

APPENDIX A

GLIM MACROS

GLIM directives for J=3 level of response variable with one explanatory variable for example 2.5.1

```
$c macro for fitting a multinomial logit model for Green's data$
$echo$
$sl 2912$c standard length of Green's data$
$num n$calc n=%nu/2$c n=1456, no.of response level$
$ass n2=n1,n1$c n1 declared in Green's data file$

$c variables notations used in macros$
$var  n w11 w22  w12
      a11 a12 a22

$ass nx=age,age$c explanatory variable$
$ass nw= vy2,vy3$c response variable are declared in Green's file$
$calc ystack=nw$

$c macros for initial values$
$macro startup
$calc %lp=%eta=%sr(0) $calc  %fv=%sr(0)/2$
$endmac

$num np1 twon$calc np1=n+1:twon=2*n$
$ass i=1...n:i2=np1...twon$

$c macro for calculating the deviance$
$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i2))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i2))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1$c variance function$
$endmac

$c macro for calculating the equation (2.35)$
$var n eta1  eta2$
$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2))/a11(i)$
$endmac

$c macro for calculating the equation (1.6)$
$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i2))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i))$

$calc %fv=%fv*n2:%dr=1$
$endmac

$c macro for calculating equation (2.15) and (2.20)$
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
      w22(i)=%fv(i2)*(1-%fv(i2)/n1(i)):
      w12(i)=-%fv(i)*%fv(i2)/n1(i)$

$calc  a11=%sqrt(w11):a12=w12/a11:
$calc  a22=%sqrt(w22-a12**2)$
$endmac
```

```
$c macro for calculating the equation (2.33)$
$var n py1 py2$
$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i):
   py2(i)=-(vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
             +(vy3(i)-%fv(n+i))/a22(i)$

$ass tem=py1,py2$
$calc %wvd=tem+%lp$c y-variable for the model$
$calc %wtd=1$
$endmac$

$var n one zero age$

$c macro for calculating the design matrix$
$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
      a4=a22*one $

$calc  a6=a11*age:a7=a12*age:
       a8=a22*age$

$ass d1=a1,zero: d2=a2,a4$
$ass dp1=a6,zero:dp2=a7,a8$
$endmac$

$yvar nw$error own devcalc$link own newlink2$
$scale 1.00$

$method * etacalc workvar$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2  1.0e-5$
$fit d1+d2+dp1+dp2-1$
$di e $
$endmac


$use startup$use loop$c considers initial values at each cycle in loop$

$use loop$c upgrade the weight matrix at each cycle for correct s.e.$
```

GLIM directives for J=3 response variable with one explanatory variable in Poisson trick approach of Francis et al, (1992) for example 2.5.2

```
$c green's data file is attached$
$ c macro for poisson trick approach of Francis et al(1992)$
$calc n1=1$

$c health declared in Green's data file$

$calc vy1=(health==1):vy2=(health==2):
          vy3=(health==3):vy4=(health==4)$
$calc vy5=(health>=3)$

$num vl$
$calc vl=3*%sl$

$ass freq=vy1,vy2,vy5$c response variable$
$var vl case group$

$calc  case=%gl(%sl,1):group=%gl(3,%sl)$
$fact case %sl group 3$

$elimimate case $error p $yvar freq$

$fit group$dis e$
$fit +group*age(case)$dis e$c age is a explanatory variable$
$return
```

GLIM directives with two explanatory variables for example 2.5.3

```
$c code notations can be found at page 134$

$c example 2.5.2 loglink(a2) 3.1 data file$

$echo$
$sle 8 $num n  $calc n=%nu/2$
$var n vy1 vy2 n1 Sex Age$
$data vy1 vy2 n1 Sex Age$
$read

9   5   20  1  1
4   1   10  1  2
3  11   15  2  1
9   6   21  2  2

$ass n2=n1,n1$
$var n w11 w22 w12
       a11 a22 a12
           py1 py2

$macro startup
$calc %lp=%eta=%sr(0) $calc %fv=%sr(0)/2$
$c calc %lp=%eta=%log(nw/(n2-nw)) $calc %fv=ystack$
$endmac

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                      )/a11(i)$
$endmac


$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i2-n))$

 $calc %fv=%fv*n2:%dr=1$endmac


$ass i=1...n $

$ass nx=Sex, Sex $
$ass nx1=Age,age $
$ass nw= vy1,vy2 $calc ystack=nw$

$yvar nw $error own  devcalc $link own newlink2
$scale 1.0$
```

```
$method * etacalc workvar$
$initial startup$
$calc a11=a12=a22=1$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 50 2  1.0e-5 $fit d1+dp1+d2+dp2+d3+dp3-1$display e$
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
          w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):

          w12(i)=-%fv(i)*%fv(i+n)/n1(i)$

$calc  a11=%sqrt(w11) :a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2)
$endmac

$macro workvar
$calc py1(i)=(vy1(i)-%fv(i))/a11(i):
   py2(i)=-(vy1(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
              +(vy2(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero $

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
         a4=a22*one $

$calc  a6=a11*Sex:a7=a12*Sex:
          a8=a22*Sex$
$calc a9=a11*age:a10=a12*age:
         a14=a22*age$

$ass d1=a1,zero: dp1=a2,a4$
$ass d2=a6,zero  :dp2=a7,a8 $
$ass d3=a9,zero :dp3=a10,a14$

$endmac$

$use startup$use loop$
$use loop$
```

GLIM directives with two explanatory variables in Poisson trick approach of Francis et al. (1992) for example 2.5.3

```
$c code notations can be found at page 136$

$echo $
$sle 4 $
$var n vy1 vy2  n Sex Age$
$data vy1 vy2  n Sex Age$
$read

6  9     20  1  1
5  4     10  1  2
1  3     15  2  1
6  9     21  2  2

$calc vy3=n - vy1 - vy2$
$num vl$
$calc  vl=3*%sl$
$ass freq=vy1,vy2,vy3$

$var vl case group$
$calc case=%gl(%sl,1) :group=%gl(3,%sl)
$fact case %sl group 3$

$eliminate case $error p $yvar freq$
$fit group$dis e$

$fit+group*sex(case)+group*age(case)$dis e$
```

GLIM directives for example chapter 3

```
$code notations can be found at page 134$
$c EEAap23.2adata(4levels)glim file$

$echo $
$sle 4368 $num n  $calc n=%nu/3 $
$ass n2=n1,n1,n1$

$var n w11 w22 w33 w12 w13 w23
      a11 a12 a13 a22 a23 a33

$ass nw= vy2,vy3,vy4$
$calc ystack=nw$

$macro startup
$calc %lp=%eta=%sr(0)$c calc %fv=%sr(0)/3$
$calc %fv=0.25$
$endmac

$num np1 np2 twon thren$
$calc np1=n+1:twon=2*n:
      np2=twon+1:thren=3*n$
$ass i=1...n:i2=np1...twon :i3=np2...thren$

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i2)-%fv(i3))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i2)-ystack(i3))$
$calc y1=1/3*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$var n eta1 eta2 eta3$
$macro etacalc
$calc %eta(i3)=%lp(i3)/a33(i)$
$calc %eta(i2)=(%lp(i2)-a23(i)*%eta(i3))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                    -a13(i)*%eta(i3))/a11(i)$
$endmac


$macro newlink
$cal expta=%exp(%eta)$
$calc %fv(i)=
         expta(i)/(1+expta(i)+expta(i2)+expta(i3))$
$calc %fv(i2)=
        expta(i2)/(1+expta(i2)+expta(i)+expta(i3))$
$calc %fv(i3)=
        expta(i3)/(1+expta(i3)+expta(i2)+expta(i))$

$calc %fv=%fv*n2:%dr=1$
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
      w22(i)=%fv(i2)*(1-%fv(i2)/n1(i)):
      w33(i)= %fv(i3)*(1-%fv(i3)/n1(i)):
      w12(i)=-%fv(i)*%fv(i2)/n1(i):
      w13(i)=-%fv(i)*%fv(i3)/n1(i):
      w23(i)=-%fv(i2)*%fv(i3)/n1(i)$
```

```
$calc  a11=%sqrt(w11):a12=w12/a11:
       a13=w13/a11$
$calc  a22=%sqrt(w22-a12**2):a23=(w23-a12*a13)/a22$
$calc  a33=%sqrt(w33-a23**2-a13**2)$
$endmac

$var n py1 py2 py3 py4 py5 py6 py7 py8$

$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i):
   py2(i)=-(vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
              +(vy3(i)-%fv(i2))/a22(i)$

 $calc py3(i)=(vy2(i)-%fv(i))*((((a12(i)*a23(i)/a33(i))/a22(i))
                    -a13(i)/a33(i))/a11(i))
                  - (vy3(i)-%fv(i2))*(a23(i)/a33(i))/a22(i)
                + (vy4(i)-%fv(i3))/a33(i)$

$calc py4=(vy2(i)-%fv(i))*(a23(i)*a12(i))/(a11(i)*a22(i)*a33(i))$
$calc py5=-(vy2(i)-%fv(i))*a13(i)/(a11(i)*a33(i))$
$calc py6=-(vy3(i)-%fv(i2))*a23(i)/(a22(i)*a33(i))$
$calc py7=(vy4(i)-%fv(i3))/a33(i)$
$calc py8=py4+py5+py6+py7$

$ass tem=py1,py2,py3$
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc  a1=a11*one:a2=a12*one:
         a3=a13*one:a4=a22*one:
         a5=a23*one:a6=a33*one$

$calc  a51=a11*age:a52=a12*age:
          a53=a13*age:a54=a22*age:
          a55=a23*age:a56=a33*age$

$ass d1=a1,zero,zero:d2=a2,a4,zero:
        d3=a3,a5,a6$
$ass dp1=a51,zero,zero:dp2=a52,a54,zero:
        dp3=a53,a55,a56$
$cal d1=d1:d2=d2:d3=d3:
             dp1=dp1:dp2=dp2:dp3=dp3$

$endmac$

$yvar nw$error own  devcalc$link own newlink$
$scale 1.00$
$method * etacalc workvar$

$Macro loop
$cycle 500 2 1.0e-5$
$fit d1+d2+d3+dp1+dp2+dp3-1$
$dis e$
$endmac

$use startup$use loop$
$use loop$
```

