TITLE Some Applications of Generalised Linear Models

AUTHOR Anthony Scallan

INSTITUTION and DATE Polytechnic of North London
July 1990
(CNAA)

Some Applications of Generalised Linear Models

Anthony Scallan

Thesis submitted in partial filfilment of the CNAA requirments for the degree of Doctor of Philosophy.
Polytechnic of North London
July 1990

Some applications of Generalised Linear Models

Anthony Scallan

Abstract

This thesis is concerned with some extensions to and applications of generalised linear models and their implementation in a statistical package. The principal extension considered is the inclusion of extra parameters in the link function of the model in order to create a family of *parametric* link functions. This technique is applied to standard link functions as well as to the family of composite link functions. The applications of such models are illustrated by reference to several examples. The techniques presented enable complicated models to be fitted in a unified and consistent manner without the need for specialist software or algorithms.

A two-stage algorithm for fitting parametric link functions is presented and a diagnostic procedure applied to this class of extended models. The applications of such models include the analysis of grouped and multivariate data. It is shown that grouped data arising from a truncated or mixture distribution can be represented as a parametric composite link function and the technique applied to extend the analysis of some previously published data sets. Following a transformation, it is shown that certain time series models may modelled using parametric composite link functions. An algorithm is presented for the fitting of such models in which the variance function of the observations may be a quite general function of the mean. A generalisation of the multivariate logistic distribution is introduced with application to the analysis of repeated measurements data.

Finally, the results of an investigation into the possible development of a statistical programming language, with particular reference to the fitting of generalised linear models, are considered. An implementation of such a language is reported and some features of the language illustrated.

Acknowledgements

Special thanks are due to a friend and mentor, Robert Gilchrist, who has been a source of inspiration and encouragement during the period that the work on this thesis was undertaken.

Thanks are also due to my wife Heather for her patience, understanding and encouragmant during the writing of this thesis.

I also gratefully acknowledge the financial support at various times of the Polytechnic of North London, the Economic and Social Research Council and Lancashire Polytechnic.

Contents

Chapter 1

Introduction

1.1 Opening comments

This thesis is concerned with the implementation and practical application of several techniques for fitting parametric models to data. Many of the techniques considered are not new, although it is hoped that they are implemented in such a way that it enables previously complicated analyses to be carried out in a simple and unified manner.

The driving force behind all these implementations has been the unifying idea of generalised linear models and the statistical package specifically written for their practical application, *GLIM* . Generalised linear models were first presented in a unified manner by Nelder and Wedderburn, [49],

who showed that many existing data analytic methods could be handled in a similar way by using a weighted least squares technique; although, it should be noted that, many other authors had worked along similar lines, see, for example, Dempster, [21]. In this context we take analysis to mean the fitting of a parametric model to data and the estimation of the model parameters by the technique of maximum likelihood. The most common example of such a model is the familiar linear regression model defined by,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n$$

Here, the ith observation, Y_i, is assumed to be linearly related to a set of regressor or explanatory variables, $x_1, \ldots, x_p$, while the ϵ_i represent a set of random error terms, usually assumed to independently distributed as $\mathcal{N}(0, \sigma^2)$. In this case, the procedure for the estimation of the unknown parameter vector, $\underline{\beta}$, is least squares and is equivalent to that of maximum likelihood.

Generalised linear models, as first introduced, extend this setup to include such diverse applicatons as log-linear modelling for the analysis of contingency tables and the analysis of proportions via probit and logit models.

The computer package, GLIM, Payne et. al., [51], first appeared in

1974 and was designed specifically for the analysis of such generalised linear models. This was achieved through a powerful and common command structure which in many respects mirrored the generality of the definition of generalised linear models. Although the package contained some facilities for basic programming structures, such as looping and branching, it is unlikely that the authors could have anticipated the explosion in research and applications which followed the release of the package. It is our feeling that the availability of GLIM has been a strong motivating force for much of the research in this area of applied statistics. In this respect, we would regard GLIM as a statistical language rather than merely as a statistical package. Although similar programming facilities appear in other packages, for example, MINITAB has a limited macro facility, none seem to have fired the imagination of so many statisticians as GLIM.

The principal model fitting algorithm in GLIM is the procedure of iteratively re-weighted least squares (IRLS). Thus, much research and ingenuity has been directed towards showing that various diverse statistical estimation problems can be expresed as one solveable using IRLS. This approach can obviously be criticised on the grounds that the IRLS approach may not be the most efficient or the most intuitively obvious. However, it can conversely be argued that least squares procdures are very familiar to statisticians, as

they form such a fundamental part of applied statistical techniques, and an unusual model expressed in such a form may be made more accessible to a wider audience than would otherwise be the case. Coupled with the simple yet powerful command structure of GLIM, it is our feeling that, in many cases, this approach, while losing little in aesthetic quality, gains much in accessibility.

The subject matter of this thesis consists largely of considering estimation problems which, in themselves, are not necessarily novel, but have previously required specialist software or subroutines to enable their application. By utilising the IRLS approach, we illustrate how they may be expressed in a form which enables the models to be fitted using the GLIM package.

In Chapter 2 we review the basic formulation and properties of generalised linear models and illustrate the derivation of the IRLS algorithm. In Chapter 3 we discuss two important extensions to the basic definition of a generalised linear model. In Chapter 4 we present a method for fitting models to data that has been grouped or truncated and compare this approach with other techniques. Chapter 5 considers the somewhat neglected topic of the analysis of multivariate data and illustrates how such models may be expressed as generalised linear models. In particular we look at the

fitting of time series models and introduce a specific model for the analysis of repeated measures data which may be fitted using standard techniques. The concluding chapter of this thesis discusses the requirements of a new statistical computing language and outlines an investigation into the format of a possible candidate based on the ideas of functional programming.

Throughout the thesis, applications are illustrated by practical examples many of which have previously been analysed in the literature. The GLIM coding used in our analysis of the examples is presented in the Appendices.

1.2 Notational Conventions

Some of the notation we shall use has already been introduced. Throughout this thesis an attempt has been made to conform to a notational style which seems to have become conventional over the past few years for the majority of papers in this subject area.

We use the standard symbols for the mathematical operations of summation ($\sum$) and integration ($\int$). The derivative of a function $a(\theta)$ with respect to (w.r.t.) θ is denoted by $\partial a(\theta)/\partial\theta$ or by $\dot{a}(\theta)$. Repeated differentiation is denoted by $\partial^2 a(\theta)/\partial\theta^2$ or $\ddot{a}(\theta)$ etc.

Parameters to be estimated by the data are denoted by greek letters

such as α, β, and γ. The estimated values of the parameters are denoted by attaching a marking to the parameter such as $\hat{\beta}$.

Random variables are denoted by capital roman letters such as X,Y etc. We use the symbol $\sim$ to denote that a random variable follows a particular distribution, e.g. $X \sim \mathcal{N}(\mu, \sigma^2)$. The expectation operator E[ ] is used to denote the expectation taken over some distribution, e.g. $E[Y] = \int y f(y;\theta) dy$.

Vectors are denoted by lower case roman letters and are underlined, e.g. $\underline{x}$. Matrices are denoted by upper case roman letters, e.g. X. Where needed, the dimension of a vector or matrix is given in the form $n \times p$, for example.

Chapter 2

GENERALISED LINEAR MODELS

In this chapter we provide a basic overview of generalised linear models and introduce some concepts and results which will be utilised in later sections.

In most modelling situations, it is usual to think of an observed or response variable as consisting of two parts:

- a random component or error term
- a systematic component of explanatory variables

In the classical linear model, it is assumed that the data arise from a

model of the form, $y_i = \mu_i + \epsilon_i, \quad i=1,\ldots,n,$ where $\mu_i = \underline{z}_i^T \underline{\beta}$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ for i=1,...,n.

Such a formulation is clearly not appropriate in many situations, for example when observations take on only integer values or when data arise from a process which is clearly non-Normal. Moreover, the relationship between the mean, $\underline{\mu}$, and the linear predictor, $\underline{\eta} = X\underline{\beta}$, may not necessarily be linear.

Nelder and Wedderburn, [49], proposed a class of models which generalised the classical linear model to cope with data from a variety of non-Normal error distributions and non-linear mean/linear predictor relationships. Such models are termed generalised linear models and are conveniently defined by the following two components:

- A family of error distributions $f_Y(y;\theta)$ representing the random or error component of the model.
- A link function $\eta = g(\mu)$, relating the systematic linear predictor to the mean of the distribution.

In the following sections we discuss in more detail the two components of a generalised linear model and illustrate how the maximum likelihood estimates of the parameters in such a model may be found using iteratively

re-weighted least squares (IRLS).

2.1 The Exponential Family

A random response variable, Y, with density $f_Y(y;\theta,\phi)$, is said to be a member of a univariate exponential family if its density function can be represented in the form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}, \tag{2.1}$$

where the range of Y does not depend on θ. The parameter θ is called the natural parameter and ϕ the scale parameter. Such families include the Normal, Gamma, Poisson, Binomial and Inverse Gaussian distributions. The parameter θ is related to the mean and variance of the response variable by

$$\begin{aligned} \mathrm{E}[Y] &= \mu &&= \dot{b}(\theta) \\ \mathrm{Var}(Y) &= \mathrm{E}[(Y-\mu)^2] &&= \phi\ddot{b}(\theta) \end{aligned} \tag{2.2}$$

These well known properties follow from differentiating under the integral sign with respect to θ.

A list of some of the distributions mentioned above is given in Table 2.1. In each case we give the form of θ and $b(\theta)$, and the normalising constant. Note that, for the Normal and Gamma distributions, it is conventional to

Distribution	θ	$b(\theta)$	$c(y,\phi)$
Normal, $Y \sim \mathrm{N}(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2}(\frac{y^2}{\sigma^2}+\ln(2\pi\sigma^2))$
Poisson, $Y \sim \mathrm{Pois}(\mu)$	$\ln(\mu)$	$\exp(\theta)$	$-\ln(y!)$
Binomial, $Y \sim B(n,p)$	$\ln\left(\frac{\mu}{n-\mu}\right)$	$\ln(1+e^{\theta})$	$\ln({}^{n}C_{y})$
Gamma, $Y \sim G(\mu,\nu)$	$-1/\mu$	$-\ln(-\theta)$	$\nu\ln(\nu y)-\ln(y)-\ln(\Gamma(\nu))$

Table 2.1: Exponential Families

use a different parameterisation for the scale parameter than that given by Equation 2.1. In particular, for the Gamma distribution, the parameterisation used in Table 2.1 has $\mathrm{var}(Y) = \mu^2/\nu$.

Wedderburn , [73], introduced the concept of quasi-likelihood as a further generalisation of the assumption that observations are a member of an exponential family. Briefly, given independent observations y_i with means μ_i and variances $V(\mu_i)$ we assume the quasi-likelihood , $K(y_i,\mu_i)$ is defined as the solution to the differential equation

$$\frac{\partial K(y_i,\mu_i)}{\partial \mu_i} = \frac{y_i-\mu_i}{V(\mu_i)}$$

Wedderburn showed that, by regarding this equation as a score function and solving the set of equations $\partial K/\partial \underline{\mu} = 0$, maximum *quasi-likelihood (MQL)* estimates could be calculated which share many of the properties

of maximum-likelihood estimates. Indeed, for certain variance functions, $V(\mu_i)$, the quasi-likelihood corresponds to the kernel of a the log-likelihood of a member of an exponential family, so that, in these cases, maximum likelihood and maximum quasi-likelihood correspond. See also McCullagh, [43], for a discussion of the theoretical properties of MQL estimation for the non-independence case. The efficiency of quasi-likelihood estimation for certain models has been considered by Hill, [37], and Firth, [23].

Another generalisation of this formulation is to consider functions of the form $a_i(\phi)$, i=1,...,n, so that the scale parameter can vary for each observation. In particular we can consider functions of the form $\frac{\phi}{w_i}$, where the w_i are known prior weights. However, recent authors, for example Aitkin, [3] and Smyth, [67] , have considered modelling the scale parameter, ϕ, as a function of a set of covariates in the same way as the mean. This is accomplished via a two-way algorithm, in which the parameters in the linear predictor for the mean are estimated for fixed values of the scale parameters, and then the parameters for the scale estimated for fixed values of the mean. Iteration in this algorithm proceeds until both sets of parameters have converged to their final values.

Such two-way or see-saw algorithms will occur throughout this thesis as many of the models to be considered have a natural separation, in terms of

sets of parameters, of the type discussed above.