GLIM directives of Poisson trick approach chapter 3

```
$c codes notation can be found at page 136$
$c and attached Green's data file$

$pick HEALTH,AGE SEL$
$sort AGE,HEALTH $

$calc  n1=1$
$calc  vy1=(health==1): vy2=(health==2):
       vy3=(health==3): vy4=(health==4)$
$num vl$
$calc vl=4*%sl$
$ass freq=vy1,vy2,vy3,vy4$

$var vl case group$
$calc  case=%gl(%sl,1) :group=%gl(4,%sl) $
$fact case %sl group 4$

$eliminate case $error p $yvay freq $

$fit group$dis e$

$fit +group*age(case)$dis e$

$return
```

Green's GLIM directives file

```
$subfile GSS84 !
! title '1984 General Social Survey'.

$sle 1456$

$data 1473 id race age sex marital satjob hapmar life postlife
  educ degree paeduc maeduc speduc income health hompop agewed sibs
$format (I4,1X,I1,I2,1X,6I1,1X,I2,I1,I2,I2,I2,I2,I1,I2,I2,I2)
$comment
/ id 1-4 race 6 age 7-8
  sex 10 marital 11 satjob 12 hapmar 13 life 14 postlife 15
  educ 17-18 degree 19 paeduc 20-21 maeduc 22-23 speduc 24-25
  income 26-27 health 28 hompop 29-30 agewed 31-32 sibs 33-34.
variable labels
  age "Age of respondent"
  health "Condition of health"

value labels
  health 1 'Excellent' 2 'Good' 3 'Fair' 4 'Poor' 9 'Missing data' /

missing value
  age, educ, paeduc, maeduc, speduc, hompop, agewed,
  sibs, income (99) /.
$ ! end comment
$din 'a:gss84.dat' AGE HEALTH $
!
! omit missing data and sort on AGE
!
$cal SEL=(AGE<99)&(HEALTH>0)&(HEALTH<8)$
$pick HEALTH,AGE SEL$
$sort AGE,HEALTH $

$calc  n1=1$
$calc  vy1=(health==1): vy2=(health==2):
           vy3=(health==3): vy4=(health==4) $


$return
```

GLIM directives for example 4.1

```
$c codes notation can be found at page 134$
$c E1Eaap23.dat(4.1)glim$
$echo $
$sle 4368 $num n  step$calc n=%nu/3 $
$ass n2=n1,n1,n1$
$var n w11 w22 w33 w12 w13 w23
       a11 a12 a13 a23 a22 a33
       py1 py2 py3

$num a$
$ass a=0.0001$

$macro startup
$c calc %lp=%sr(0): %eta=%log(0.5 ) $calc  %fv=0.25$
$calc %lp=%eta=%sr(0)$c calc %fv=%sr(0)/4 $
$c calc %lp=%eta=%log(nw/(n2-nw)) $c calc %fv=ystack
$endmac

$num np1 np2 twon thren $calc np1=n+1:twon=2*n:
                 np2=twon+1:thren=3*n
$ass j=1...%nu:i2=np1...twon :i3=np2...thren$

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n)-%fv(i+2*n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n)-ystack(i+2*n))$
$calc y1=1/3*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i3)=%lp(i3)/a33(i3-2*n)$
$calc %eta(i2)=(%lp(i2)-a23(i2-n)*%eta(i3))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                    -a13(i)*%eta(i3))/a11(i)$
$endmac

$macro newlink
$calc  %fv(i)=(1+a*%eta(i))**(1/a)/
               (1+
                (1+a*%eta(i))**(1/a)+(1+a*%eta(i2))**(1/a)+
                (1+a*%eta(i3))**(1/a)  )$

$calc %fv(i2)=(1+a*%eta(i2))**(1/a)/
                 (1+
                   (1+a*%eta(i))**(1/a)+(1+a*%eta(i2))**(1/a)+
                   (1+a*%eta(i3))**(1/a)  )$

$calc  %fv(i3)=(1+a*%eta(i3))**(1/a)/
               (1+
                (1+a*%eta(i))**(1/a)+(1+a*%eta(i2))**(1/a)+
                 (1+a*%eta(i3))**(1/a)  )$

 $calc %fv=%fv*n2:%dr=1$endmac
$ass i=1...n $
$ass nw= vy2,vy3,vy4$
$calc ystack=nw$
```

```
$yvar nw $error own  devcalc $link own newlink
$scale 1.00$
$method * etacalc workvar$
$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2 1.0e-5 $fit d1+d2+d3+dp1+dp2+dp3-1$dis e$
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i))*(1/1+a*%eta(i))**2:
           w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i))*
                                    (1/1+a*%eta(i2))**2$
$calc  w33(i)= %fv(i+2*n)*(1-%fv(i+2*n)/n1(i))*
                                     (1/1+a*%eta(i3))**2:
          w12(i)=-%fv(i)*%fv(i+n)/n1(i)*(1/1+a*%eta(i))*
                                   (1/1+a*%eta(i2))$
 $calc  w13(i)=-%fv(i)*%fv(i+2*n)/n1(i)*(1/1+a*%eta(i))*
                                   (1/1+a*%eta(i3)):
          w23(i)=-%fv(i+n)*%fv(i+2*n)/n1(i)*(1/1+a*%eta(i2))*
                                     (1/1+a*%eta(i3))$

$calc  a11=%sqrt(w11) :a12=w12/a11:
          a13=w13/a11$
$calc  a22=%sqrt(w22-a12**2):a23=(w23-a12*a13)/a22$
$calc a33=%sqrt(w33-a23**2-a13**2)$
$endmac

$var n eta1 eta2 eta3$

$macro workvar
$calc py1(i)=((vy2(i)-%fv(i))/a11(i))*(1/(1+a*%eta(i))):
   py2(i)=-((vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i))*(1/(1+a*%eta(i)))
              +((vy3(i)-%fv(n+i))/a22(i))*(1/(1+a*%eta(i2)))$
 $calc  py3(i)=((vy2(i)-%fv(i))*((a12(i)*a23(i)/a33(i)/a22(i)
                    -a13(i)/a33(i))/a11(i)))*(1/(1+a*%eta(i)))
                   - ((vy3(i)-
%fv(n+i))*((a23(i)/a33(i))/a22(i)))*(1/(1+a*%eta(i2)))
                  + (vy4(i)-%fv(2*n+i))/a33(i)*(1/(1+a*%eta(i3)))$

$ass tem=py1,py2,py3$
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$
$macro newdesign
$calc one=1:zero=0$
$calc  a1=a11*one:a2=a12*one:
       a3=a13*one:a4=a22*one:
       a5=a23*one:a6=a33*one$
$calc  a51=a11*age:a52=a12*age:
       a53=a13*age:a54=a22*age:
       a55=a23*age:a56=a33*age$
$ass d1=a1,zero,zero:d2=a2,a4,zero:
     d3=a3,a5,a6$
$ass dp1=a51,zero,zero:dp2=a52,a54,zero:
      dp3=a53,a55,a56$
$cal d1=d1:d2=d2:d3=d3:
     dp1=dp1:dp2=dp2:dp3=dp3$
$endmac$

$use startup$use loop$
$use loop$
```

GLIM directives for example 4.2

```
$c codes notation can be found at page 134$
$c E1Eaap23.dat(4.2)glim$
$echo $
$sle 4368 $num n  step$calc n=%nu/3 $

$ass n2=n1,n1,n1$
$var n w11 w22 w33 w12 w13 w23
       a11 a12 a13 a23 a22 a33
       py1 py2 py3

$num a b c$
$ass a=0.0001:b=0.0001:c=0.0001$

$macro startup
$c calc %lp=%sr(0): %eta=%log(0.5 ) $calc  %fv=0.25$
$calc %lp=%eta=%sr(0)$c calc %fv=%sr(0)/4 $
$c calc %lp=%eta=%log(nw/(n2-nw)) $c calc %fv=ystack
$endmac

$num np1 np2 twon thren $calc np1=n+1:twon=2*n:
                  np2=twon+1:thren=3*n
$ass j=1...%nu:i2=np1...twon :i3=np2...thren$

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n)-%fv(i+2*n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n)-ystack(i+2*n))$
$calc y1=1/3*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i3)=%lp(i3)/a33(i3-2*n)$
$calc %eta(i2)=(%lp(i2)-a23(i2-n)*%eta(i3))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                     -a13(i)*%eta(i3))/a11(i)$
$endmac

$macro newlink
$calc  %fv(i)=(1+a*%eta(i))**(1/a)/
               (1+
                (1+a*%eta(i))**(1/a)+(1+b*%eta(i2))**(1/b)+
                (1+c*%eta(i3))**(1/c)  )$
$calc %fv(i2)=(1+b*%eta(i2))**(1/b)/
                 (1+
                   (1+a*%eta(i))**(1/a)+(1+b*%eta(i2))**(1/b)+
                   (1+c*%eta(i3))**(1/c)  )$
$calc  %fv(i3)=(1+c*%eta(i3))**(1/c)/
               (1+
                (1+a*%eta(i))**(1/a)+(1+b*%eta(i2))**(1/b)+
                 (1+c*%eta(i3))**(1/c)  )$

 $calc %fv=%fv*n2:%dr=1$endmac

$ass i=1...n $
$ass nw= vy2,vy3,vy4$
$calc ystack=nw$
$yvar nw $error own  devcalc $link own newlink
$scale 1.00$
$method * etacalc workvar$
```

```
$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2 1.0e-5 $fit d1+d2+d3+dp1+dp2+dp3-1$
$dis e $
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i))*(1/1+a*%eta(i))**2:
          w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i))*
                                    (1/1+b*%eta(i2))**2$
$calc  w33(i)= %fv(i+2*n)*(1-%fv(i+2*n)/n1(i))*
                                     (1/1+c*%eta(i3))**2:
        w12(i)=-%fv(i)*%fv(i+n)/n1(i)*(1/1+a*%eta(i))*
                                   (1/1+b*%eta(i2))$
 $calc  w13(i)=-%fv(i)*%fv(i+2*n)/n1(i)*(1/1+a*%eta(i))*
                                   (1/1+c*%eta(i3)):
          w23(i)=-%fv(i+n)*%fv(i+2*n)/n1(i)*(1/1+b*%eta(i2))*
                                     (1/1+c*%eta(i3))$

$calc  a11=%sqrt(w11) :a12=w12/a11:
          a13=w13/a11$
$calc  a22=%sqrt(w22-a12**2):a23=(w23-a12*a13)/a22$
$calc a33=%sqrt(w33-a23**2-a13**2)$
$endmac

$var n eta1 eta2 eta3$

$macro workvar
$calc py1(i)=((vy2(i)-%fv(i))/a11(i))*(1/(1+a*%eta(i))):
   py2(i)=-((vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i))*(1/(1+a*%eta(i)))
              +((vy3(i)-%fv(n+i))/a22(i))*(1/(1+b*%eta(i2)))$
 $calc  py3(i)=((vy2(i)-%fv(i))*((a12(i)*a23(i)/a33(i)/a22(i)
                    -a13(i)/a33(i))/a11(i)))*(1/(1+a*%eta(i)))
                   - ((vy3(i)-
%fv(n+i))*((a23(i)/a33(i))/a22(i)))*(1/(1+b*%eta(i2)))
                  + (vy4(i)-%fv(2*n+i))/a33(i)*(1/(1+c*%eta(i3)))$

$ass tem=py1,py2,py3$
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc  a1=a11*one:a2=a12*one:
          a3=a13*one:a4=a22*one:
          a5=a23*one:a6=a33*one$
$calc  a51=a11*age:a52=a12*age:
           a53=a13*age:a54=a22*age:
           a55=a23*age:a56=a33*age$
$ass d1=a1,zero,zero:d2=a2,a4,zero:
         d3=a3,a5,a6$
$ass dp1=a51,zero,zero:dp2=a52,a54,zero:
         dp3=a53,a55,a56$
$cal d1=d1:d2=d2:d3=d3:
              dp1=dp1:dp2=dp2:dp3=dp3$

$endmac$

$use startup$use loop$
$use loop$
```

GLIM directives for example 5.1

```
$c codes notation can be found at page 134$
$sl 1742$
$echo $
$num n  $calc n=%nu/2$
$var eta1 eta2$
$ass n2=n1,n1$
$var  n w11 w22  w12
        a11 a12 a22 py1 py2

$macro startup
$calc %lp=%eta=%sr(0) $calc  %fv=%sr(0)/2$
$endmac

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1$
$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2))/a11(i)$
$endmac

$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i2-n))$

$calc %fv=%fv*n2 : %dr=1$
$endmac

$ass i=1...n $

$c ass nx=age,age$
$ass nw= vy2,vy3 $calc ystack=nw$
$yvar nw $error own  devcalc $link own newlink2

$method * etacalc workvar$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2  1.0e-5$
$fit d1+d2+dp1+dp2-1$
$di e $
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
          w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):
          w12(i)=-%fv(i)*%fv(i+n)/n1(i)$
```

```
$calc  a11=%sqrt(w11):a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2)$
$endmac

$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i):
   py2(i)=-(vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
             +(vy3(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $

$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac$

$var n one zero $

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
         a4=a22*one $

$calc a3=a11*ag$

$calc  a8=a12*ag:a9=a22*ag$

$ass d1=a1,zero: dp1=a2,a4$
$ass d2=a3,zero  :dp2=a8,a9 $


$endmac$

$use startup$use loop$
```

GLIM directive for example 5.3 with improved design matrix (a)

```
$c codes notation can be found at page 134$
$sl 1742$
$echo $
$num n  $calc n=%nu/2$
$var eta1 eta2$
$ass n2=n1,n1$
$var  n w11 w22  w12
        a11 a12 a22  py1 py2

$macro startup
$calc %lp=%eta=%sr(0) $calc  %fv=%sr(0)/2$
$endmac

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1$
$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2))/a11(i)$
$endmac

$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i2-n))$

$calc %fv=%fv*n2 : %dr=1$
$endmac

$ass i=1...n $
$c ass nx=age,age$
$ass nw= vy2,vy3 $calc ystack=nw$
$yvar nw $error own  devcalc $link own newlink2

$method * etacalc workvar$
$initial startup$
$calc a11=a12=a22=1$

$macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2  1.0e-5$
$fit d1+d2+dp1+dp2-1$
$extract  %pe$
$pr 'para1 estimate' %pe$
$di e $
$endmac
```

```
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
          w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):
          w12(i)=-%fv(i)*%fv(i+n)/n1(i)$

$calc  a11=%sqrt(w11) :a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2) $
$endmac

$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i):
   py2(i)=-(vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
              +(vy3(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $

$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac$

$var n one zero $

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
         a4=a22*one $

$calc a3=a11*ag$
$calc a5=a1+a2$
$calc  a8=a12*ag:a9=a22*ag$
$calc a6=a3+a8$

$ass d1=a5,a4: dp1=a2,a4$
$ass d2=a6,a9  :dp2=a8,a9 $

$endmac$
$use startup$use loop$
```

GLIM directives for example 5.3 with new design matrix (b)

```
$c codes notation can be found at page 134$
$sl 1742$
$echo $
$num n  $calc n=%nu/2$
$var eta1 eta2$
$ass n2=n1,n1$
$var  n w11 w22  w12
       a11 a12 a22  py1 py2

$macro startup
$calc %lp=%eta=%sr(0) $calc  %fv=%sr(0)/2$
$endmac

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1$
$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2))/a11(i)$
$endmac

$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
         expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
        expta(i2)/(1+expta(i2)+expta(i2-n))$

$calc %fv=%fv*n2 : %dr=1$
$endmac

$ass i=1...n $
$c ass nx=age,age$
$ass nw= vy2,vy3 $calc ystack=nw$
$yvar nw $error own  devcalc $link own newlink2

$method * etacalc workvar$
$initial startup$
$calc a11=a12=a22=1$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2  1.0e-5$
$fit d1+d2+d3+dp1+dp2+dp3-1$
$extract  %pe$
$pr 'para1 estimate' %pe$
$di e $
$endmac
```

```
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
          w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):
          w12(i)=-%fv(i)*%fv(i+n)/n1(i)$

$calc  a11=%sqrt(w11) :a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2) $
$endmac

$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i):
   py2(i)=-(vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
              +(vy3(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $

$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac$

$var n one zero $

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
      a4=a22*one $

$calc a3=a11*ag$

$calc a5=a11*ag*ag$

$calc a6=a12*ag*ag :a7=a22*ag*ag$

$calc  a8=a12*ag:a9=a22*ag$

$ass d1=a1,zero: dp1=a2,a4$
$ass d2=a3,zero  :dp2=a8,a9 $
$ass d3=a5,zero: dp3=a6,a7$


$endmac$

$use startup$use loop$
```

GLIM directives for four levels of response variable with six explanatory variables for data in section 6.3

```
$c codes notation can be found at page 134$
$echo $
$sle 1098 $num n $calc n=%nu/3 $
$ass n2=n1,n1,n1$

$var n w11 w22 w33 w12 w13 w23
      a11 a12 a13 a22 a23 a33
      py1 py2 py3

$macro startup
$calc %lp=%eta=%sr(0)$c calc %fv=%sr(0)/3$
$calc %fv=0.25$
$endmac

$num np1 np2 twon thren $calc np1=n+1:twon=2*n:
        np2=twon+1:thren=3*n
$ass j=1...%nu:i2=np1...twon :i3=np2...thren$

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n)-%fv(i+2*n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n)-ystack(i+2*n))$
$calc y1=1/3*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i3)=%lp(i3)/a33(i3-2*n)$
$calc %eta(i2)=(%lp(i2)-a23(i2-n)*%eta(i3))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                    -a13(i)*%eta(i3))/a11(i)$
$endmac

$macro newlink
$calc expta=%exp(%eta)$
$calc  %fv(i)=expta(i)/(1+expta(i)+expta(i2)+expta(i3))
$calc %fv(i2)=expta(i2)/(1+expta(i)+expta(i2)+expta(i3))
$calc  %fv(i3)=expta(i3)/(1+expta(i)+expta(i2)+expta(i3))
 $calc %fv=%fv*n2:%dr=1$endmac

$ass i=1...n $

$ass nw= vy2,vy3,vy4 $
$calc ystack=nw$
$yvar nw $error own  devcalc $link own newlink
$scale 1.00$
$method * etacalc workvar$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2 $

$fit d1+d2+d3+dp1+dp2+dp3+dp4+dp5+dp6+dp7+dp8+
                        dp9+dp10+dp11+ dp12+dp13+dp14+
                        dp15+dp16+dp17+dp18-1$

$dis e $endmac
```

```
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
      w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i))$

$calc w33(i)= %fv(i+2*n)*(1-%fv(i+2*n)/n1(i)):
      w12(i)=-%fv(i)*%fv(i+n)/n1(i)$

 $calc w13(i)=-%fv(i)*%fv(i+2*n)/n1(i):
       w23(i)=-%fv(i+n)*%fv(i+2*n)/n1(i)$

$calc a11=%sqrt(w11) :a12=w12/a11:
      a13=w13/a11$
$calc a22=%sqrt(w22-a12**2)  :a23=(w23-a12*a13)/a22
$calc a33=%sqrt(w33-a23**2-a13**2)$
$endmac

$var n eta1 eta2 eta3$

$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i)$

$calc py2(i)=-(vy2(i)-%fv(i))*(a12(i)/(a22(i)*a11(i)))

          +(vy3(i)-%fv(n+i))/a22(i)$

 $calc  py3(i)=(vy2(i)-%fv(i))*((a12(i)*a23(i))/(a33(i)*a22(i)*a11(i)))

               -(vy2(i)-%fv(i))*(a13(i)/(a33(i)*a11(i)))

               -(vy3(i)-%fv(n+i))*(a23(i)/(a33(i)*a22(i)))

               +(vy4(i)-%fv(2*n+i))/a33(i)$

$ass tem=py1,py2,py3$
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc  a1=a11*one:a2=a12*one:
          a3=a13*one:a4=a22*one:
          a5=a23*one:a6=a33*one$

$calc  a51=a11*koe: a52=a12*koe$
$calc  a53=a13*koe: a54=a22*koe$
 $calc a55=a23*koe: a56=a33*koe$

$calc  a61=a11*pnl: a62=a12*pnl$
$calc  a63=a13*pnl: a64=a22*pnl$
$calc   a65=a23*pnl: a66=a33*pnl$

$calc  a71=a11*par: a72=a12*par$
$calc  a73=a13*par: a74=a22*par$
$calc  a75=a23*par: a76=a33*par$

$calc a81=a11*foc: a82=a12*foc$
$calc a83=a13*foc: a84=a22*foc$
$calc a85=a23*foc: a86=a33*foc$

$calc a91=a11*spo: a92=a12*spo$
$calc a93=a13*spo: a94=a22*spo$
$calc a95=a23*spo: a96=a33*spo$
```

```
$calc a41=a11*fol: a42=a12*fol$
$calc a43=a13*fol: a44=a22*fol$
$calc a45=a23*fol: a46=a33*fol$

$ass d1=a1,zero,zero:d2=a2,a4,zero:
        d3=a3,a5,a6$

$ass dp1=a51,zero,zero:dp2=a52,a54,zero:
        dp3=a53,a55,a56$

$cal d1=d1:d2=d2:d3=d3:
             dp1=dp1:dp2=dp2:dp3=dp3$

$ass dp4=a61,zero,zero:dp5=a62,a64,zero$
$ass dp6=a63,a65,a66$

$ass dp7=a71,zero,zero:dp8=a72,a74,zero $
$ass dp9=a73,a75,a76$

$cal dp4=dp4:dp5=dp5:dp6=dp6:
             dp7=dp7:dp8=dp8:dp9=dp9$

$ass dp10=a81,zero,zero:dp11=a82,a84,zero$
$ass dp12=a83,a85,a86$

$ass dp13=a91,zero,zero:dp14=a92,a94,zero$
$ass dp15=a93,a95,a96$

$ass dp16=a41,zero,zero:dp17=a42,a44,zero$
$ass dp18=a43,a45,a46$

$calc dp10=dp10:dp11=dp11:dp12=dp12:dp13=dp13$
$calc dp14=dp14:dp15=dp15:dp16=dp16:dp17=dp17$
$calc dp18=dp18$


$endmac$

$use startup$use loop$
$use loop$
```