2.2 The Systematic Component

For a given set of covariates, $\mathbf{x}_i$, the relationship between the mean μ_i and the linear predictor, $\eta_i = \mathbf{x}_i^T\beta$, is defined by the link function as

$$\eta_i = g(\mu_i),$$

where g is assumed to be monotonic and twice differentiable, ensuring that g^{-1} exists. The classical linear model has g as the identity function.

An important special case is when the link function is defined in terms of the natural parameter, θ. This leads to the link function. $\eta = \theta = b^{-1}(\mu) = g(\mu)$, which is known as the 'natural' link function for that family of distributions indexed by θ. For the Binomial distribution, where we typically observe y_i successes out of n_i trials , with probability of success, p_i, the natural link function is the logit defined by,

$$\eta_i = \ln\left(\frac{p_i}{1-p_i}\right)$$

and $\mu_i = n_i p_i$.

The use of such natural link functions is both intuitively and theoretically appealing. In many cases, the natural link leads to a simple interpretation

of the relationship between the mean and linear predictor. For example, the logit link described above immediately leads to the use of odds and log-odds, while the natural link for the Poisson distribution, that is the natural logarithm, leads to the analysis of cell expectations in contingency tables expressed as sums of row and column effects. The estimates derived from the use of natural link functions also have theoretically desirable properties which are discussed further in Section 2.4.

Link functions may be generalised in two major ways. Firstly, the function may depend on extra unknown parameter(s) not contained in the linear predictor, thus creating a "parametric" family of link functions. An example is the Box-Cox link with unknown exponent defined by,

$$\eta = \begin{cases} \frac{\mu^{\alpha}-1}{\alpha} & \alpha \neq 0 \\ \ln(\mu) & \alpha = 0 \end{cases} \tag{2.3}$$

Of interest is the estimation of the unknown parameter α and, hence, the most appropriate form of link function suggested by the data.

The second major generalisation is to assume that the mean of an observation may be related to more than one linear predictor. Such models were proposed originally by Thompson and Baker, [71] and are known as *composite* link functions. In following chapters we show how such link func-

tions may be used to specify several non-standard models in the generalised linear model framework and, hence, be fitted using standard techniques for GLM's.

2.3 Fitting Generalised Linear Models

In this section we first describe the basic technique of *iteratively reweighted least squares*, IRLS, which may be used to fit a generalised linear model. We then describe how this techinique can be extended in different ways to fit various non-standard generalised linear models. In order to fit a generalised linear model, the following components must first be specified,

- the frequency (error) distribution of the response variable, y,

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

- the set of explanatory variables which form the linear predictors for each observation, i.e. the matrix X such that $\underline{\eta} = X^T\underline{\beta}$.

- the link function relating the mean of the distribution of y to the linear predictor, i.e. $\eta = g(\mu)$.

By fitting such a model we mean the estimation of the vector of unknown parameters, $\underline{\beta}$. Note that, it is assumed the matrix X is of full rank,

otherwise there is an obvious non-uniqueness in the estimates for $\underline{\beta}$.

The estimation techniqe employed is maximum likelihood, which, as well as being intuitively appealing, has many desirable and tractable theoretical properties. We will outline the iteratively reweighted least squares estimation procedure for a standard generalised linear model, described above, and then discuss extensions to such procedures for more general models.

2.3.1 Iteratively Reweighted Least Squares

The log likelihood function for a sample of independent observations $y_1, \ldots, y_n$ from an exponential family with systematic component described above is,

$$l(\underline{\beta}; \underline{y}) = \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\}/\phi + \sum_{i=1}^{n} c(y_i, \phi) \qquad (2.4)$$

and its derivative, the *score function*, is

$$s(\underline{\beta}) = \frac{\partial l}{\partial \underline{\beta}} = \sum_{i=1}^{n} \left\{y_i - \dot{b}(\theta_i)\right\} \frac{\partial \theta_i}{\partial \underline{\beta}}/\phi \qquad (2.5)$$

Use of the chain rule and results derived in section 2.1 above gives,

$$\begin{aligned} \frac{\partial \theta_i}{\partial \beta} &= \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta} \\ &= \frac{1}{\ddot{b}(\theta_i)}\frac{1}{\dot{g}(\mu_i)}x_i \end{aligned}$$

and

$$\phi\ddot{b}(\theta_i) = \mathrm{var}(Y_i) \;=\; V_i$$

$$\dot{b}(\theta_i) = \mathrm{E}[y_i] \;=\; \mu_i$$

so that the score function can be written as

$$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i) x_i / V_i \delta_i \qquad (2.6)$$

where $\delta_i = \dot{g}(\mu_i)$. In order to find the root of this equation using Fisher scoring,we need the expected second derivative of the log-likelihood. It is well known that,

$$\begin{aligned} E\left[\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right] &= -E\left[\frac{\partial l}{\partial\beta}\frac{\partial l}{\partial\beta^T}\right] \\ &= -E\left[\sum_{i=1}^{n}(y_i - \mu_i)x_i/V_i\delta_i \;\; \sum_{j=1}^{n}(y_j - \mu_j)x_j^T/V_j\delta_j\right] \end{aligned}$$

Since the y_i are assumed independent, the expectations of $(y_i - \mu_i)(y_j - \mu_j)$ are zero for $i \neq j$. Thus,

$$\begin{aligned} E\left[\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right] &= -\sum_{i=1}^{n} x_i x_i^T / V_i \delta_i^2 \\ &= -\sum_{i=1}^{n} w_i x_i x_i^T \\ &= -X^T W X \end{aligned}$$

where W is a diagonal matrix with elements w_i given by

$$w_i \;=\; \frac{1}{V_i \delta_i^2}$$

Using the Fisher-scoring algorithm, the (r+1)th estimate of $\underline{\beta}$ is obtained from the previous estimate via the relation,

$$\underline{\hat{\beta}}_{r+1} = \underline{\hat{\beta}}_r + (X^T W_r X)^{-1} s(\underline{\hat{\beta}}_r)$$

where all quantities on the right hand side are evaluated at $\underline{\hat{\beta}}_r$. Now, rewriting Equation 2.6 as $s(\underline{\beta}) = X^T W \underline{u}$, where $u_i = (y_i - \mu_i)\delta_i, i = 1, \ldots, n$, the updated estimate of $\underline{\beta}$ can be found using,

$$\begin{aligned}
\underline{\hat{\beta}}_{r+1} &= \underline{\hat{\beta}}_r + (X^T W_r X)^{-1} X^T W_r \underline{u}_r \\
&= (X^T W_r X)^{-1} (X^T W_r X \underline{\hat{\beta}}_r + X^T W_r \underline{u}_r) \\
&= (X^T W_r X)^{-1} X^T W_r (X \underline{\hat{\beta}}_r + \underline{u}_r) \\
&= (X^T W_r X)^{-1} X^T W_r \underline{z}_r
\end{aligned}$$

where,

$$\begin{aligned}
\underline{z}_r &= X \underline{\hat{\beta}}_r + \underline{u}_r \\
&= \underline{\eta}_r + \underline{u}_r
\end{aligned}$$

Thus, at each iteration of the algorithm, the updated estimates of $\underline{\beta}$ can be found from a weighted least squares fit on an iterative dependent variable, $\underline{z}$, with weight matrix, W. A similar results holds for the Newton-Raphson algorithm with observed rather than expected information. The

only difference in the two approaches is in the estimate of the asymptotic variance matrix of $\hat{\underline{\beta}}$, given by $(X^TWX)^{-1}$. In small samples there may be a considerable difference between the two, although, with the natural link function for a particular exponential family, the two methods give identical results.

Another important consequence of this formulation is the ready use of the observations as starting values. Since the algorithm only effectively employs $\underline{\eta}$ and $\underline{\mu}$, rather than the individual elements of $\underline{\beta}$, starting values for the linear predictors can be easily found by equating observed to expected values to give, $\underline{\eta}_0 = g(\underline{y})$.

2.3.2 Some Generalisations

The basic weighted least squares algorithm outlined in Section 2.3.1 has been extended by various authors to include models with error distributions not in the exponential family, non-linear parameterisations, and dependent observations.

A straightforward extension is to allow the inclusion of extra unknown parameters in the link function creating a *parametric* family of links, such as the Box-Cox link defined by Equation 2.3. The fitting of such models is discussed in Section 3.1.1.

Thompson and Baker, [71], introduced the idea of a composite link function, in which the mean of each observation may be related to more than one linear predictor. For the simplest case of a *linear* composite link, we may write $\underline{\mu} = C\underline{\gamma}$, where $g^{-1}(\underline{\gamma}) = \underline{\eta}$ and, in general, C is an $n \times m$ matrix. Thompson and Baker showed that the weighted least squares routine could be adapted by defining iterative explanatory variable CHX, dependent variable $CH\underline{z}$, where $H = [\partial\underline{\gamma}/\partial\underline{\eta}]$, and weight matrix $W = \text{diag}[V_i^{-1}]$. Applications of such models are discussed in Chapters 4 and 5.

Similar adaptations of the algorithm can be made for non-linear composite links of the form $\mu_i = c_i(\underline{\gamma})$, where the c_i, $i = 1, \ldots, n$ are non-linear vector functions. An obvious obstacle to the implementation of such models is the need to calculate the design matrix at each iteration. In the absence of general matix operations in some packages, in particular, GLIM, this has limited the general use of such models to situations in which the matrix C has a relatively simple form.

Several authors, for example Green, [33], Jorgenson, [40], and Stirling, [68], have considered quite general models with non-linear parameterisations and error distributions not of exponential family type, and shown that the weighted least squares algorithm can be used to fit such models. A general

formulation is to consider a log-likelihood, $l(\underline{\eta})$ in which the $n \times 1$ vector $\underline{\eta}$ is in turn a function of the p-vector of unknown regression parameters, i.e. $\underline{\eta} = \underline{\eta}(\underline{\beta})$. Using a similar notation as in Section 2.3, it is easily shown that the updated estimates of $\underline{\beta}$ can be found using

$$\begin{aligned} \underline{\hat{\beta}}_{r+1} &= (D_r^T W_r D_r)^{-1} D_r^T W_r (D_r \underline{\hat{\beta}}_r + \underline{u}_r) \\ &= (D_r^T W_r D_r)^{-1} D_r^T W_r \underline{z}_r \end{aligned} \tag{2.7}$$

where

$$D = \left[\frac{\partial \underline{\eta}}{\partial \underline{\beta}}\right]$$

$$W = E\left[\frac{\partial l}{\partial \underline{\eta}}\left(\frac{\partial l}{\partial \underline{\eta}}\right)^T\right] \quad \text{, and}$$

$$\underline{u} = \frac{\partial l}{\partial \underline{\eta}}$$

When the distribution of $\underline{y}$ is of exponential family form and the mean of y is related to the linear predictor ,$\underline{\eta}$, through a link function, g, these equations simplify to the weighted least squares algorithm described in section 2.3.1.

2.4 Inference for Generalised Linear Models

In this section we outline some basic distributional results for maximum likelihood estimates in general and for the estimates of β in a generalised

linear model in particular. The standard result for maximum likelihood estimates, and in particular generalised linear models, is that $\hat{\underline{\beta}}$ has a limiting Normal distribution with,

$$\begin{aligned} \mathrm{E}[\hat{\underline{\beta}}] &= \underline{\beta} \\ \mathrm{var}(\hat{\underline{\beta}}) &= (X^T W X)^{-1} \end{aligned}$$

Note that, for finite or small samples, $\hat{\underline{\beta}}$ is typically biased and its variance different from that given by equation 2.8. Shenton and Bowman, [66], discuss the small sample behaviour of maximum likelihood estimates.

The sampling distribution of $\hat{\underline{\beta}}$ is used primarily for testing hypotheses concerning individual parameters or linear combinations of parameters, for example

$$\mathcal{H}_0 \; : \; C\underline{\beta} = \underline{0}$$

where C is a $q \times m$ matrix of constants. A particular case is where C corresponds to $C = (0, \ldots, 1, \ldots, 0)$, so that we are testing,

$$\mathcal{H}_0 \; : \; \beta_j = 0$$

Using the results in equation 2.8 above, it is easy to show that the

limiting Normal distribution of $C\underline{\hat{\beta}}$ has moments,

$$\mathrm{E}[C\underline{\hat{\beta}}] = C\underline{\beta}$$

$$\mathrm{var}(C\underline{\hat{\beta}}) = C(X^TWX)^{-1}C^T$$

so that under the null hypothesis, $\mathcal{H}_0$

$$\underline{\hat{\beta}}^T C^T (C(X^TWX)^{-1}C^T)^{-1} C\underline{\hat{\beta}} \tag{2.8}$$

has a limiting χ^2_q distribution. The particular form of C for testing $\mathcal{H}_0 : \beta_j = 0$, leads to the usual test statistic

$$\frac{\hat{\beta}_j^2}{(X^TWX)^{-1}_{jj}} \tag{2.9}$$

which has a limiting χ^2_1 distribution.