GLIM directives for different parameters of four levels of response variable with six explanatory variables data for section 6.5

```
$c codes notation can be found at page 134$
$echo $
$sle 1098 $num n  step$calc n=%nu/3 $
$ass n2=n1,n1,n1$
$var n w11 w22 w33 w12 w13 w23
       a11 a12 a13 a22 a23 a33
       py1 py2 py3

$num a b c $
$ass a=0.0001:b=0.0001:c=0.0001$

$macro startup
$calc %lp=%eta=%sr(0)$calc  %fv=0.25$
$endmac

$num np1 np2 twon thren $calc np1=n+1:twon=2*n:
         np2=twon+1:thren=3*n
$ass j=1...%nu:i2=np1...twon :i3=np2...thren$

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n)-%fv(i+2*n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n)-ystack(i+2*n))$
$calc y1=1/3*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i3)=%lp(i3)/a33(i3-2*n)$
$calc %eta(i2)=(%lp(i2)-a23(i2-n)*%eta(i3))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                     -a13(i)*%eta(i3))/a11(i)$
$endmac

$macro newlink
$calc  %fv(i)=(1+a*%eta(i))**(1/a)/
               (1+
                (1+a*%eta(i))**(1/a)+(1+b*%eta(i2))**(1/b)+
                (1+c*%eta(i3))**(1/c)  )$
$calc %fv(i2)=(1+a*%eta(i2))**(1/b)/
                 (1+
                   (1+a*%eta(i))**(1/a)+(1+a*%eta(i2))**(1/a)+
                   (1+a*%eta(i3))**(1/a)  )$
$calc  %fv(i3)=(1+a*%eta(i3))**(1/a)/
               (1+
                (1+a*%eta(i))**(1/a)+(1+b*%eta(i2))**(1/b)+
                 (1+c*%eta(i3))**(1/c)  )$

 $calc %fv=%fv*n2:%dr=1$endmac

$ass i=1...n $

$ass nw= vy2,vy3,vy4 $
$calc ystack=nw$

$yvar nw $error own  devcalc $link own newlink
$scale 1.00$
$method * etacalc workvar$
```

```
$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2 $

$fit d1+d2+d3+dp1+dp2+dp3+dp4+dp5+dp6+dp7+dp8+
                        dp9+dp10+dp11+ dp12+dp13+dp14+
                        dp15+dp16+dp17+dp18-1$

$dis e $
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i))*(1/1+a*%eta(i))**2:
          w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i))*
                                  (1/1+b*%eta(i2))**2$
$calc  w33(i)= %fv(i+2*n)*(1-%fv(i+2*n)/n1(i))*
                                   (1/1+c*%eta(i3))**2:
        w12(i)=-%fv(i)*%fv(i+n)/n1(i)*(1/1+a*%eta(i))*
                                 (1/1+b*%eta(i2))$
 $calc  w13(i)=-%fv(i)*%fv(i+2*n)/n1(i)*(1/1+a*%eta(i))*
                                 (1/1+c*%eta(i3)):
        w23(i)=-%fv(i+n)*%fv(i+2*n)/n1(i)*(1/1+b*%eta(i2))*
                                   (1/1+c*%eta(i3))$

$calc  a11=%sqrt(w11) :a12=w12/a11:
          a13=w13/a11
$calc   a22=%sqrt(w22-a12**2)  :a23=(w23-a12*a13)/a22

$calc a33=%sqrt(w33-a23**2-a13**2)
$endmac

$var n eta1 eta2 eta3$

$macro workvar
$calc py1(i)=((vy2(i)-%fv(i))/a11(i))*(1/(1+a*%eta(i))):
   py2(i)=-((vy2(i)-%fv(i))*(a12(i)/a22(i))/a11(i))*(1/(1+a*%eta(i)))
             +((vy3(i)-%fv(n+i))/a22(i))*(1/(1+b*%eta(i2)))$
 $calc  py3(i)=((vy2(i)-%fv(i))*((a12(i)*a23(i)/a33(i)/a22(i)
                   -a13(i)/a33(i))/a11(i)))*(1/(1+a*%eta(i)))
                  - ((vy3(i)-
%fv(n+i))*((a23(i)/a33(i))/a22(i)))*(1/(1+b*%eta(i2)))
                 + (vy4(i)-%fv(2*n+i))/a33(i)*(1/(1+c*%eta(i3)))$

$ass tem=py1,py2,py3$

$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc  a1=a11*one:a2=a12*one:
          a3=a13*one:a4=a22*one:
          a5=a23*one:a6=a33*one$

$calc  a51=a11*koe: a52=a12*koe$
$calc  a53=a13*koe: a54=a22*koe$
$calc a55=a23*koe: a56=a33*koe$

$calc  a61=a11*pnl: a62=a12*pnl$
$calc  a63=a13*pnl: a64=a22*pnl$
$calc   a65=a23*pnl: a66=a33*pnl$
```

```
$calc  a71=a11*par: a72=a12*par$
$calc  a73=a13*par: a74=a22*par$
$calc  a75=a23*par: a76=a33*par$

$calc a81=a11*foc: a82=a12*foc$
$calc a83=a13*foc: a84=a22*foc$
$calc a85=a23*foc: a86=a33*foc$

$calc a91=a11*spo: a92=a12*spo$
$calc a93=a13*spo: a94=a22*spo$
$calc a95=a23*spo: a96=a33*spo$

$calc a41=a11*fol: a42=a12*fol$
$calc a43=a13*fol: a44=a22*fol$
$calc a45=a23*fol: a46=a33*fol$

$ass d1=a1,zero,zero:d2=a2,a4,zero:
     d3=a3,a5,a6$

$ass dp1=a51,zero,zero:dp2=a52,a54,zero:
     dp3=a53,a55,a56$

$cal d1=d1:d2=d2:d3=d3:
     dp1=dp1:dp2=dp2:dp3=dp3$

$ass dp4=a61,zero,zero:dp5=a62,a64,zero$
$ass dp6=a63,a65,a66$

$ass dp7=a71,zero,zero:dp8=a72,a74,zero $
$ass dp9=a73,a75,a76$

$cal dp4=dp4:dp5=dp5:dp6=dp6:
     dp7=dp7:dp8=dp8:dp9=dp9$

$ass dp10=a81,zero,zero:dp11=a82,a84,zero$
$ass dp12=a83,a85,a86$

$ass dp13=a91,zero,zero:dp14=a92,a94,zero$
$ass dp15=a93,a95,a96$

$ass dp16=a41,zero,zero:dp17=a42,a44,zero$
$ass dp18=a43,a45,a46$

$calc dp10=dp10:dp11=dp11:dp12=dp12:dp13=dp13$
$calc dp14=dp14:dp15=dp15:dp16=dp16:dp17=dp17$
$calc dp18=dp18$


$endmac$

$use startup$use loop$
$use loop$
```

GLIM Directives for improved design matrix for section 6.6

```
$c codes notation can be found at page 134$
$c GLIM Directives for improved design matrix example 6.5$
$echo $
$sle 1098 $num n  step$calc n=%nu/3 $
$ass n2=n1,n1,n1$

$var n w11 w22 w33 w12 w13 w23
       a11 a12 a13 a22 a23 a33
       py1 py2 py3

$macro startup
$calc %lp=%eta=%sr(0)$
$calc  %fv=0.25$
$endmac

$num np1 np2 twon thren $calc np1=n+1:twon=2*n:
         np2=twon+1:thren=3*n
$ass j=1...%nu:i2=np1...twon :i3=np2...thren$

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n)-%fv(i+2*n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n)-ystack(i+2*n))$
$calc y1=1/3*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i3)=%lp(i3)/a33(i3-2*n)$
$calc %eta(i2)=(%lp(i2)-a23(i2-n)*%eta(i3))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                     -a13(i)*%eta(i3))/a11(i)$
$endmac

$macro newlink
$calc expta=%exp(%eta)$

$calc  %fv(i)=expta(i)/(1+expta(i)+expta(i2)+expta(i3))
$calc %fv(i2)=expta(i2)/(1+expta(i)+expta(i2)+expta(i3))
$calc  %fv(i3)=expta(i3)/(1+expta(i)+expta(i2)+expta(i3))

$calc %fv=%fv*n2:%dr=1$endmac

$ass i=1...n $
$ass nw= vy2,vy3,vy4 $
$calc ystack=nw$

$yvar nw $error own  devcalc $link own newlink
$scale 1.00
$method * etacalc workvar$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 500 2 1.0e-4 $

$fit p1+p2+p3+pd1+pd2+pd3+pd4+pd5+pd6+pd7+pd8+
                        pd9+pd10+pd11+pd12+pd13+pd14+
                        pd15+pd16+pd17+pd18-1$

$dis e $endmac
```

```
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i))/n1(i)$
$calc w22(i)=%fv(i+n)*(1-%fv(i+n))/n1(i)$
$calc w33(i)= %fv(i+2*n)*(1-%fv(i+2*n))/n1(i)$
$calc w12(i)=-%fv(i)*%fv(i+n)/n1(i)$
$calc w13(i)=-%fv(i)*%fv(i+2*n)/n1(i)$
$calc w23(i)=-%fv(i+n)*%fv(i+2*n)/n1(i)$

$calc a11=%sqrt(w11):a12=w12/a11:
      a13=w13/a11$
$calc a22=%sqrt(w22-a12**2):a23=(w23-a12*a13)/a22$
$calc a33=%sqrt(w33-a23**2-a13**2)$
$endmac

$var n eta1 eta2 eta3$

$macro workvar
$calc py1(i)=(vy2(i)-%fv(i))/a11(i)$

$calc py2(i)=-(vy2(i)-%fv(i))*(a12(i)/(a22(i)*a11(i)))

              +(vy3(i)-%fv(n+i))/a22(i)$

 $calc  py3(i)=(vy2(i)-%fv(i))*((a12(i)*a23(i))/(a33(i)*a22(i)*a11(i)))

                    -(vy2(i)-%fv(i))*(a13(i)/(a33(i)*a11(i)))

                   -(vy3(i)-%fv(n+i))*(a23(i)/(a33(i)*a22(i)))

                  + (vy4(i)-%fv(2*n+i))/a33(i)$

$ass tem=py1,py2,py3$
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac
$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
      a3=a13*one:a4=a22*one:
      a5=a23*one:a6=a33*one$

$calc  a51=a11*koe: a52=a12*koe$
$calc  a53=a13*koe: a54=a22*koe$
$calc a55=a23*koe: a56=a33*koe$

$calc  a61=a11*pnl: a62=a12*pnl$
$calc  a63=a13*pnl: a64=a22*pnl$
$calc   a65=a23*pnl: a66=a33*pnl$

$calc  a71=a11*par: a72=a12*par$
$calc  a73=a13*par: a74=a22*par$
$calc  a75=a23*par: a76=a33*par$

$calc a81=a11*foc: a82=a12*foc$
$calc a83=a13*foc: a84=a22*foc$
$calc a85=a23*foc: a86=a33*foc$

$calc a91=a11*spo: a92=a12*spo$
$calc a93=a13*spo: a94=a22*spo$
$calc a95=a23*spo: a96=a33*spo$

$calc a41=a11*fol: a42=a12*fol$
$calc a43=a13*fol: a44=a22*fol$
$calc a45=a23*fol: a46=a33*fol$
```

```
$calc a7=a1+a2+a3:a8=a4+a5:a9=a2+a3$

$ass p1=a7,a8,a6:p2=a9,a8,a6:
     p3=a3,a5,a6$

$calc a57=a51+a52+a53:a58=a54+a55:a59=a52+a53$

$ass pd1=a57,a58,a56:pd2=a59,a58,a56:
     pd3=a53,a55,a56$

$cal p1=p1:p2=p2:p3=p3:
     pd1=pd1:pd2=pd2:pd3=pd3$

$calc a67=a61+a62+a63:a68=a64+a65:a69=a62+a63$

$ass pd4=a67,a68,a66:pd5=a69,a68,a66$
$ass pd6=a63,a65,a66$

$calc a77=a71+a72+a73:a78=a74+a75:a79=a72+a73$

$ass pd7=a77,a78,a76:pd8=a79,a78,a76 $
$ass pd9=a73,a75,a76$

$cal pd4=pd4:pd5=pd5:pd6=pd6:
     pd7=pd7:pd8=pd8:pd9=pd9$

$calc a87=a81+a82+a83:a88=a84+a85:a89=a82+a83$

$ass pd10=a87,a88,a86:pd11=a89,a88,a86$
$ass pd12=a83,a85,a86$

$calc a97=a91+a92+a93:a98=a94+a95:a99=a92+a93$

$ass pd13=a97,a98,a96:pd14=a99,a98,a96$$
$ass pd15=a93,a95,a96$

$calc a47=a41+a42+a43:a48=a44+a45:a49=a42+a43$

$ass pd16=a47,a48,a46:pd17=a49,a48,a46$
$ass pd18=a43,a45,a46$

$calc pd10=pd10:pd11=pd11:pd12=pd12:pd13=pd13$
$calc pd14=pd14:pd15=pd15:pd16=pd16:pd17=pd17$
$calc pd18=pd18$
$endmac$


$use startup$use loop$
$use loop$
```

GLIM Directives for example 7.3.1

```
$c codes notation can be found at page 134$
$sle 10 $num n  $calc n=%nu/2$
$var n vy1 vy2 n1 x$
$data vy1 vy2 n1 x$
$read
0.00000000001  0.00000000001  40 1
2  0.00000000001  40 2
9  5  40 3
13 6  40 4
20 10 40 5

$calc vy3=n1-vy2-vy1$
$ass n2=n1,n1$
$var n w11 w22  w12
     a11 a12 a22
     py1 py2

$macro startup
$calc %lp=%eta=%sr(0) $calc %fv=%sr(0)/2$
$c calc %lp=%eta=%log(nw/(n2-nw)) $calc %fv=ystack$
$endmac$

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                    )/a11(i)$
$endmac

$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i2-n))$

$calc %fv=%fv*n2:%dr=1$
$endmac

$ass i=1...n $
$ass nx=x,x$
$$ass nw= vy1,vy2 $calc ystack=nw$

$yvar nw $error own  devcalc $link own newlink2
$scale 1.00$
$method * etacalc workvar$
```

```
$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 50 2  1.0e-4 $fit dp1+d1+dp2+d2-1$
$display e$
$extract %pe$
$endmac

$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
      w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):
      w12(i)=-%fv(i)*%fv(i+n)/n1(i)$
$calc  a11=%sqrt(w11) :a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2)$
$endmac

$macro workvar
$calc py1(i)=(vy1(i)-%fv(i))/a11(i):
   py2(i)=-(vy1(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
             +(vy2(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
         a4=a22*one $

$calc  a6=a11*x:a7=a12*x:
          a8=a22*x$

$ass d1=a1,zero: dp1=a2,a4$
$ass d2=a6,zero  :dp2=a7,a8 $
$endmac$

$use startup$use loop$
$use loop$


$macro like_del
$c some routines to calculate the log likelihhod kernel devc
   to test for removal of points ,
   also devc times 2 divided by p$
$c to use this, first use it with the overall fit.
    Then weight out points and fit again with loop
    and reuse this macro to see the effect of point deletion
$num devc devc2p$
$calc dumi=%gl(5,1)$
$tab the %fv total for dumi into tfvt$
$cal tfvt=n1-tfvt$

$ass fv=%fv, tfvt$

$ass yv=vy1,vy2,vy3$
$calc  devc = %cu(yv*%log(fv))$
$calc devc2p=devc*2/%pl$
$pr devc devc2p$
$ENDMAC$

$use like_del$
$look fv yv$
$print yv$
```

GLIM Directives for weighted out observation in example 7.3.1

```
$c codes notation can be found at page 134$
$sle 10 $num n  $calc n=%nu/2$
$var n vy1 vy2 n1 x$
$data vy1 vy2 n1 x$
$read
0.00000000001  0.00000000001  40 1
2  0.00000000001  40 2
9  5  40 3
13 6  40 4
20 10 40 5

$CALC VY3=N1-VY2-VY1$
$ass n2=n1,n1$
$var n w11 w22  w12
     a11 a12 a22 py1 py2

$macro startup
$calc %lp=%eta=%sr(0) $calc %fv=%sr(0)/2$
$c calc %lp=%eta=%log(nw/(n2-nw)) $calc %fv=ystack$
$endmac$

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                      )/a11(i)$
$endmac

$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i2-n))$

 $calc %fv=%fv*n2:%dr=1$
$endmac

$ass i=1...n $
$ass nx=x,x$
$$ass nw= vy1,vy2 $calc ystack=nw$

$yvar nw $error own  devcalc $link own newlink2
$scale 1.00$
$method * etacalc workvar$
$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 50 2  1.0e-4 $fit dp1+d1+dp2+d2-1$
$display e$
$extract %pe$endmac
```

```
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
      w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):
      w12(i)=-%fv(i)*%fv(i+n)/n1(i)$

$calc  a11=%sqrt(w11) :a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2)$
$endmac

$macro workvar
$calc py1(i)=(vy1(i)-%fv(i))/a11(i):
   py2(i)=-(vy1(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
             +(vy2(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
      a4=a22*one $

$calc  a6=a11*x:a7=a12*x:
        a8=a22*x$

$ass d1=a1,zero: dp1=a2,a4$
$ass d2=a6,zero  :dp2=a7,a8 $
$endmac$

$calc w=1$
$calc w(1)=w(6)=0$
$weight w$

$use startup$use loop$
$use loop$

$macro like_del
$c some routines to calculate the log likelihhod kernel devc
   to test for removal of points ,
    also devc times 2 divided by p$
$c to use this, first use it with the overall fit.
    Then weight out points and fit again with loop
    and reuse this macro to see the effect of point deletion
$num devc devc2p$
$calc dumi=%gl(5,1)$
$tab the %fv total for dumi into tfvt$
$cal tfvt=n1-tfvt$

$ass fv=%fv, tfvt$

$ass yv=vy1,vy2,vy3$
$calc  devc = %cu(yv*%log(fv))$
$calc devc2p=devc*2/%pl$
$pr devc devc2p$
$ENDMAC$

$use like_del$
$look fv yv$
$print yv$
```