In certain cases, notably when the link function is the canonical link, exact results are possible. In particular the use of such links leads to a set of sufficient statistics for the parameter vector $\underline{\beta}$. However, such link functions may not necessarily arise in practice or be applicable to a particular data set so that, in general, we will have to rely on asymptotic and approximate results.

The limiting distribution of $\underline{\hat{\beta}}$ will be a good approximation if the log-likelihood function is reasonably quadratic. Although, asymptotically, all likelihoods have this shape, for small samples this property may not hold

closely. This has implications in terms of the fitting algorithm, as well as the asymptotic distribution distribution of the parameter estimates, since the rate of convergence of the Netwton-Raphson algorithm depends on the quadratic nature of the log-likelihood.

The situation may be improved on both counts via transformations of the form $\gamma = \gamma(\beta)$. For example Anscombe, [4], considers the problem of finding a reparameterisation $\gamma = \gamma(\beta)$ which leads to an approximate Normal likelihood for $\hat{\gamma}$.

The likelihood ratio method of testing hypotheses and calculating confidence intervals is, in general, preferable to the approach based on the limiting distribution of $\underline{\hat{\beta}}$. One reason for this is the *invariance* property of the likelihood ratio in that inferences regarding $\underline{\beta}$ correspond directly to inferences regarding $\tau(\underline{\beta})$, where $\tau()$ is an arbitrary 1-1 transformation.

In terms of generalised linear models, the most frequently expressed form of the likelihood ratio statistic is the (scaled) deviance, defined by

$$\begin{aligned} D &= 2[l(\tilde{\underline{\mu}}; \underline{y}) - l(\hat{\underline{\mu}}; \underline{y})] \qquad (2.10) \\ &= 2[\sum_i l_i(\tilde{\mu}_i; y_i) - l_i(\hat{\mu}_i; y_i)] \end{aligned}$$

where $\tilde{\mu}$ denotes estimation in the extreme case by allowing a parameter for each observation, so that $\tilde{\mu}_i = y_i$. Thus the deviance measures the fit of the

current model, with corresponding parameter estimates $\hat{\mu}_i$, relative to the best possible model. In particular, we have $\hat{\mu}_i = g(\underline{x}_i^T \hat{\underline{\beta}})$. Standard results show that the limiting distribution of D is χ^2_{n-p}, where p is the dimension of $\underline{\beta}$

For some distributions, notably the Normal and Gamma, the deviance involves an unknown scale parameter, denoted by ϕ in Equation 2.1, which must be estimated before any tests can be carried out. The usual method is to estimate ϕ by the residual mean deviance from the maximal model, or the largest model under consideration, and scale the deviance by this value. This procedure typically leads to hypothesis tests based on the F-distribution rather than the χ^2.

The principal use of the deviance is in making model comparisons of the form

$$\mathcal{H}_0 \ : \underline{\beta}^T = [\beta_1, \ldots, \beta_q, 0, \ldots, 0]$$

against a more general hypothesis

$$\mathcal{H}_1 \ : \underline{\beta}^T = [\beta_1, \ldots, \beta_p], \quad \text{where } q < p < n.$$

The difference in deviances between the two models is

$$D = D_0 - D_1 \ = \ 2[l(\underline{\tilde{\mu}}; \underline{y}) - l(\underline{\tilde{\mu}}_0; \underline{y})] - 2[l(\underline{\tilde{\mu}}; \underline{y}) - l(\underline{\tilde{\mu}}_1; \underline{y})]$$

$$= 2[l(\underline{\hat{\mu}}_1; \underline{y}) - l(\underline{\hat{\mu}}_0; \underline{y})]$$

$$\sim \chi^2_{p-q}$$

which identically is the likelihood ratio test of the two hypotheses. Large values of D (i.e. greater than the upper α% point of the χ^2_{p-q} distribution) lend support to $\mathcal{H}_1$ on the basis that it provides a significantly better description of the data.

The usual application of this procedure is the sequential addition (or deletion) of regressor variables, usually one at a time, until the change in the deviance becomes non-significant. This procedure is obviously a generalisation of the subset selection procedures applicable in multiple linear regression and available as options in various statistical packages , although not in GLIM.

Recent work, Aitkin [1] and Whittaker [74], has focussed on the selection of factors in the analysis of contingency tables and, in particular, the representation of conditional independence models by graphical models.

2.5 Model Checking

An important aspect of data analysis is the process of verification and evaluation of the assumptions made when fitting the model. In most practical

situations, the process of model fitting is an iterative one involving model choice, model fitting and model verification. Only in very simple cases is it likely that one pass through these stages will be sufficient to highlight the important features of the data under study.

The most frequently used form of model checking involves the analysis of residuals. A residual may be define in a general form as some function on the product space of the observed and fitted values. Thus, for the ith observation, we define $r_i = r(y_i, \hat{\mu}_i), i = 1, \ldots, n$. The history of residual analysis in Normal theory linear models is long and widespread, and many of the concepts and techniques carry over in a natural way to the study of generalised linear models.

The most common form of residuals are generally referred to as standardised and studentised residual respectively, and may be defined as,

$$\text{standardised residual} = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(y_i)}}$$

$$\text{studentised residual} = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(y_i - \hat{\mu}_i)}}$$

In most cases, the variance expressions will be replaced by an estimate from the fitted model. The standardised residuals, for a given error distribution, correspond to the signed square root of the Pearson chi-squared statistic for the ith observation and are the residuals routinely output by

GLIM following a model fit. Gilchrist, [27], showed how studentised residuals could be calculated easily using the output from a model fit. A more general definition is that of the transformed residual defined by,

$$\text{transformed residual} \quad = \quad \frac{\tau(y_i) - \tau(\hat{\mu}_i)}{\text{var}(\tau(y_i))},$$

in which the transformation $\tau(.)$ is chosen in order to improve the Normality of the sampling distribution of the residuals, see for example Cox and Snell, [18]. Another generalised form of residual is the deviance residual defined by,

$$\text{deviance residual} \quad = \quad \text{sgn}(y_i - \hat{\mu}_i)\sqrt{D_i},$$

where

$$\text{sgn}(x) = \begin{cases} +1 & X > 0 \\ -1 & x < 0 \end{cases}$$

and D_i is the ith component of the deviance function defined in Equation 2.10. Pierce and Schafer, [55], consider the sampling distributions of deviance residuals in detail and conclude that, in general, they compare favourably with the best transformed residuals for specific models.

We now consider briefly the ways in which residual analysis may be utilised to highlight depeartures from the assumptions made when fitting a generalised linear model. The principal asumptions may be summarised as

i) Observations arise from some distribution $f(y, \theta)$.

ii) Observations are independent.

iii) One (or more observations) is an outlier.

Note that we have deliberately not included assumptions about the form of the link function $\eta = g(\mu)$ as this issue is considered in detail in Chapter 3.

To test these assumptions, residual plots may be utilised as follows,

i) Plots of the ordered residuals $r_{[i]}$ against the order statistics of some reference distribution, usually the Normal.

i) Plots of r_i against $\hat{\mu}_i$.

iii) Plots of r_i against i.

iv) Plots of partial residuals against $\underline{z}$, where $\underline{z}$ is a regresor variable for potential inclusion in the model.

In the first plot, the configuration expected is a straight line. In plots ii) and iii), any systematic tendencies or unusual points may be indicative of departures from the assumptions in the model. For example, residuals increasing in absolute value with the mean are suggestive of an incorrect mean/variance realationship in the fitted model. The final plot is less easy

to predict since its form may depend in many ways on the variables being considered and those already fitted in the model. It is probably most useful as a screening device when there are a large number of candidate regressor variables for inclusion in the model, see for example Landwehr *et al*, [42], for a discussion of these techniques in logistic regression.

Another important diagnostic procedure may be termed case deletion methods. By these we mean to quantify the effect of individual observations, or subsets of observations, on the estimated parameters in the model. Thus, in an obvious notation, we consider the effect of deleting the ith point in terms of the change from $\hat{\underline{\beta}}$ to $\hat{\underline{\beta}}_{(i)}$. Pregibon, [57], derived a one-step approximation to estimate the change in the values of the fitted parameters following the deletion of a point from a fit. This idea has also been discussed by Williams, [77]. In particular we can consider the *influence* curve for an observation, that is a plot of $\underline{\hat{\beta}(w)}_{(i)}$ against w, where $0 \leq w \leq 1$ is the weight given to the ith point in the fit and $\underline{\hat{\beta}(w)}_{(i)}$ is the vector of parameter estimates when the ith point is given weight w. Typically, the calculation of the influence function for all values of w is computationally expensive and it is more usual to consider the empirical influence function given by,

$$\Delta\underline{\hat{\beta}}_{(i)} = \underline{\hat{\beta}}(1)_{(i)} - \underline{\hat{\beta}}(0)_{(i)}$$

which is esentially a trade off between the influence function at $w = 1$ and the function at $w = 0$. As mentioned above, these techniques are considered in Chapter 3 in connection with the estimation of the form of the link function.

This concludes our overview of generalised linear models. The notation and methods introduced in this Chapter will be developed further in subsequent chapters as we consider the application of generalised linear modelling techniques to various non-standard analyses. In the next Chapter, we consider how generalised linear models may be extended in two important ways.

Chapter 3

EXTENSIONS TO THE BASIC MODEL

In this chapter we will discuss generalisations of the basic formulation of a generalised linear model discussed in Chapter 2.

The first generalisation is the introduction of extra parameters into the link function to create a "family" of parametric link functions. For a given data set, it is possible to estimate which link function represents the relationship between the mean and linear predictor. Secondly we will consider the determination of the appropriate error distribution for a given data set. In particular, this means the characterisation of a distribution by the mean–

variance relationship.

3.1 Parametric Link Functions

3.1.1 Formulation

Given a set of observations $y_1, \ldots, y_n$, having a distribution in the exponential family, we will assume that the relationship between the mean of y_i and the linear predictor, $\eta_i = \underline{x}_i^T \beta$, can be represented by,

$$\mu_i = h(\eta_i, \underline{\alpha}) \tag{3.1}$$

where, in general, $\underline{\alpha}$, is a set of unknown parameters not contained in the linear predictor. This is known as a *parametric link function.*

Pregibon [56], indicated how the weighted least squares algorithm might be modified to allow for the estimation of the extra parameters $\underline{\alpha}$ as well as the β_j's. Briefly, the technique consists of fitting an extra explanatory variable to the model for each extra parameter to be estimated. The extra variables can be calculated at each cycle , so that the $\underline{\beta}$ and $\underline{\alpha}$ parameters are estimated simultaneously , or, alternatively, after the β_j have been estimated for a fixed value of $\underline{\alpha}$. However, either of these techniques can fail for some problems; the first seems prone to failure when the link contains an unknown

exponent; the latter seems prone to failure when estimating links with an unknown asymptote.

An alternative approach is to consider the estimation of the two sets of parameters $\underline{\alpha}$ and $\underline{\beta}$ separately in a two stage process. This is the approach discussed in Scallan et al, [65], and only a brief outline of the technique is given here.

Suppose we denote the joint log-likelihood by $l(\underline{\beta}, \underline{\alpha})$, then the maximum likelihood estimate of $\underline{\beta}$ for a fixed value of $\underline{\alpha}$ is given by the solution to the equation

$$\frac{\partial l(\underline{\beta}, \underline{\alpha})}{\partial \underline{\beta}} = 0 \tag{3.2}$$

In many cases a fixed value of $\underline{\alpha}$ will represent a link function which can be easily specified in a computer package, for example using *GLIM*'s *OWN* facility. The estimates $\underline{\hat{\beta}}$ will, in general, be a function of $\underline{\alpha}$ and we will denote this relationship by $\underline{\hat{\beta}}(\alpha)$. We can then replace $\underline{\beta}$ in the likelihood by $\underline{\hat{\beta}}(\alpha)$ and, by regarding the likelihood as a function of $\underline{\alpha}$ alone, find the maximum likelihood estimate of $\underline{\alpha}$ as the solution to the equation,

$$\frac{\partial}{\partial \underline{\alpha}} l(\underline{\hat{\beta}}(\alpha), \underline{\alpha}) = 0 \tag{3.3}$$

Note that, in general, the solution of equation 3.3 involves the calculation of $\partial \underline{\hat{\beta}} / \partial \underline{\alpha}$. In [65] it was shown that the values of $\partial \underline{\hat{\beta}}(\underline{\alpha}) / \partial \underline{\alpha}$ could be found

using weighted least squares on a derived dependent variable. Given these values, Equation 3.3 may be solved using weighted least squares to give an updated estimate for $\underline{\alpha}$. This new value is then used to calculate updated values of $\hat{\underline{\beta}}(\alpha)$ and so on until both sets of parameters converge. This is a technique that was applied by Richards, [58], in the context of non-linear regression analysis; the common theme being that, for fixed values of $\underline{\alpha}$ in the applications described by Richards, the resulting model was one of straightforward multiple regression. Thus, the technique discussed in this section is the extension of the method to generalised linear models.

3.1.2 Applications

As well as proving useful in the estimation process, the quantities $\partial\hat{\underline{\beta}}(\underline{\alpha})/\partial\underline{\alpha}$ have other applications, in particular the calculation of the joint asymptotic covariance matrix of $\hat{\underline{\alpha}}$ and $\hat{\underline{\beta}}$. This matrix is shown in Equation 3.4

$$V = \hat{\phi}\begin{bmatrix} A^{-1} & A^{-1}C^T \\ CA^{-1} & B^{-1} + CA^{-1}C^T \end{bmatrix} \tag{3.4}$$

where A is the information matrix for the $\underline{\alpha}$ parameters, B is the information matrix for the $\underline{\beta}$ parameters given $\hat{\underline{\alpha}}$ and C is a matrix of the quantities $\frac{\partial\beta_i}{\partial\alpha_j}$.

Example 3.1 Carrot tops

Y	3.57	6.25	9.54	16.91	24.51	33.78	50.00	62.05	69.34	67.09	69.34
t	-2.15	-1.50	-0.85	-0.08	0.52	1.10	2.28	3.23	4.00	4.65	5.00

Table 3.1: Weight of Carrot tops

To note the effect of this adjustment on the covariance matrix of the $\underline{\beta}$ parameters, consider the data analysed in [65] and displayed in Table 3.1. This shows the weight of carrot tops y_k at time t_k, k=1,...,11, relative to some base time. A (log) logistic model for E[ln(Y_k)] of the form

$$\mu_k = \ln(\alpha) - \ln(1 + \exp\{-(\beta_0 + \beta_1 t_k)\}$$

fits the data well and gives parameter estimates $\hat{\alpha} = 73.26$, $\hat{\beta}_0 = -1.130$ and $\hat{\beta}_1 = 0.8542$. The estimated matrices , A and B, defined above are,

$$A^{-1} = [2280] \quad B^{-1} = \begin{bmatrix} 0.3484 & 0.1277 \\ 0.1277 & 0.1700 \end{bmatrix}$$

wth $\hat{\phi} = 0.000797$ and $C^T = [-0.0234, -0.00656]$ estimated using the technique described earlier. Thus, using Equation 3.4, the adjusted covariance matrix of $\hat{\underline{\beta}}$ is given by,

$$\hat{\phi} \begin{bmatrix} 1.601 & 0.4779 \\ 0.4779 & 0.2680 \end{bmatrix}$$

which can be confirmed directly as it is is straightforward to estimate all three parameters simultaneously for this model, see Scallan, [61] for details.

Several authors have proposed specific models for the analysis of binary response data. For example, Stukel, [70], considers a parametric logistic model of the form

$$p = \frac{\exp(h(\eta))}{1 + \exp(h(\eta))}$$

where, for $\eta \geq 0$,

$$h(\eta) = \begin{cases} \alpha_1^{-1}(e^{\alpha_1|\eta|} - 1) & \alpha_1 > 0 \\ \eta & \alpha_1 = 0 \\ -\alpha_1^{-1}\ln(1 - \alpha_1 \mid \eta \mid) & \alpha_1 < 0 \end{cases}$$

and, for $\eta \leq 0$,

$$h(\eta) = \begin{cases} \alpha_2^{-1}(e^{\alpha_2|\eta|} - 1) & \alpha_2 > 0 \\ \eta & \alpha_2 = 0 \\ \alpha_1^{-1}\ln(1 - \alpha_2 \mid \eta \mid) & \alpha_2 < 0 \end{cases}$$

The $h()$ functions govern the behaviour of the logistic function in either tail, depending on the value of (α_1, α_2). This is an example of a model in which, for fixed values of the link function parameters (α_1, α_2), we have a standard generalised linear model which may be fitted using an own model. In [70], it was suggested that the estimates of (α_1, α_2) could be found using a two-dimensional search method over possible parameter values. However, this is clearly a candidate model for the formulation given by Equation 3.1.

Example 3.2 Faults in fabric

To illustrate these ideas, consider the data in Table 3.2 which shows the number of faults in a piece of fabric of given length. A detailed analysis of this data was given by Hinde, [38], who showed that the data exhibits a significant degree of overdispersion relative to that which might be expected from fitting a Poisson regression model with ln(length) as explanatory variable. However, our interest centres on the determination of the appropriate form of link function for the data by introducing a parametric family of links.

The family of link functions we consider is given by Equation 3.5, i.e. the Box-Cox link.

$$\eta = \begin{cases} \frac{\mu^{\alpha}-1}{\alpha} & \alpha \neq 0 \\ \ln(\mu) & \alpha = 0 \end{cases} \tag{3.5}$$

The joint maximum likelihood estimates of α and $\underline{\beta}$ can be found easily using the two stage algorithm described above as $\hat{\alpha} = -1.337$ and $\hat{\underline{\beta}}^T = [0.2992, 0.06315]$ with deviance 60.798. The estimated variance covariance matrices for $\hat{\alpha}$ and $\hat{\underline{\beta}}$, and the vector $[\partial\hat{\underline{\beta}}/\partial\alpha]$ are as follows,

$$\text{var}(\hat{\alpha}) = [0.6012], \quad \text{var}(\hat{\underline{\beta}}) = \begin{bmatrix} 0.005834 & \\ -0.0008707 & 0.00013 \end{bmatrix} \quad \frac{\partial\hat{\underline{\beta}}^T}{\partial\alpha} = [-0.4759, 0.1423]$$

length of roll	faults	length of roll	faults
551	6	543	8
651	4	842	9
832	17	905	23
375	9	542	9
715	14	522	6
868	8	122	1
271	5	657	9
630	7	170	4
491	7	738	9
372	7	371	14
645	6	735	17
441	8	749	10
895	28	495	7
458	4	716	3
642	10	952	9
492	4	417	2

Table 3.2: Fabric Fault Data

Using the results derived above, the adjusted covariance matrix for $\hat{\underline{\beta}}$ is given by,

$$\text{var}(\hat{\underline{\beta}}) = \begin{bmatrix} 0.005834 & \\ -0.0008707 & 0.00013 \end{bmatrix} + \begin{bmatrix} 0.1362 & \\ -0.0407 & 0.0122 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1420 & \\ -0.0416 & 0.01233 \end{bmatrix}$$

Clearly, if α is treated as a parameter to be estimated, the variances of the estimates for $\hat{\underline{\beta}}$ are overwhelmed by the adjustment necessary because of the estimation of the link function. It is our experience that link function parameters, especially "power" parameters, are rarely estimated precisely. Moreover, as noted by several authors, the high correlations between the $\underline{\alpha}$ and $\underline{\beta}$ parameters may make the usual asymptotic covariance matrix somewhat unreliable. An alternative method of constructing confidence intervals for the link function parameters is via the use of profile likelihoods. Several examples of this technique are given in [65], and readers are referred to that paper for further details.

3.2 Generalising the Mean-Variance relationship

When fitting a generalised linear model it is assumed that error distribution of the observations is specified. In particular, as a consequence of results derived in section 2.1, this means making assumptions about the mean/variance relationship of the data. In keeping with the spirit of the techniques discussed in Section 3.1.1, it would clearly be desireable to formulate a model in which this relationship could be estimated or tested.

In order to do so, we can introduce a parametic form of the mean/variance relationship as follows. We assume that the probability density function of a random variable Y is of the form given in equation 2.1 and that the mean/variance relationship is given by,

$$\text{var}(Y) = \phi\mu^{\gamma} = \phi V_{\gamma}(\mu) \tag{3.6}$$

Using results from Section 2.1, if this relationship holds , the following differential equations must be satisfied.

$$\begin{aligned} \dot{b}(\theta) &= \mu \\ \ddot{b}(\theta) &= \mu^{\gamma} \end{aligned} \tag{3.7}$$

Gilchrist et al, [29], showed that the solution to these equations, apart from

arbitrary constants, is given by,

$$\theta = \begin{cases} \frac{\mu^{2-\gamma}}{1-\gamma} & \gamma \neq 1 \\ \ln(\mu) & \gamma = 1 \end{cases}$$

$$b(\theta) = \begin{cases} \exp(\theta) & \gamma = 1 \\ -\ln(-\theta) & \gamma = 2 \\ \frac{[(1-\gamma)\theta]^{\frac{2-\gamma}{1-\gamma}}}{2-\gamma} & \text{otherwise} \end{cases} \tag{3.8}$$

With the mean-variance relationship specified by equation 3.6, it is easy to show that the deviance function is given by

$$D(y;\mu) = \begin{cases} 2(y\ln(y/\mu) - (y-\mu)) & \gamma = 1, \\ 2(y/\mu - \ln(y/\mu) - 1) & \gamma = 2, \\ \frac{2(y^{2-\gamma} - (2-\gamma)y\mu^{(1-\gamma)} + (1-\gamma)\mu^{2-\gamma})}{(1-\gamma)(2-\gamma)} & \text{otherwise} \end{cases} \tag{3.9}$$

The deviance functions defined by Equation 3.9 may be used to test for the difference between competing models for a fixed value of γ since the maximum quasi-likelihood estimates have an asymptotic Normal distribution. However, attention is often focussed on the determination of an appropriate value of γ, and, hence, the mean/variance relationship, for a particular data set.We note in passing that the estimation of γ will often be dependent on the form of link function chosen and the regressor variables included in the model.

However, it is clearly not possible to use the deviance function given above in order to discriminate between models with different values of γ, since, each value of γ introduces a different scale.

A method for estimating the form of the mean/variance relationship was suggested by Nelder and Pregibon, [48]. They introduced an extended quasi-likelihood function, which, for a single observation has the form

$$Q_\gamma^+ = -\tfrac{1}{2}\ln(2\pi\phi V_\gamma(y)) - \tfrac{1}{2}D(y;\mu) \qquad (3.10)$$

where, in particular, in particular, $D(y;\mu)$ is given by Equation 3.9. It can be shown that, for certain values of γ, this function is very similar (exact in the Normal case) to appropriate members of an exponential family with the same mean/variance relationship. Nelder and Pregibon give several examples of using this extended likelihood function to estimate the value of γ for several data sets and report generally favourable results. The asymptotic properties of extended quasi-likelihood estimates are considered in relation to the method of pseudo-likelihood by Davidian and Carroll, [20].

3.3 Diagnostic Procedures

The role of diagnostic checks in assessing the fit of a model has already been discussed with particular reference to residual analysis. In this section

we discuss the application of a certain form of diagnostic procedure with reference to some of the model generalisations discussed in this chapter. The procedure we will utilise primarily is that of case-deletion, that is, assessing the effects individual points on the parameter estimates in the fitted model. The procedure is easily demonstrated for the classical linear regression model where it is well known that,

$$\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)} = (X^T X)^{-1} \underline{x}_i^T \frac{r_i}{1 - h_{ii}}. \tag{3.11}$$

Here $\hat{\underline{\beta}}_{(i)}$ means the vector of parameter estimates with the ith data point deleted, h_{ii} is the ith diagonal element of the hat matrix, $H = X(X^T X)^{-1} X^T$ and $\underline{r} = (\underline{y} - X\hat{\underline{\beta}}$. Thus the quantity $\Delta\hat{\underline{\beta}} = \hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)}$, gives an indication of the effect of the ith data point on the estimated model parameters. Plots of $\Delta\hat{\underline{\beta}}_i / \text{s.e.}\hat{\beta}_i$ against the index number of the observations are useful in detecting influential points.