GLIM Directives for example 7.3.2

```
$c codes notation can be found at page 134$
$sle 8 $num n  $calc n=%nu/2$
$var n vy1 vy2 n1 x$
$data vy1 vy2 n1 x$
$read
6  9  20 1
5  4  10 2
1  3  15 1
6  9  21 2

$CALC VY3=N1-VY2-VY1$
$ass n2=n1,n1$
$var n w11 w22  w12
     a11 a12 a22 py1 py2

$macro startup
$c calc %lp=%eta=%sr(0) $c calc %fv=%sr(0)/2$
$calc %lp=%eta=%log(nw/(n2-nw)) $calc %fv=ystack$
$endmac$

$num np1  twon  $calc np1=n+1:twon=2*n:
$ass j=1...%nu:i2=np1...twon $

$macro devcalc
$var n mu1 y1
$calc mu1(i)=(n2(i)-%fv(i)-%fv(i+n))$
$calc y1(i)=(n2(i)-ystack(i)-ystack(i+n))$
$calc y1=1/2*(y1*%log(mu1/y1))$
$ass y1stack=y1,y1$
$calc %di=ystack*%log(%fv/ystack)+y1stack$
$calc %di=-2*%di$
$calc %va=1
$endmac

$macro etacalc
$calc %eta(i2)=(%lp(i2))/a22(i2-n)$
$calc %eta(i)=(%lp(i)-a12(i)*%eta(i2)
                     )/a11(i)$
$endmac

$macro newlink2
$cal expta=%exp(%eta)$
$calc %fv(i)=
          expta(i)/(1+expta(i)+expta(i+n))$
$calc %fv(i2)=
         expta(i2)/(1+expta(i2)+expta(i2-n))$

 $calc %fv=%fv*n2:%dr=1$
$endmac

$ass i=1...n $
$ass nx=x,x$
$$ass nw= vy1,vy2 $calc ystack=nw$

$yvar nw $error own  devcalc $link own newlink2
$scale 1.00$
$method * etacalc workvar$

$Macro loop
$use cholesk$use workvar$use newdesign$
$cycle 50 2  1.0e-4 $fit d1+d2+dp1+dp2-1$
$display e$
$endmac
```

```
$macro cholesk
$calc w11(i)=%fv(i)*(1-%fv(i)/n1(i)):
      w22(i)=%fv(i+n)*(1-%fv(i+n)/n1(i)):
      w12(i)=-%fv(i)*%fv(i+n)/n1(i)$

$calc  a11=%sqrt(w11) :a12=w12/a11:
$calc   a22=%sqrt(w22-a12**2)$
$endmac

$macro workvar
$calc py1(i)=(vy1(i)-%fv(i))/a11(i):
   py2(i)=-(vy1(i)-%fv(i))*(a12(i)/a22(i))/a11(i)
             +(vy2(i)-%fv(n+i))/a22(i) $

$ass tem=py1,py2 $
$calc %wvd=tem+%lp$
$calc %wtd=1$
$endmac

$var n one zero$

$macro newdesign
$calc one=1:zero=0$
$calc a1=a11*one:a2=a12*one:
         a4=a22*one $

$calc  a6=a11*x:a7=a12*x:
          a8=a22*x$

$ass d1=a1,zero: dp1=a2,a4$
$ass d2=a6,zero  :dp2=a7,a8 $
$endmac$

$use startup$use loop$
$use loop$

$macro like_del
$c some routines to calculate the log likelihhod kernel devc
   to test for removal of points ,
    also devc times 2 divided by p$
$c to use this, first use it with the overall fit.
    Then weight out points and fit again with loop
    and reuse this macro to see the effect of point deletion

$num devc devc2p$
$calc dumi=%gl(4,1)$
$tab the %fv total for dumi into tfvt$
$cal tfvt=n1-tfvt$
$ass fv=%fv, tfvt$
$ass yv=vy1,vy2,vy3$
$calc  devc = %cu(yv*%log(fv))$
$calc devc2p=devc*2/%pl$
$pr devc devc2p$
$ENDMAC$

$use like_del$
```

GLIM Directives for example section 7.5

```
$c codes notation can be found at page 136$
$sle 5$data y1 y2 y3$
$read
0  0  40
2  0  38
9  5  26
13 6  21
20 10 10

$cal x=%gl(5,1)$

$num vl$calc vl=3*%sl$
$var vl case group w$
$ass freq=y1,y2,y3$

$calc case=%gl(%sl,1): group=%gl(3,%sl)$
$calc group=4-group$

$fact case %sl group 3$

$elim case$error p$yvar freq$

$fit group:+group*x(case)$
$di e$

$extract %cd %lv %di$
$look %fv %yv group freq$


$num i1 i2 i3 devr devf$
$calc devf=2*%cu(%yv*%log(%fv))/%pl$
$print devf$

$calc i1=1$

$macro fvs
$calc i2=i1+5:i3=i1+10$
$calc w=1:w(i1)=w(i2)=w(i3)=0$weigh w$

$fit .$
$di e$extract %pe$

$calc fv=%fv$

$calc fv(i1)=40*%exp(%pe(1)+%pe(4)*i1)
             /(1+%exp(%pe(1)+%pe(4)*i1)+%exp(%pe(2)+%pe(5)*i1))$

$calc fv(i2)=40*%exp(%pe(2)+%pe(5)*i1)
             /(1+%exp(%pe(1)+%pe(4)*i1)+%exp(%pe(2)+%pe(5)*i1))$

$calc fv(i3)=40
             /(1+%exp(%pe(1)+%pe(4)*i1)+%exp(%pe(2)+%pe(5)*i1))$

$calc devr=2*%cu(%yv*%log(fv))/4$

$print devr$
$endmac$


$use fvs$

$look %fv fv  %yv$
```

APPENDIX B

Weight matrix $W_{ij} = E(-\frac{\partial^2 l}{\partial \eta_i \partial \eta_j})$ **for *J=3* response variable**

The W_{ij} for J=3 level of response variable can be summarised in tabular form as follows:

∂/∂	$\eta_1^{(2)}$	$\eta_2^{(2)}$	.....	$\eta_m^{(2)}$	$\eta_1^{(3)}$	$\eta_2^{(3)}$	.....	$\eta_m^{(3)}$
$\eta_1^{(2)}$	$\mu_{12}(1-\frac{\mu_{12}}{n_1})$	0	....	0	$-\frac{\mu_{12}\mu_{13}}{n_1}$	0	.....	0
$\eta_2^{(2)}$	0	$\mu_{22}(1-\frac{\mu_{22}}{n_2})$	....	0	0	$-\frac{\mu_{22}\mu_{23}}{n_2}$	.....	0
.	.....				.....			
.	.....				.....			
$\eta_m^{(2)}$	0	0	.....	$\mu_{m2}(1-\frac{\mu_{m2}}{n_m})$	0	0	.....	$-\frac{\mu_{m2}\mu_{m3}}{n_m}$
$\eta_1^{(3)}$					$\mu_{13}(1-\frac{\mu_{13}}{n_1})$	0	.....	0
$\eta_2^{(3)}$					0	$\mu_{23}(1-\frac{\mu_{23}}{n_2})$	.....	0
.					.....			
.					.....			
$\eta_m^{(3)}$					0	0	.....	$\mu_{m3}(1-\frac{\mu_{m3}}{n_m})$

Derivatives of log-likelihood with respect to the logit link function

APPENDIX C

Weight matrix W_{ij} for $J=4$ response variable

W_{ij} for 4-level of response variable

∂/∂	$\eta_1^{(2)}$ $\eta_2^{(2)}$ ..... $\eta_m^{(2)}$	$\eta_1^{(3)}$ $\eta_2^{(3)}$ ..... $\eta_m^{(3)}$	$\eta_1^{(4)}$ $\eta_2^{(4)}$ ..... $\eta_m^{(4)}$
$\eta_1^{(2)}$	$\mu_{12}(1-\frac{\mu_{12}}{n_1})$ 0 0	$-\frac{\mu_{12}\mu_{13}}{n_1}$ 0 0	$\frac{\mu_{12}\mu_{14}}{n_1}$ 0 0
$\eta_2^{(2)}$	0 $\mu_{22}(1-\frac{\mu_{22}}{n_2})$ 0	0 $-\frac{\mu_{22}\mu_{23}}{n_2}$ 0	0 $-\frac{\mu_{22}\mu_{24}}{n_2}$ 0
$\vdots$	.....	.....	.....
$\eta_m^{(2)}$	0 0 ... $\mu_{m2}(1-\frac{\mu_{m2}}{n_m})$	0 0 ..... $-\frac{\mu_m\mu_{m3}}{n_m}$	0 0 ..... $-\frac{\mu_{m2}\mu_{m4}}{n_n}$
$\eta_1^{(3)}$		$\mu_{13}(1-\frac{\mu_{13}}{n_1})$ 0 0	$-\frac{\mu_{13}\mu_{14}}{n_1}$ 0 0
$\eta_2^{(3)}$		0 $\mu_{23}(1-\frac{\mu_{23}}{n_2})$ 0	0 $-\frac{\mu_{23}\mu_{24}}{n_2}$ 0
$\vdots$		.....	.....
$\eta_m^{(3)}$		0 0 ..... $\mu_{m3}(1-\frac{\mu_{m3}}{n_m})$	0 0 .... $-\frac{\mu_{m3}\mu_{m4}}{n_m}$
$\eta_1^{(4)}$			$\mu_{14}(1-\frac{\mu_{14}}{n_1})$ 0 0
$\eta_2^{(4)}$			0 $\mu_{24}(1-\frac{\mu_{24}}{n_2})$ 0
$\vdots$			.....
$\eta_m^{(4)}$			0 0 .... $\mu_{m4}(1-\frac{\mu_{m4}}{n_m})$

W_{ij} is a diagonal weight matrix namely I_{3n}

APPENDIX D

Alternative approach to find W_{ij}

Alternative approach to find $W_{ij} = \mathrm{E}(\dfrac{\partial^2 l}{\partial \eta_i \partial \eta_j})$

The standard approach used in this study to find the weight matrix is by using the log-likelihood function and the detail is given in section (4.3) but here we will give an alternate form to find W_{ij} . This approach is much easier to find any form of a derivative of the weight matrix for any link function. The derivatives given in section (4.3) are given by using the log - likelihood general approach but these derivatives can also be found very easily as follows:

$$\frac{\partial l_i}{\partial \eta_{i2}} = \frac{\partial l_i}{\partial \theta_{i2}} \times \frac{\partial \theta_{i2}}{\partial \eta_{i2}},$$

where $\dfrac{\partial l_i}{\partial \theta_{i2}}$ can be found from the equation (1.8) and since

$$\frac{\partial \eta_{i2}}{\partial \theta_{i2}} = \frac{ae^{a\theta_{i2}}}{a} = e^{a\theta_{i2}} = (1 + a\eta_{i2}) .$$

Or $$\frac{\partial \theta_{i2}}{\partial \eta_{i2}} = \frac{1}{1 + a\eta_{i2}} \quad \text{since} \quad \theta_{i2} = \frac{1}{a}\log(1 + a\eta_{i2})$$

Thus $$\frac{\partial l_i}{\partial \eta_{i2}} = \frac{\partial l_i}{\partial \theta_{i2}} \times \frac{1}{1 + a\eta_{i2}} = \frac{\partial l_i}{\partial \theta_{i2}} \times e^{-a\theta_{i2}} .$$

and $$\frac{\partial^2 l_i}{\partial^2 \eta_{i2}} = \frac{\partial}{\partial \theta_{i2}}\{\frac{\partial l_i}{\partial \theta_{i2}} \times e^{-a\theta_{i2}}\}\frac{\partial \theta_{i2}}{\partial \eta_{i2}}$$

$$= \{\frac{\partial^2 l_i}{\partial^2 \theta_{i2}} e^{-a\theta_{i2}} - ae^{-a\theta_{i2}} \frac{\partial l_i}{\partial \theta_{i2}}\} e^{-a\theta_{i2}}$$

$$= \{\frac{\partial^2 l_i}{\partial^2 \theta_{i2}} - a\frac{\partial l_i}{\partial \theta_{i2}}\} e^{-2a\theta_{i2}} = \{\frac{\partial^2 l_i}{\partial^2 \theta_{i2}} - a\frac{\partial l_i}{\partial \theta_{i2}}\} \frac{1}{(1+a\eta_{i2})^2}$$