The extension of this procedure to generalised linear models is straightforward since we know that at the maximum likelihood estimate,

$$\hat{\underline{\beta}} = (X^T W X)^{-1} X^T W \underline{z}$$

where $W = \text{diag}[\delta_i^2 V_i]$ and $\underline{z} = \underline{\eta} + (\underline{y} - \underline{\mu}) \otimes \underline{\ell}$. Rewriting this equation as,

$$\hat{\underline{\beta}} = ((W^{1/2}X)^T (W^{1/2}X))^{-1} (W^{1/2}X)^T W^{1/2} \underline{z},$$

it is easy to show that a one-step approximation to $\Delta\underline{\hat{\beta}}$ is given by,

$$\Delta\underline{\hat{\beta}} = (X^T W X)^{-1} \underline{x}_i^T \frac{r_i}{\delta_i V_i^{1/2}(1 - h_{ii})} \qquad (3.12)$$

where $r_i = (y_i - \mu_i)/V_i^{1/2}$, h_{ii} is the ith diagonal element of the corresponding hat matrix for a generalised linear model and V is the variance function of Y. This procedure seems to have first been used by Pregibon, [57], in the analysis of logistic regression models and the application to generalised linear models discussed by Williams, [77]. In the following sections we show how this procedure may be applied to some of the extended models discussed in this Chapter.

Wiliams also discusses diagnostic procedures in terms of the changes of the deviance values following the deletion of an observation from a fit. In particular, by considering a Taylor series expansion of the deviance, $D = \sum d_i^2$, it is easy to show that,

1. The decrease in $\sum_{j \neq i} d_j^2$ is approximately $\phi h_i (1 - h_i)^{-1} r_i^2$.
2. The increase in d_i^2 is approximately $\phi h_i (2 - h_i)(1 - h_i)^{-2} r_i^2$.
3. The increase in $D = \sum_j d_j^2$ is approximately $\phi h_i (1 - h_i)^{-2} r_i^2$.

As these results rely on the approximation to changes in the parameter estimates, they may also be applied to the more general models discussed

below . These results are reported more fully elsewhere.

We also note that these techniques may be applied to situations in which more than one observation is excluded from the model at a time. In particular, we may consider the estimation of the quantity $\Delta\hat{\underline{\beta}}_L$, where , in an obvious notation, the subscript L refers to the change in parameter estimates when observations $y_{i_1}, \ldots, y_{i_l}$ are excluded from the model. Following Pregibon, [57], this quantity may be approximated by

$$\Delta\hat{\underline{\beta}}_L = (X^T W X)^{-1} X_L^T W_L^{-1/2} (I - H_L)^{-1} \underline{r}_L \tag{3.13}$$

As noted in [57], all the quantities required are available after a model fit except $(1 - H_L)^{-1}$. A partial solution to this problem is offered by Scallan, [64], who shows how the elements of H_L, and hence $(I - H_L)$, may be calculated following a model fit by an extra iteration involving auxiliary variables. For moderately sized values of l, the inversion of $(I - H_L)$ should not be too difficult even for a package such as GLIM which does not explicitly support matrix operations.

3.3.1 Composite Link Functions

The principle of a composite link function was introduced in Section 2.2. We can define one extended form of composite link function known as a

bilinear function, although, in keeping with the terminology of this chapter, we prefer the term parametric composite link function, as follows,

$$\begin{aligned} \underline{\mu} = \underline{\eta}_o &= C\gamma \\ &= \textstyle\sum_{j=1}^{r} \alpha_j C_j \gamma \end{aligned} \tag{3.14}$$

where, $\gamma = h(\underline{\eta}_\beta)$, $\underline{\eta}_\beta = X\underline{\beta}$ and the C_j, $j = 1, \ldots, r$, are known $n \times m$ matrices. The $r \times 1$ and $p \times 1$ parameter vectors $\underline{\alpha}$ and $\underline{\beta}$ may be estimated using iteratively re-weighted least squares with design matrix X^* and dependent variable z^* given by,

$$\begin{aligned} X^* &= [C_1\gamma, \ldots, C_r\gamma : CHX] \\ z^* &= \eta_o + CH\eta_\beta + (\underline{y} - \underline{\mu}) \end{aligned}$$

where $H = \text{diag}[\partial\gamma/\partial\eta]$ is an $m \times m$ matrix, and with weights $W = \text{diag}[1/V_i]$. Now, since at the mle we have $(\hat{\underline{\alpha}}, \hat{\underline{\beta}}) = (X^{*T}X)^{-1}X^{*T}Z^*$, we can use the general result given in Equation 3.12 to find approximations for the changes in the value of $(\hat{\underline{\alpha}}, \hat{\underline{\beta}})$ when the ith case is deleted. Thus we have,

$$\Delta(\hat{\underline{\alpha}}, \hat{\underline{\beta}})_i = (X^{*T}X^*)^{-1}x_i^* \frac{r_i}{(1 - h_{ii})V_i^{1/2}} \tag{3.15}$$

where h_{ii} is the ith diagonal element of the hat matrix and $r_i = (y_i - \mu_i)/V_i^{1/2}$. As noted by Gilchrist, [28], the diagonal elements of the hat matrix for a composite link model may be found easily following a fit.

As a simple example consider the data shown in Table 3.3, which shows the number of coal miners, R out of N, diagnosed as suffering from a wheeze at different ages. Ekholm and Palgrem, [22], analysed this data by considering the possibility of misdiagnosis. Burns, [8], showed how this model could be fitted in GLIM using a parametric link formulation. Briefly, we assume the probability of being diagnosed as having a wheeze, p_k is given by,

$$p_k = (1 - \epsilon_1)\gamma_k + \epsilon_0(1 - \gamma_k),$$

where, $\epsilon_0 = P(\text{diagnosed wheeze} \mid \text{healthy})$, $\epsilon_1 = P(\text{diagnosed healthy} \mid \text{wheeze})$ and $\gamma_k = P(\textit{true}\ \text{wheeze}) = \exp(\eta_k)/(1 + \exp(\eta_k))$. In the notation above $(\epsilon_0, \epsilon_1) \equiv (\alpha_1, \alpha_2)$.

The maximum likelihood estimates of the parameters in the model are shown in Table 3.4. As noted in [8], the estimate for ϵ_0 has a negative value and large standard error, which suggests more appropriate model would be one in which $\epsilon_0 = 0$. However, this example is used purely to illustrate the technique and we will not be concerned the suitability of the fitted model.

The standardised changes in the parameter estimates for this data set are shown in Table 3.5. For each parameter, the first column shows the exact change and the second the one step approximation. As can be seen from Table 3.5, there is good agreement between the exact and approximate

Age	R	N
22.5	104	1952
27.5	128	1791
32.5	231	2113
37.5	378	2783
42.5	442	2274
47.5	593	2393
52.5	649	2090
57.5	631	1750
62.5	504	1136

Table 3.3: Number of miners diagnosed *wheezy*

changes in most cases. These quantities are illustrated graphically in Figures 3.1 to 3.4. In general, the approximate changes are underestimates of the actual changes, although the relative magnitude of the changes is reasonably similar in most cases. The only observation which appears to have a significant effect on the estimates when deleted is observation number 9, in which deletion causes a large shift in the values of all the parameter estimates. To an extent this may be expected since a change in the link function parameters may well alter the whole nature of the form of the link

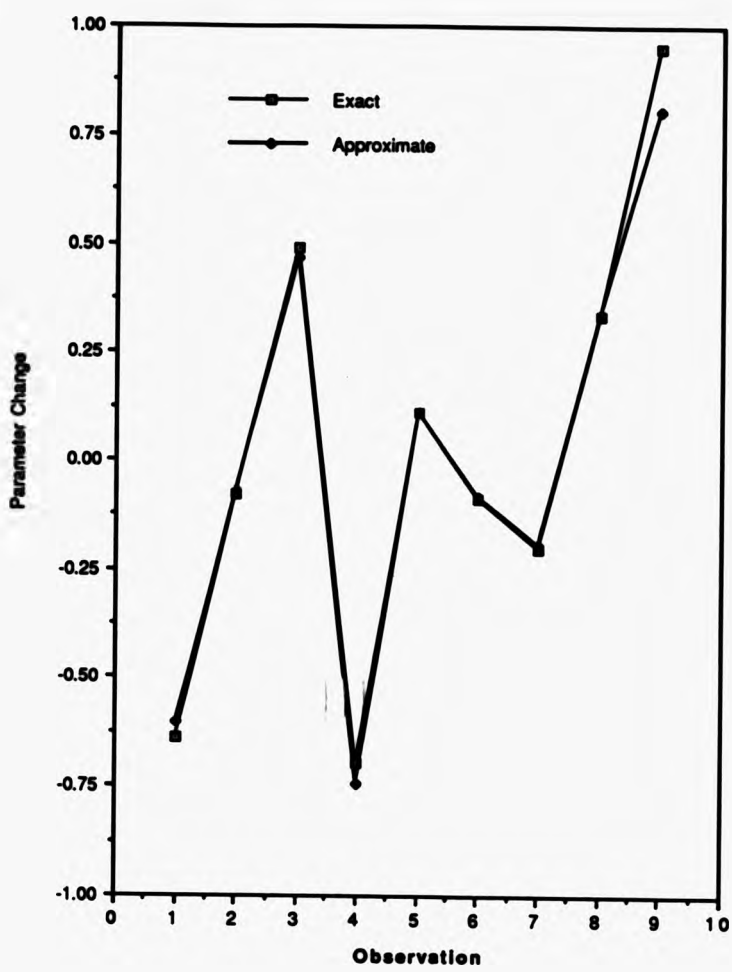

Figure 3.1: Standardised Parameter changes - β_1

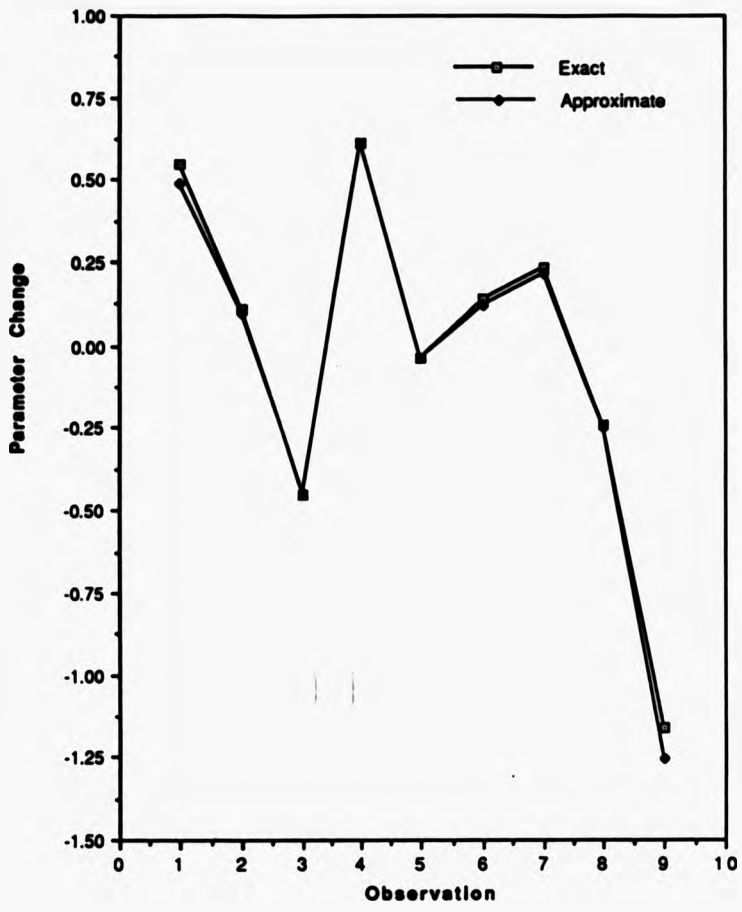

Figure 3.2: Standardised Parameter changes - β_2

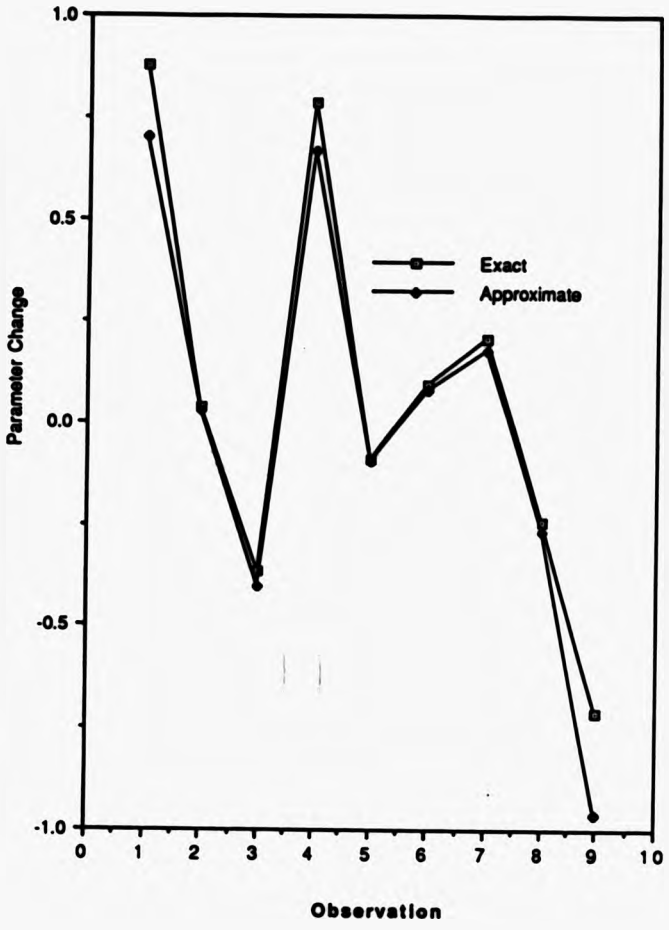

Figure 3.3: Standardised Parameter changes - ϵ_0

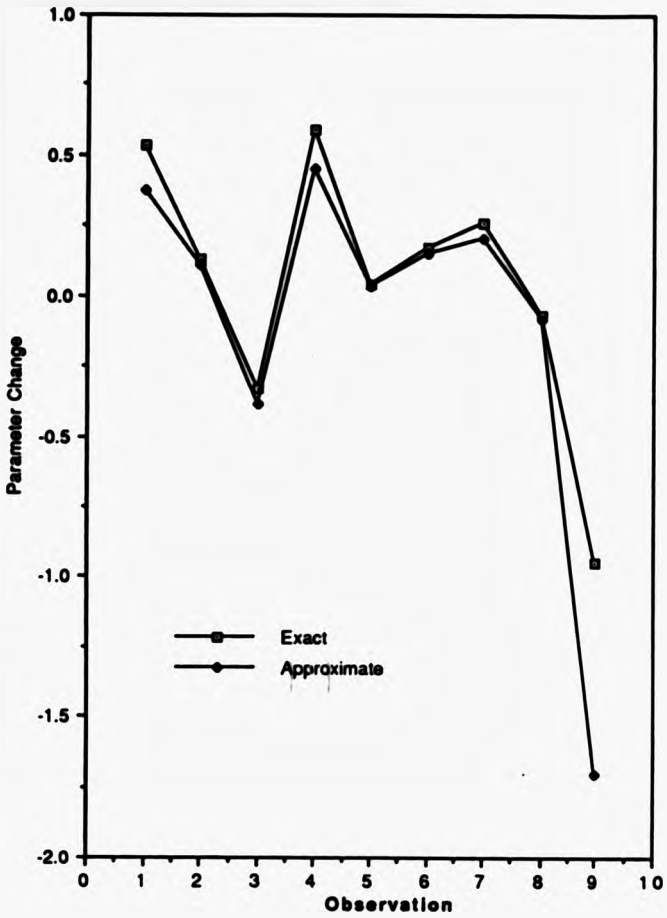

Figure 3.4: Standardised Parameter changes - ϵ_t

Parameter	estimate	s.e.
β_1	-4.090	0.6132
β_2	0.07324	0.01932
ϵ_0	-0.004567	0.02942
ϵ_1	0.2940	0.2022

Table 3.4: Parameter estimates for miners data

function and, consequently, the values of the regression parameters, $\underline{\beta}$.

3.3.2 Generalised IRLS models

In principle there is no reason why the techniques illustrated in this section cannot be applied to the general IRLS formulation described in Equation 2.7. To illustrate the idea we consider a special case of this formulation in which the model still contains a linear predictor, $\underline{\eta}$, although observations need not necessarily be a member of an exponential family. Thus, we assume that we have independent observations, y_i with log likelihoods,

$$\log p(y_i \mid \eta_i) = l(y_i, \eta_i) \quad \text{where} \quad \underline{\eta} = X\underline{\beta}$$

Stirling, [68], showed that such models could be fitted with iterative weight matrix $W = \text{diag}[w_i]$ and explanatory variable $\underline{z}$ given by,

$$w_i = -l''(y_i, \eta_i)$$

Case	Parameter							
	β_1		β_2		ϵ_0		ϵ_1	
1	-0.638	-0.605	0.551	0.492	0.878	0.701	0.536	0.372
2	-0.080	-0.073	0.111	0.099	0.039	0.027	0.125	0.108
3	0.489	0.466	-0.454	-0.445	-0.363	-0.403	-0.333	-0.385
4	-0.698	-0.743	0.613	0.614	0.786	0.669	0.591	0.448
5	0.109	0.112	-0.034	-0.036	-0.091	-0.097	0.037	0.033
6	-0.088	-0.082	0.137	0.123	0.093	0.081	0.172	0.146
7	-0.202	-0.194	0.235	0.215	0.208	0.180	0.257	0.206
8	0.334	0.340	-0.241	-0.249	-0.244	-0.265	-0.074	-0.086
9	0.951	0.806	-1.163	-1.254	-0.713	-0.961	-0.952	-1.704

Table 3.5: Comparison of exact and approximate changes in parameter estimates for single case deletion in miners' data

$$z_i = \eta_i - \frac{l'(y_i, \eta_i)}{l''(y_i, \eta_i)}$$

where differentiation is with respect to η. Note that, in some cases, the observed information matrix, W, may be replaced by its expectation.

We note, in particular, that much attention has been focussed on case deletion measures of influence for Cox's proportional hazards model, see for example Storer and Crowley, [69], and Pettitt and Bin Daud, [54].

As an example consider observations y_i which follow an exponential distribution with mean $\lambda_i^{-1} = \exp^{-\eta_i}$ and may be right censored. Thus, the exact failure time y_i is known exactly only if $y_i < t_i$, otherwise it is recorded as censored with $y_i \geq t_i$. Stirling showed that this model results in IRLS formulae

$$w_i = \begin{cases} y_i e^{\eta_i} & y_i < t_i \\ t_i e^{\eta_i} & y_i \geq t_i \end{cases}$$

and

$$z_i = \begin{cases} \eta_i + (y_i e^{\eta_i})^{-1} - 1 & y_i < t_i \\ \eta_i - 1 & y_i \geq t_i \end{cases}$$

which corrects a small mistake in the original paper.

We have applied these formulae to the data analysed by Gehan, [26], displayed in Table 3.6. The data consists of 42 observations of remission

Control	1(2)	2(2)	3	4(2)	5(2)	8(4)
	11(2)	12(2)	15	17	22	23
Treatment	6(3)	6^c	7	9	10	10^c
	11^c	13	16	17^c	19^c	20^c
	22	23	25^c	$32^c(2)$	34^c	35^c

Table 3.6: Gehan data

Parameter	Estimate	s.e.
β_1	-2.159	0.2182
β_2	-1.527	0.3984

Table 3.7: Parameter estimates for Gehan data

times in weeks of leukaemia patients. A randomized treatment group was treated with 6-mercaptopurine, the other group was a control. Note that figures in brackets refer to multiplicities of observations and ^{c} denotes a censored observation. The parameter estimates for this model are displayed in Table 3.7. Again, using the result that, at the maximum likelihood estimate, we have

$$\hat{\underline{\beta}} = (X^T W X)^{-1} X^T W \underline{z}$$

we can use the techniques of this Section to find the one-step approximations to the changes in $\hat{\underline{\beta}}$ on the deletion of each observation. These changes

are summarised in Table 3.8. It is interesting to note from Table 3.8 the effect on the parameter estimates of a censored observation compared to an uncensored observation. In most cases, the sign of the change is reversed between censored and uncensored observations. However, it is clear that, for observations which have a substantial effect on the parameter estimates, the censoring is irrelevant in terms of the magnitude of the change.

This model can also be fitted by using an auxiliary Poisson model, a technique that has been used to fit various survival distribution models using GLIM (see Section 5.3.4 for a discussion of a technique for fitting the logistic distribution to censored data). Since these models also rely on a weighted least squares approach, similar diagnostic techniques may be employed. We have compared the auxiliary Poisson model with the direct approach for the Gehan data set and found the results to be similar. The difference is probably explained by the use of the expected information in the auxiliary Poisson model. We are currenlty investigating the use of these techniques in other applications.

3.3.3 Tests on Link Functions

Another important topic discussed in this Chapter has been the idea of a parametric link function. It is clearly desireable to assess whether one or

	Group						
	Control				Treatment		
Observation	β_1		β_2		Observation	β_2	
1(2)	0.1983	0.2027	-0.1086	-0.1110	6(3)	0.2533	0.2666
2(2)	0.1729	0.1763	-0.0947	-0.0965	6[c]	-0.0423	-0.0472
3	0.1474	0.1498	-0.0807	-0.0821	7	0.2462	0.2587
4(2)	0.1217	0.1234	-0.0667	-0.0676	9[c]	-0.0637	-0.0708
5(2)	0.0959	0.0969	-0.0525	-0.0531	10	0.2247	0.2351
8(4)	0.0176	0.0176	-0.0096	-0.0097	10[c]	-0.0709	-0.0787
11(2)	-0.0621	-0.0617	0.0340	0.0338	11[c]	-0.0781	-0.0865
12(2)	-0.0890	-0.0881	0.0487	0.0483	13	0.2031	0.2115
15	-0.1706	-0.1674	0.0934	0.0917	16	0.1812	0.1879
17	-0.2258	-0.2203	0.1237	0.1207	17[c]	-0.1218	-0.1337
22	-0.3668	-0.3325	0.2009	0.1931	19[c]	-0.1365	-0.1494
23	-0.3955	-0.3789	0.2166	0.2076	20[c]	-0.1439	-0.1573
					22	0.1369	0.1407
					23	0.1294	0.1328
					25[c]	-0.1812	-0.1966
					32[c](2)	-0.2343	-0.2517
					34[c]	-0.2497	-0.2674
					35[c]	-0.2575	-0.2753

Table 3.8: Case deletion diagnostics for Gehan data

more observations may have high or undue influence in the estimation of the parameters in the link function.

The two-stage estimation technique introduced in Section 3.1.1 involves the iterative solution of a single weighted least squares equation to find updated estimates of the link function parameters, $\underline{\alpha}$. Again, drawing on the analogy with the general result of the effect of single case deletions on parameter estimates, Equation 3.12, it seems reasonable to apply this technique to the estimation of the link function parameters.

As an example, consider again the data on the number of faults in fabric displayed in table 3.2. For the Box-Cox parametric link function, the estimate of the link function parameter for the full data set is $\alpha = -1.337$. Table 3.9 shows the fully iterated an one-step approximations to the changes in the values of the link function parameter following the deletion of each observation. For each observation, the first entry shows the exact change in the parameter estimate and the second, the one-step approximation.

Although the agreement between the exact and approximate values is not as close as those observed in the other applications considered in this Section, those observations with a large effect do seem to be highlighted by the one-step approximations. For example, observations 19, 23 and 31 produce the largest changes in the value of the link function parameter, and

this is in agreement with the approximate changes given by the one-step approximations.

For comparison, we have also calculated these quantities using the auxiliary variable technique described by Pregibon. The results of the two methods are displayed graphically in Figures 3.5 and 3.6. As can be seen, the one-step approximations using Pregibon's auxiliary variable technique are not as accurate as those obtained from the two-stage algorithm. However, both techniques seem to pick out reasonably well the influential observations.

This data was analysed in [38] as an example of an overdispersed Poisson model. The deviance he obtained for a Normal/Poisson compound fitted to the data was 50.98 on 30 df. The deviances for various models deleting the most influential observations in terms of the link function parameters are shown in Table 3.10. It is interesting to note that the improvement in the deviance achieved by fitting the compound model is matched, and in some cases exceeded, by fitting a more general link function and deleting influential observations.