Now $$\frac{\partial l_i}{\partial \theta_{i3}\partial \theta_{i2}} = \frac{\partial}{\partial \theta_{i3}}\{y_{i2} - \frac{n_i e^{\theta_{i2}}}{1 + e^{\theta_{i2}} + e^{\theta_{i3}}}\}$$

$$= -n_i e^{\theta_{i2}}\{-e^{\theta_{i3}}(1+e^{\theta_{i2}}+e^{\theta_{i3}})^{-2}\}$$

$$= n_i e^{\theta_{i2}} e^{\theta_{i3}} p_{i1}^2 = n_i p_{i1} e^{\theta_{i2}} p_{i1} e^{\theta_{i3}}$$

$$= \mu_{i2}\frac{\mu_{i3}}{n_i} = \frac{\mu_{i2}\mu_{i3}}{n_i}.$$

Therefore $$E(-\frac{\partial^2 l_i}{\partial^2 \eta_{i2}}) = E\{-\frac{\partial^2 l_i}{\partial^2 \theta_{i2}} + a\frac{\partial l_i}{\partial \theta_{i2}}\}\frac{1}{(1 + a\eta_{i2})^2}$$

$$= E\{-\frac{\partial^2 l_i}{\partial^2 \theta_{i2}}\}\frac{1}{(1 + a\eta_{i2})^2}$$

Since $$E\{a\frac{\partial l_i}{\partial \theta_{i2}}\} = 0.$$

Or $$E(-\frac{\partial^2 l_i}{\partial^2 \eta_{i2}}) = \mu_{i2}(1 - \frac{\mu_{i2}}{n_i})\frac{1}{(1 + a\eta_{i2})^2}.$$

Similarly $$E(-\frac{\partial^2 l_i}{\partial \eta_{i3}\partial \eta_{i2}}) = E(-\frac{\partial^2 l_i}{\partial \theta_{i3}\partial \theta_{i2}})\frac{1}{(1 + a\eta_{i2})(1 + a\eta_{i3})}$$

$$= -\frac{\mu_{i2}\mu_{i3}}{n_i}\frac{1}{(1+a\eta_{i2})(1+a\eta_{i3})}.$$

The other derivatives needed to complete the weight matrix can be found using own link function. This approach is much easier and can be applied in general for any link function to be considered.

References

[1] Andrew, D. F. and Pregibon, D. (1987): Finding the outliers that matter. *J. R. Statist. Soc, B,* **40**, 85-93.

[2] Abdelbasit, K. M. and Plackett, R. L. (1981): Experimental design for categorized data. *International Statistical Review*, **49**, 111-126.

[3] Abdelbasit, K. M. and Plackett, R. L. (1983): Experimental design for binary data. *J. Amer. Statist. Assoc.*, **78**, 90-98.

[4] Agresti, A. (1984): *Analysis of Ordinal Categorical Data*. John Wiley, New York.

[5] Agresti, A. (1990): *Categorical data analysis*. John Wiley, New York.

[6] Aitkin, M., Anderson, D. A., Francis, B. and Hinde, J. P. (1989): *Statistical modelling in GLI,* Oxford University Press, Oxford.

[7] Aitkin. M. A. and Francis, B. J. (1992): Fitting the multinomial logit model with continuous covariates GLIM. *Computational Statistics & Data Analysis*, **14**, 89-97.

[8] Archer, L., Hutchings, M., Ross, A. et al. (2003): *Higher Education and Social Class*. London; Falmer Routledge.

[9] Atkinson, A. C. (1981): Likelihood Ratios, Posterior odds and Information Criteria. *Journal of Econometrics*, **16,** 15-20.

[10] Atkinson, A. C. (1985): *Plots, Transformations and Regression*. Oxford Statistical Science Series, Oxford.

[11] Barnett, V. and Lewis, T. (1994): *Outliers Statistical Data*. John Wiley, New York.

[12] Bratton, K. A. and Ray, L. P. (2002): Descriptive representation, policy outcomes, and municipal day - care coverage in Norway. *American Journal of Political Science*, **46**, 428-437.

[13] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980): *Regression Diagnostis: Identifying Influential Data and Sources of Collinearity*. John Wiley, New York.

[14] Bennett, S. and Whitehead, J. (1981): Fitting logistic and log-logistic regression models to censored data using GLIM. *GLIM Newsletter*, **4**, 12-19.

[15] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975): *Discrete Multivariate Analysis. M. I. T.* Press, Cambridge Mass.

[16] Box, G. E. P. and Cox, D. R. (1964): An analysis of transformations. *J. R. Statist. Soc. B*, **26**, 211 - 251.

[17] Burn, R. (1982): Log-linear models with composite link functions in genetics. In *GLIM82, Proceedings International Conference Generalized Linear Models, Lecture Notes in Statistics* **14**. Gilchrist, R. (Ed), 144-154, Springer - Verlag, New York.

[18] Candy, S. G. (1986): Fitting a Parametric Log-linear Hazard Function to Grouped Survival Data. *GLIM Newsletter*, **13**, 28 - 31.

[19] Canes-Wrone et al.(2002): Out of Step, Out of Office: Electoral Accountability and house Members Voting. *American Political Review*, **96**, 127 - 140.

[20] Chatterjee, S. and Hidi, A. S. (1986): Influential observations, high-leverage points, and outliers in linear regression (with discussion). *Statistical Science,* **1**, 379 - 416.

[21] Chatterjee, S. and Hadi, A. S. (1988): *Sensitivity Analysis in Linear Regression.* John Wiley, New York.

[22] Collier, T., Gilchrist, R. and Phillips, D. (2003): Who plans to go to university? Statistical Modelling of potential working class participants. *Educational Research and Evaluation* 9, 239 - 263

[23] Cook, R. D. (1977): Detection of influential observations in linear regression. *Technometrics*, **19**, 15 – 19.

[24] Cook, R. D. and Weisberg, S. (1982): *Residuals and Influence in Regression*. Chapman Hall, New York.

[25] Cook, R. D. (1986): Assessment of local influence. *J. R. Statist. Soc. B,* **48**, 133-169.

[26] Copenhaver, T. and Mielke, P. (1977): Quantit analysis: a quantal assay refinement. *Biometrics*, Mar;33(1):175 - 86

[27] Cox, D. R. and Hinkley, D. V. (1968): A note on the efficiency of least-squares estimates. *J. R. Statist. Soc. B*, **30**, 284 - 289.

[28] Cox, D. R. (1970): *The Analysis of Binary Data*. Chapman Hall, London.

[29] Cox, D. R. (1972): Regression models & life tables. *J. R.. Statist. Soc. B*, **34**, 87-203.

[30] Cox, D. R. (1975): Partial likelihood, *Biometrika*, **64,** 269 - 276.

[31] Cox. D. R. and Snell, E. J. (1989): *The Analysis of Binary Data.* (2nd Ed.) Chapman Hall, London.

[32] Crichton, N. and Hinde, J. (1992): Investigation of an ordered logostic model for consumer debt. *In proceedings of GLIM92 and the 7th International Workshop on Statistical Modelling*, Lecture Notes in Statistics, **78**, Springer Verlag, Berlin.

[33] Critchley, F., Atkinson, R. A., Lu, G., and Biazi, E. (2001): Influence analysis based on the sensitivity function. *J. R. Statist. Soc.* **63**, 307 - 323.

[34] Critchley, F. and Vitietto, C. (1991): The influence of observations on the misclassification probability estimates in linear discriminate analysis. *Biometricka*, **78,** 677 - 690.

[35] Daniel, C. and Snell, E. J. (1971): *Fitting equations in data*. John Wiley, New York.

[36] Davidian, M. and Carroll, R.(1987): Variance function estimation. *J. Amer. Statist. Assoc.*, **82**, 1079-1091.

[37] Davidian, M. and Carroll, R. (1988): A note on extended quasi-likelihood. *J. R. Statist. Assoc. B,* **50**, 74-82.

[38] Davis, R. A., Dunsmuir, W. T. M. and Wang, Y. (1999): Modelling Time Series of Count Data. In: Ghosh, S. (ed.), *Asymptotics, Nonparametrics and Time Series*, chapter 3, 63-113, Marcel Dekker.

[39] Decarli, A., Francis, B. J. , Gilchrist, R. and Seeber, G. U. H. (Eds) (1989): *Statistical Modelling: Proceedings of the GLIM89 and the 4'International Workshop on Statistical Modelling*. Lecture Notes in Statistics, **57**, Springer - Verlag, Berlin.

[40] Demiror, D., Guvenir, H. A. and Itler, N. (1998): Learning differential diagnosis of erythemato - squamous diseases using voting feature interval. *Artificial Intelligence in Medicine* 13, 147-165.

[41] Dempster, A. P. and Gasko-Green, M. (1981): New tools for residual analysis. *The Annals of Statistics*, Vol.9 **5**, 945 - 959.

[42] Dobson, A. J. (1990): *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

[43] Draper, N, R. and Smith, H. (1981): *Applied Regression Analysis.* (2nd Ed.) John Wiley, New York.

[44] Dunn, P. K. and Smyth, G. K. (1996): Randomized Quantile Residuals. *Journal Computational Graph, Statist.* **5**, 236 - 244.

[45] Efron, B. (1982): Nonparametric estimates of standard error: the jacknife, the bootstrap and other methods. *Biometrika,* **68**, 589.

[46] Ellis, S. P. and Morgenthaler, S. (1992): Leverage and breakdown in L_1 regression. *J. Amer. Statist. Assoc.*, **87**, 143 - 148.

[47] Everitt, B. S. (1977): *The Analysis of Contingency Tables.* Chapman &Hall, London.

[48] Fahrmeir, L. and Klinger, J. (1994): Estimation and testing generalized linear models under inequality restrictions. *Statistical papers* **35,** 211 - 229.

[49] Fahrmeir, L. and Tutz, G. (1994): *Multivariate Statistical Modelling.* Springer-Verlag, New York.