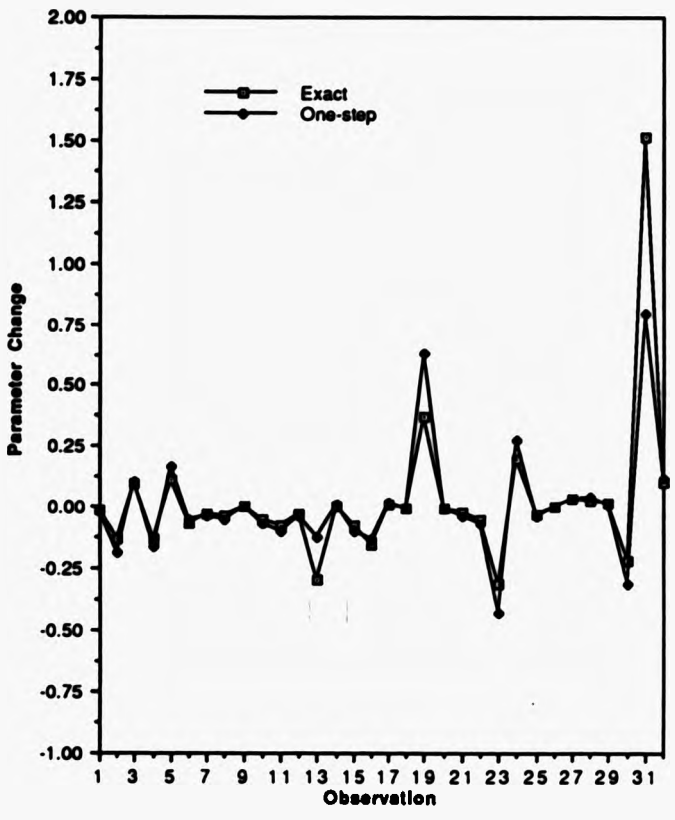

Figure 3.5: Link function parameter changes , two-stage method

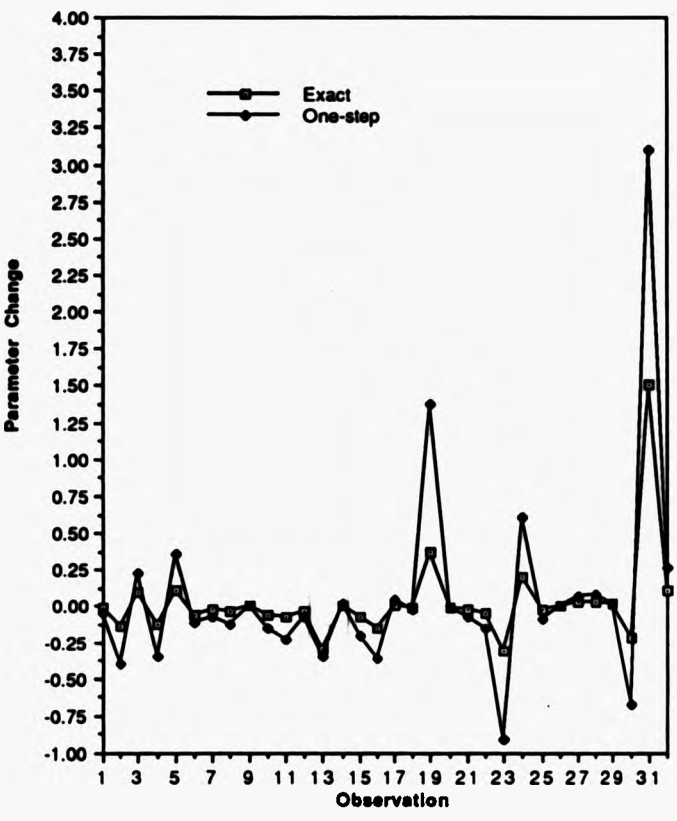

Figure 3.6: Link function parameter changes, Pregibon's method

Observation	1	2	3	4	5	6	7	8
Exact	-0.012	-0.133	0.093	-0.127	0.112	-0.066	-0.025	-0.037
One-step	-0.022	-0.182	0.104	-0.165	0.165	-0.056	-0.034	-0.054
Observation	9	10	11	12	13	14	15	16
Exact	-0.001	-0.055	-0.072	-0.032	-0.294	0.003	-0.077	-0.151
One-step	-0.001	-0.071	-0.103	-0.037	-0.122	0.007	-0.096	-0.133
Observation	17	18	19	20	21	22	23	24
Exact	0.008	-0.002	0.368	-0.002	-0.024	-0.049	-0.314	0.195
One-step	0.019	-0.007	0.627	-0.004	-0.034	-0.070	-0.425	0.279
Observation	25	26	27	28	29	30	31	32
Exact	-0.026	0.000	0.035	0.026	0.015	-0.216	1.513	0.107
One-step	-0.037	-0.000	0.037	0.040	0.010	-0.309	0.796	0.124

Table 3.9: Exact and approximate changes in link function parameters for Fabric data

	Observation deleted			
Link function	None	19	23	31
Log	64.6	64.0	54.8	61.9
Box-Cox	60.8	58.1	51.0	47.6

Table 3.10: Deviance values for fabric data

Chapter 4

Analysis of Grouped Data

4.1 Introduction

The formulation of composite link functions has already been described briefly in Section 2.3.2. In this chapter we illustrate the application of composite link functions to the analysis of grouped data in which the the data may arise from an underlying truncated or mixture distribution. Moreover, we show how the model fitting facilities of a package such as *GLIM* enable potentially complex models to be fitted to such data.

In their original article on composite link functions, Thompson and Baker [71] illustrated the application of such models to the analysis of grouped Normal data. The formulation they used was as follows.

Let Y_i denote the frequency of the ith class, $i = 1, \ldots, n$. We assume the expected frequency is given by,

$$E[Y_i] = \mu_i = \begin{cases} N\Phi(\eta_1) & i = 1 \\ N(\Phi(\eta_i) - \Phi(\eta_{i-1})) & i = 2, \ldots, n \end{cases}$$

where $\eta_i = (u_i - \mu)/\sigma)$ and $N = \sum_i Y_i$. Here u_i represents the upper bound of the ith class and μ and σ are the mean and standard deviation, respectively, of the underlying distribution. By letting $\gamma_i = \Phi(\eta_i)$, it is clear that we can write, $\underline{\mu} = NC\underline{\gamma}$, where

$$C = \begin{bmatrix} 1 & 0 & \ldots & \ldots & 0 \\ -1 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & -1 & 1 \end{bmatrix}$$

Which is exactly the form of a linear composite link function. The parameter vector estimated by this formulation is $\beta^T = [-\mu/\sigma, 1/\sigma]$ although estimates of μ and σ can obviously be obtained via a simple transformation. Vanderhoeft, [72], describes a technique for obtaining approximate confidence intervals for such parameter transformations. In the next Section we show how this technique can be adapted to fit models to grouped data which arise from an underlying truncated distribution.

The parameterisation of such models for grouped data has been discussed by Burridge, [9] and [10]. Burridge showed that, if the density function of the observations was log-concave, then so was the log-likelihood function of the parameters in a regression-type model for both grouped and un-grouped data.

The extension of this result to data which arises from an underlying truncated or mixture distribution is open to doubt. It is well known, for example, that Normal mixture distributions admit singularities in the likelihood function at each observation, so that it is unlikely conditions will be found that result in a well-behaved likelihood function, even for grouped data. In the case of truncated distributions, it is easy to find examples in which the likelihood function is not concave even for a simple scale/location model.

4.2 Truncated Distributions

In a recent article, McClaren, Brittenham and Hasselbad [46], gave an example of using the EM algorithm to fit a truncated log-Normal distribution to grouped data. In this section we illustrate the formulation of such models in terms of composite links and illustrate the technique using an example

given in [46].

4.2.1 Formulation

Assuming that the sum of the frequencies is N, we can derive the model as a composite link function by assuming that the frequency in each cell, $y_1, \ldots, y_n$, has mean , μ_i, given by

$$\mu_i = N\frac{(\gamma_i - \gamma_{i-1})}{(\gamma_n - \gamma_0)} \quad , i=1,\ldots,n, \tag{4.1}$$

where $\gamma_i = F(\eta_i)$, $\eta_i = (u_i - \mu)/\sigma$, and $F(.)$ is some distribution function. In particular, this means that we assume that observations outside range (u_0, u_n) are unrecorded or lost. The parameter vector of interest is $\beta^T = [\mu, \sigma]$. In the implementation of the algorithm we have chosen to estimate μ and σ directly rather than use the parameterisation discussed in Section 4.1, namely $-\mu/\sigma$ and $1/\sigma$. In our experience, there is little benefit to be gained from the latter and the former gives asymptotic standard errors directly, rather than relying on parameter transformations. The u_i represent the upper bounds of each category and u_n and u_0 represent the upper and lower truncation points respectively. In the terminology of [71], this is a non-linear composite link function, and can be fitted with iterative explanatory

variable,

$$
\begin{aligned}
X^{*} &= CHX \\
X^{T} &= \begin{bmatrix} -1/\sigma & \cdots & -1/\sigma \\ (\mu - u_0)/\sigma^2 & \cdots & (\mu - u_n)/\sigma^2 \end{bmatrix}
\end{aligned} \tag{4.2}
$$

and iterative dependent variable,

$$
\underline{Z}^{*} = CH\underline{\eta}^{*} + (\underline{y} - \underline{\mu}) \tag{4.3}
$$

where $C = [\partial \underline{\mu}/\partial \underline{\gamma}]$, $H = \text{diag}[\partial \underline{\gamma}/\partial \underline{\eta}]$ and $\underline{\eta}^{*} = X\beta$. Note that both C and H must be updated at each iteration unlike a linear composite link where C is a constant matrix.

However, the $n \times (n+1)$ matrix C has a relatively simple form as shown in Equation 4.4.

$$
\frac{\partial \underline{\mu}}{\partial \underline{\gamma}} = \frac{N}{\alpha^2}
\begin{bmatrix}
\delta_1 - \alpha & \alpha & 0 & \cdots & 0 & -\delta_1 \\
\delta_2 & -\alpha & \alpha & \ddots & \vdots & \vdots \\
\vdots & 0 & \ddots & \ddots & 0 & \vdots \\
\vdots & \vdots & \ddots & -\alpha & \alpha & -\delta_{n-1} \\
\delta_n & 0 & \cdots & 0 & -\alpha & \alpha - \delta_n
\end{bmatrix} \tag{4.4}
$$

where $\alpha = \gamma_n - \gamma_0$ and $\delta_k = \gamma_k - \gamma_{k-1}$. Thus, for any arbitrary vector, $\underline{w}^{T} = [w_0, w_1, \ldots, w_n]$, then the kth element of $C\underline{w}$, $k = 1, \ldots, n$, is given by,

$$
\delta_k(w_0 - w_n)/\alpha^2 + (w_k - w_{k-1})/\alpha\,, \tag{4.5}
$$

Thus the algorithm is relatively straightforward and efficient to implement because it essentially only involves differencing vectors rather than matrix multiplication. Macros for fitting this model are listed in Appendix A.1

In [71], it was commented that N should be regarded as an extra unknown parameter to be estimated. Furthermore, in the example considered in that article, it was found that the maximum likelihood estimate of N was given by the observed sample total. A more formal justification of this procedure is as follows.

In the notation introduced earlier, if $\gamma_n \neq 1$ or $\gamma_0 \neq 0$, then the underlying distribution may be regarded as truncated and we can model the expected number of observations in the ith interval as in Equation 4.1. Thompson and Bakers approach is to assume that the expected frequency in the ith interval is given by,

$$\mu_i = \hat{N}(\gamma_i - \gamma_{i-1}), \qquad i=1,\ldots,n \tag{4.6}$$

It is easy to show that the likelihood equation for N leads to the estimator, $\hat{N} = N/\alpha$, in particular, if $\alpha = \gamma_n - \gamma_0 = 1$, the estimate of N is given by the observed sample total. This was the case in both examples considered by Thompson and Baker. The likelihood equations for the other parameters

in the model, i.e. (μ, σ), may be written as,

$$\text{Truncated Model} \quad : \quad (CHX)^T \underline{s} = 0$$

$$\text{T and B Model} \quad : \quad (C_T HX)^T \underline{s} = 0$$

where $s_i = (y_i - \mu_i)/\mu_i$, $i = 1, \ldots, n$ and C_T is the $n \times (n+1)$ matrix

$$C_T = \hat{N} \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ldots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & -1 & 1 \end{bmatrix}$$

Given the estimator $\hat{N} = N/\alpha$, it is straightforward to verify that any estimate which satisfies the likelihood equations for the Thompson and Baker model also satisfies the equations for the truncated model. This shows that, treating N as an extra unknown parameter, is implicitly fitting an underlying truncated distribution to the data.

Although the link derivative matrix for the truncated model is slightly more complicated than that for the Thompson and Baker approach, the latter involves the fitting of an extra variable and, for this reason, we prefer the former method in this case. However, we have used this method of estimating the overall sample size to fit models to data involving truncated mixture distributions. In this case, the approach treating N as an extra

parameter is computationally easier because of the more complicated form of the link derivative matrix. Details of the method are given in Section 4.3.1.

Applications

In this section we illustrate the fitting of truncated distributions to an example data set and discuss, briefly, some other applications.