[50] Fahrmeir, L., Francis, B., Gilchrist, R.,and Tutz, G. (1992) (Eds): *Advances in GLIM and Statistical Modelling. Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling,* **78**, 6-12.

[51] Fox, John (2002): *An R and S-Plus Companion to Applied Regression.* Sage Publications, International Educational and Professional Publisher.

[52] Flowerdew, R. and Aitkin, M. (1982): A method of fitting the gravity model based on the Poisson distribution. *Journal Regional Science* **22**, 191-202.

[53] Francis, B. Green, M. and Clarke, M. (1992): Model fitting Application in GLIM4. *In Proceedings of GLIM92 and the 7th International Workshop On Statistical Modelling.* Lecture Notes in Statistics, **78,** Springer Verlag, Berlin.

[54] Francis, B., Green, M. and Payne, C. (eds) (1993): *GLIM 4: The Statistical System for Generalized Linear Interactive Modelling.* Clarendon Press, Oxford.

[55] Francis. B. Green, M and Bradley, M. (1990): GLIM4 - developments in model fitting. In: *Proceedings of the 9th Symposium in Computational Statistics, COMPSTAT90,* Physica-Verlag, Heldelberg.

[56] Gilchrist, R. (Ed.) (1982): GLIM 82: *Proceedings of the International Conference on Generalized Linear Models.* Springer-Verlag, New York.

[57] Gilchrist, R. and Scallan, A. J. (1984): Parametric Link Functions in Generalized Linear Models. *In: COMPSTATS 84*, Wien., 203-208.

[58] Gilchrist, R., Francis, B. and Whittaker, J. (Ed.) (1985):*Proceedings of International Conference on Generalized Linear Models. Springer-Verlag*, New York.

[59] Gilchrist, R. (1981): Estimation of the parameters of the gamma distribution by means of a maximal invariant. *Commun. in Statist.* **11**, 1095-1110.

[60] Green, P. J. (1984): Iteratively re-weighted least squares for maximum likelihood estimation and some robust and resistant alternatives.*J. R. Statist. SocB,***46**,149-192.

[61] Green, M. Francis, B. and Bradley, M. (1989): GLIM4 - structure and development in Statistical modelling-*Proceedings of GLIM89 and the 4th International Workshop On Statistical Modelling.* Lecture Notes in Statistics, **57,** Springer-Verlag, Berlin.

[62] Hadi, A. S. (1992): Identifying multiple outliers in multivariate data. *J. R. Statist. Soc.* **54**, 761-771.

[63] Hadi, A. S. (1992): A new measure of over potential influence in linear regression. *Computational Statistics and Data Analysis* **14**, 1-27.

[64] Hoaglin, D. C. and Welsch, R. E. (1987): The hat-matrix in regression and ANOVA. *The American Statistician* **32**, 17-22.

[65] Hinkley, D. V., Reid, N. and Snell, E. J. (Ed.) (1990): *Statistical Theory and Modelling.* Chapman Hall, London.

[66] Huber, P. J. (1981): *Robust Statistics*. John Wiley, New York.

[67] Hutchison, D. (1986): Ordinal Variable Regression Using the McCullagh (proportional Odds) Model. *GLIM Newsletter*, **9**, 9-17.

[68] Jackson, J. E. (2002): A seemingly unrelated regressionmodel for analyzing multiparty elections, *Political analysis* **10**(1), 49-65.

[69] Jϕrgensen, B.(1987):Exponential dispersion models. *J R. Statist.Soc.B* **49**, 127-162.

[70] Katz, J. N. and King, G. (1999): A statistical model for multinomial electoral data. *American political science review* **93**, 15-32.

[71] Kendal, M. and Stuart, A. (1967):*The Advance Theory of Statistics.* Griffin, London.

[72] Lawless, J. F. (1987): Negative binomial and mixed Poisson Regression. *Scand. J. Statist.* **15,** 209-225.

[73] Lawrance, R. J. (1995): Deletion influence and masking in regression. *J. R. Statist. Soc.* **57**, 181- 189.

[74] Lee, Y. and Nelder, J. A. (1996): Hierachical Generalized Linear Models (with Discussion). *J. R. Statist. Soc. B* **58,** 619-679.

[75] Lindsey, J. K. (1989): *Analysis of Categorical Data using GLIM.* Lecture Notes in Statistics, **56**, Springer-Verlag, Berlin.

[76] Lindsey, J. K. (1991): *The Analysis of Stochastic Processes Using GLIM.* Lecture Notes in Statistics, **72**, Springer-Verlag, Berlin.

[77] Lu, J., Ko, D. and Chang, T. (1997): The standardised influence matrix and its applications. *J. Am. Statist. Ass.*, **92**, 1572-1580.

[78] McCullagh, P. (1980): Regression models for ordinal data. *J. R. Statist. Soc. B* **42**, 109-142.

[79] McCullagh, P. (1983): Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

[80] McCullagh, P. and Nelder, J. A. (1989): *Generalized Linear Models.* (2nd Ed.) Chapman Hall, London.

[81] McCulloch, C. E. and Meeter, D. (1983): Discussion of ''outlier...s'', by Beckman R.J. and Cook R. D. *Technometrics* **25**, 234-242.

[82] McDonagh, E. (2002): Political citizenship: The gender paradox, *American Political Science Review* **96**, 535-552.

[83] Monroe, B. L. and Ross, A. G. (2002): Electoral systems and unimagined consequences: partisan effects of distracted proportional representation. *American Journal of Political Science* **46**, 67-89.

[84] Morgan, B. (1985): The cubic logistic model for quantal assay data. *Applied Statistics* 34:22, 105-113.

[85] Owen, A. (1988): Empirical likelihood ratio confidence intervals for a single function. *Biometrika* **75**, 237- 249.

[86] Payne, C. D. et al. (1993): *The GLIM System Release 4.* Oxford University Press, Oxford.

[87] Pregibon, D. (1980): Goodness of link tests for generalized linear models. *Appl. Statist.* **29**, 15-24.

[88] Pregibon, D. (1981): Logistic regression diagnostics. *Ann. Statist.* **9,** 705-724.

[89] Pregibon, D. (1982): Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **32**, 485-498.

[90] Pregiban, D. (1984): Data analytic methods for matched case – control studies. *Biometrics* **40**, 639- 651.

[91] Preisser, J. S. and Qaqish, B. F. (1996): Deletion diagnostics for generalised estimating equations. *Biometrics* **83**, 551 – 562.

[92] Preisser , J. S. , Qaqish, B. F. and Perin J. (2008) : Miscellanea ; A note on deletion diagnostics for estimating equations. *Biometrika* **95**, 509 – 513.

[93] Qaqish, B. F. (2003); A family of multivariate binary distributions for simulating correlated binary variables woth specified marginal means and correlations. *Biometrika* **90**, 455 - 463.

[94] Reese, R. A. (1980): Status of the GLIM Library at the University of Hull. *GLIM Newsletter*, **2**, 12-13.

[95] Rocke, D. M. and woodruff, D. L. (1996): Identification of outliers in multivariate data. *J. Am. Statist. Ass.* **91**, 1047-1061.

[96] Rousseeuw, P. J. (1984): Least median of squares regression. *J. Am. Statist. Ass.,***79**, 871-880.

[97] Scallan, A., Gilchrist, B. and Green, M. (1984): Fitting parametric link functions in generalized linear models. *Comput. Statist. Data Ana.* **2**, 37-49

[98] Schafer, D. W. (1987): Covariate measurement errors in generalized linear models. *Biometrika* **74**, 385-391.

[99] Searle, S. R. (1982): *Matrix algebra useful for statistics*. John Wiley, New York.

[100] Seber, G. A. F. (1977): *Linear Regression Analysis*. John Wiley, New York.

[101] Sherman, J. and Morrison, W. J. (1950): Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statist.* **21,** 124 - 127.

[102] Stukel, T. (1988): Generalizes Logistic models. *J. Amer Statist. Assoc.* **83**, 426-431.

[103] Swan, A. V. and Francis, B. J. (1992): Medical Application in GLIM4. In *Proceedings of GLIM92 and the 7th International Workshop On Statistical Modelling.* Lecture Notes in Statistics, **78,** Springer-Verlag, Berlin.

[104] Tanabe, K. and Sagae, M. (1992): An exact Cholesky decomposition and the generalized inverse on the variance-covariance matrix of multinomial distribution, with applications. *J. R. Statist. Soc. B*, **54**, 211-221.

[105] Theil, H. (1965): The analysis of disturbances in regression analysis. *J. Amer. Statist . Assoc.* **60**, 1067-1079.

[106] Tomz, M., Tucker, J. A. and Wittenberg, J. (2202).: An easy and accurate regression model for multiparty electoral data. *Political analysis* **10**(1) , 66-83.

[107] Velleman, P. F. and Welsch, R. E. (1981): Efficient computing of regression diagnostics. *The Amercian Statistician* **35**, 234-242.

[108] Wang, P. C. (1987): Residual plots for detecting nonlinearity in generalized linear models. *Technometrics* **29**, 435-438.

[109] Walter, R. M. and Jaseet, S. S. (2004): Robust estimation and outlier detections for over - dispersed multinomial models for count data. *American Journal of Political Science* **vol, 46, No.2**, April 2004. 392-411

[110] Wedderburn, R. W. M. (1974): Quasi-likelihood function, generalized linear models and Gauss-Newton method. *Biometrik* **64**, 439-447.

[111] Weisberg, S. (1983): Some principles for regression diagnostics and influence analysis, discussion of paper by Hocking, R. R. *Technometrics* **25**, 240-244.

[112] Weisberg, S. (1985): *Applied Linear Regression.* John Wiley, New York.

[113] Welsch, R. E. (1982): *Influence functions and regression diagnostics. In Modern Data Analysis* (eds R. L. Launer and A. F. Siegel). Academic Press, New York.

[114] Welsch, R. E. (1983): Discussion of developments in linear regression methodology: 1959-1982, by Hocking, R. R. *Technometrics* **25**, 245-246.

[115] Whitehead, J. (1980): Fitting Cox's regression model to survival data using GLIM. *Applied Statistics* **29**, 268-275.

[116] Wetherill, G. B.(1986): *Regression Analysis applications*. Chapman &Hall, London.

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

[118] Williams, D. A. (1983): The use of the deviance to test the goodness of fit of a logistic-linear model in binary data. *GLIM Newsletter* **6**, 60-62.

[119] Williams, D. A. (1984): Residuals in generalized linear models. Proceedings of the 12th International *Biometrics* Conference Tokyo, 59-68.

[120] Williams, D. A. (1987): Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* **36**, 181-191.