Example 4.1 Blood Cell Analysis

In [46] a doubly truncated log-Normal distribution was fitted to grouped data derived from an analysis of red blood cell size. Sample number 28, in that paper, was grouped into 34 cells of width 1.3125 fl, and is shown in Table 4.1. The ease with which such models can be fitted using the composite link approach are illustrated in the following example output from a *GLIM* session.

```
!Read in cell frequencies
$units 34$data freq$read
 32  45  44  72  97 103 136 132 142 162 167 188
215 194 204 195 199 204 187 150 169 160 150 131
103  91  96  79  78  57  45  41  35  37
```

Cell Volume	Frequency	Cell volume	Frequency
67.3125–68.6250	32	89.6250–90.9375	204
68.6250–69.9375	45	90.9375–92.2500	187
69.9375–71.2500	44	92.2500–93.5625	150
71.2500–72.5625	72	93.5625–94.8750	169
72.5625–73.8750	97	94.8750–96.1875	160
73.8750–75.1875	103	96.1875–97.5000	150
75.1875–76.5000	136	97.5000–98.8125	131
76.5000–77.8125	132	98.8125–100.1250	103
77.8125–79.1250	142	100.1250–101.4375	91
79.1250–80.4375	162	101.4375–102.7500	96
80.4375–81.7500	167	102.7500–104.0625	79
81.7500–83.0625	188	104.0625–105.3750	78
83.0625–84.3750	215	105.3750–106.6875	57
84.3750–85.6875	194	106.6875–108.0000	45
85.6875–87.0000	204	108.0000–109.3125	41
87.0000–88.3125	195	109.3125–110.6250	35
88.3125–89.6250	199	110.6250–111.9375	37

Table 4.1: Red Blood Cell Volumes

```
!Calculate upper bounds of cells
$ca ub=67.3125+%gl(34,1)*1.3125
!Set up fit with log-Normal errors
$calc ub=%log(ub) : %o=1$
!Read in starting values (calculated from sample moments)
$data 2 sval$read 4.475 0.1123
!Fit model including all 34 cells
$use fitm$
 Enter range of cells to be included and lower bound  :-
$DIN? 1 34 4.209
-- model re-initialised
scaled deviance = 22.818 at cycle  3
           d.f. = 32

 $di e$
         estimate        s.e.     parameter
    1       4.479    0.002295     MEAN
    2      0.1283    0.002241     SD
    scale parameter taken as  1.000
```

The estimates for the model are as given in [46], except that our formulation also gives asymptotic standard errors of the estimates. Note that the degrees of freedom displayed in the fit should be reduced by one since the *GLIM* package does not take into account the constraint that the sum of observed frequncies and fitted values are equal under this formulation.

Example 4.2 Mixture Distributions

Hasselbad, [36] describes a convenient method due to Hald, [34], for the calculation of starting values for fitting mixture distributions to grouped data based on fitting a sequence of truncated distributions. Briefly, we assume there exist a number of cut-off points so that nearly all the sample of the $(j+1)$th component of the mixture distribution lies to the right of one of these points and some of the jth component lies to the left. Since, for the smallest cut-off point, only the observations from the first component will lie to its left, we can estimate its mean and variance by fitting a truncated distribution at this point. Given these estimates, the expected frequency lying beyond the first cut-off point can be subtracted from the observed frequency, and the process repeated for the remaining components in a similar way.

This procedure has been exploited by Wilcox and Russell, [76], who fitted a series of truncated distributions in order to estimate the proportion of excess births in the lower tail of birthweight distributions.

4.3 Mixture Distributions

Mixture models can arise in a variety of situations. For example, in [45], a two component mixture model was fitted to grouped, truncated data using the EM algorithm when analysing the volume of red blood cells. Aitkin, [2], has considered the analysis of mixture distributions using the EM algorithm in GLIM.

An important applicaton of the methods described in this section is the analysis of birthweight, which various studies have analysed by assuming a predominately Normal distribution for the majority of births, but with additional births in the lower tail - an obvious application of some form of mixture distribution; see, for example Pethybridge et al., [52].

We show how grouped data having an underlying finite mixture distribution can be specified as a composite link function. Moreover, the flexibility of the formulation means that it is straightforward to consider ,for example, mixtures of different distributions or of truncated distributions. Further,

the model specification facilities of GLIM , in particular, the use of factors, enable models to be fitted which may be classified by variables such as social class or nationality.

4.3.1 The Composite Link Formulation

We will assume that observations, x, which appear in grouped form, arise from an underlying mixture distribution with, in general, r components. Thus, the distribution function of an observation x is given by,

$$F(x) = \sum_{i=1}^{r} p_i F_i(x, \mu_i, \sigma_i) \ , \tag{4.7}$$

where $0 < p_i < 1$, $i = 1, \ldots, r, \sum_{i=1}^{r} p_i = 1$, and μ_i and σ_i, i=1,...,r, represent location and scale parameters respectively.

Suppose that the N original observations have been formed into n categories, with the upper bound of each category given by u_j, $j = 1, \ldots, n$ and the lower bound by u_0. We can assume that the number of observations in each category, y_j, has a Poisson distribution with mean, θ_j, given by $\theta_j = N(F(u_j) - F(u_{j-1}))$, j=1,...,n. Note that, unless $F(u_n) = 1$ and $F(u_0) = 0$, we regard N as an extra unknown parameter to be estimated. However, if necessary, this can be easily accommodated in the formulation given below.

We can represent this as a composite link function by writing,

$$\begin{aligned} E[Y] &= \underline{\ell} \\ &= N\left[\sum_{i=1}^{r}(C(p_i)\underline{\gamma}_i)\right] \end{aligned} \tag{4.8}$$

where $C(\alpha)$, say, is an $nx(n+1)$ matrix of the form,

$$C(\alpha) = \begin{bmatrix} -\alpha & \alpha & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & -\alpha & \alpha \end{bmatrix}$$

and $\underline{\gamma}_i = F_i(\underline{\eta}_i)$, where $\eta_{ij} = (u_j - \mu_i)/\sigma_i$, for i=1,...,r and j=0,...,n. This model forms a parametric composite link function because the link matrix, C, contains the values of the mixing parameters, $\underline{p}$, as extra unknown parameters .

The principle of the estimation method is described in detail in section 4.1 , it suffices here to say that we fit a model with Poisson errors, iterative explanatory variable, X^*, given by

$$X^* = [CHX \ : \ D\underline{p} \ : \ DN] \tag{4.9}$$

where,

$$CHX = \left[C(p_1)H_1X \ \dots \ C(p_r)H_rX \right]$$

$$D\underline{\ell} = \left[\partial\underline{\ell}/\partial p_1 \ \ldots \ \partial\underline{\ell}/\partial p_{r-1} \right]$$

$$DN = [\partial\underline{\ell}/\partial N] \quad \text{and,}$$

$$H_i = \operatorname{diag}\left[\partial\underline{\ell}_i/\partial\underline{\eta}_i\right] \quad i = 1, \ldots, r.$$

As with the truncated distributions considered in the previous Section, it is relatively straightforward to consider estimating $\underline{\mu}$ and $\underline{\sigma}$ directly by using an iterative design matrices of the form,

$$X_i^T = \begin{bmatrix} -1/\sigma_1 & \ldots & -1/\sigma_1 \\ (\mu_1 - u_0)/\sigma_1^2 & \ldots & (\mu_1 - u_n)/\sigma_1^2 \\ \vdots & & \vdots \\ -1/\sigma_r & \ldots & -1/\sigma_r \\ (\mu_r - u_0)/\sigma_r^2 & \ldots & (\mu_r - u_n)/\sigma_r^2 \end{bmatrix}$$

This is the same approach as was taken in section 4.2.1 when estimating the parameters for a truncated distribution. As before we calculate the values of $\partial\underline{\eta}/\partial\underline{\mu}$ and $\partial\underline{\eta}/\partial\underline{\sigma}$ at each iteration.

The principal advantage of the composite link formulation is that models can be simultaneously fitted to data classified by some factor, and parameter estimates calculated for each level of the factor.In order to do this in GLIM , suppose A represents a factor with a levels , then, to estimate parameters for

each level of the factor, we fit the design matrix $A.X^*$. This is a technique first considered by Roger, [59]. A consequence of this is that we can constrain the values of certain parameters and, hence, test whether they are the same for each level of the factor.

4.3.2 Starting Values

We have experienced difficulty in convergence of the algorithm for some data sets, in which the component distributions are not well separated, when the starting values are not close to their final estimates. The high correlation between parameter estimates means that small changes in the values of the mixture parameters can lead to large changes in the values of the other parameters.

In this case there seem to be two possibilities. Firstly, it is possible to estimate the mixing parameters and the location and scale parameters in a two stage algorithm, i.e. fixing the value of $\underline{p}$, finding estimates of $\underline{\mu}$ and $\underline{\sigma}$ and repeating this for a grid of values of $\underline{p}$. Once the approximate maximum likelihood estimate of $\underline{p}$ has been found, the full model can be fitted. Alternatively, the mixing parameters $\underline{p}$ can be constrained to lie in the interval (0,1) by writing $p_i = \exp(\delta_i)/(1+\exp(\delta_i))$,i=1...,r and estimating the δ_i. If necessary, once approximate convergence has been obtained, the

algorithm can switch back to estimating p directly.

Starting values for the parameter estimates can be found using a variety of methods. For example, Bhattachyarya, [6], describes a simple graphical method for Normal mixtures. For the logistic distribution, initial values can be found by assuming Normality and scaling the variance estimates. An obvious alternative in the two component case are moment estimators with observations replaced by class midpoints.

4.3.3 Examples

Example 4.3 Analysis of Blood Cell Volume

Mclachlan and Jones, [45], considered an analysis of the volume of cows' red blood cells following innoculation by a parasite which causes a form of anaemia. The observed counts of red blood cell volumes 21 and 23 days after innoculation are listed in table 4.2. Note that the grouping intervals are slightly different to those published (private communication), in that the lower and upper truncation values are 28.8 fl and 158.8 fl respectively, and the grouping is over 18 intervals of equal width 7.2 fl.

Lower	Time after Innoculation		Lower	Time after Innoculation	
Bound	21 Days	23 days	Bound	21 Days	23 days
28.8	10	9	93.6	54	67
36.0	21	32	100.8	53	44
43.2	51	64	108.0	54	36
50.4	77	69	115.2	44	30
57.6	70	56	122.4	36	24
64.8	50	68	129.6	29	21
72.0	44	88	136.8	21	14
79.2	40	93	144.0	16	8
86.4	46	87	151.2	13	7

Table 4.2: Red Blood Cell Volumes

In [45] a two component log-Normal mixture distribution was fitted to the observed frequencies. The assumption of a mixture distribution was justified informally by probability plots, in particular, the $\Phi - p$ versus Q proposed by Fowlkes, [25]. These plots indicate the possible presence of a Normal mixture, even for the second data set, where there is no clear evidence of bimodality in the histogram.

Figure 4.1 shows the results of Bhattacharya's graphical technique for finding parameter estimates for the 21 days group. Briefly, the configuration of the plot results in the grouping of points indicating the number of components in the mixture. A straight line fitted through the ith group, $i = 1, \ldots, r$, results in parameter estimates given by Equation 4.11,

$$\hat{\mu}_i = -\frac{\beta_{i0}}{\beta_{i1}} + \frac{w}{2} \tag{4.10}$$

$$\hat{\sigma}_i^2 = -\frac{w}{\beta_{i1}} - \frac{w^2}{12}, \tag{4.11}$$

where β_{i0} and β_{i1} are the intercept and slope of the lines fitted through the ith group respectively, and w is the interval width.

Application of the formulae in Equation 4.11 for this data gives the estimates $\underline{\hat{\mu}}^T = (4.109, 4.628)$ and $\underline{\hat{\sigma}}^T = (0.300, 0.301)$. Note that, since the intervals have been transformed to the log scale, we have taken w to be the width of the middle interval on the log scale. These estimates may be used as

	Time After Innoculation			
	21 Days		23 Days	
Parameter	Estimate	s.e.	Estimate	s.e.
p	0.46	(0.052)	0.17	(0.050)
μ_1	4.08	(0.038)	3.86	(0.048)
σ_1	0.24	(0.024)	0.17	(0.029)
μ_2	4.72	(0.024)	4.47	(0.026)
σ_2	0.20	(0.022)	0.28	(0.022)

Table 4.3: Estimates - Blood Cell Data (Full Model)

starting values for the maximum likelihood estimates. Parameter estimates for the complete data are easily found using the techniques described in section 4.3.1 by specifying a factor with two levels representing the two time periods. The estimates are given in Table 4.3 with asymptotic standard errors in brackets.

The deviance for this model is 15.87 on 24 degrees of freedom. Of obvious interest is a test of whether the value of the mixing parameter is different over the two time periods. A comparison of the approximate confidence intervals calculated from the asymptotic standard errors suggests that this is highly likely. For example, approximate 95% confidence intervals for the

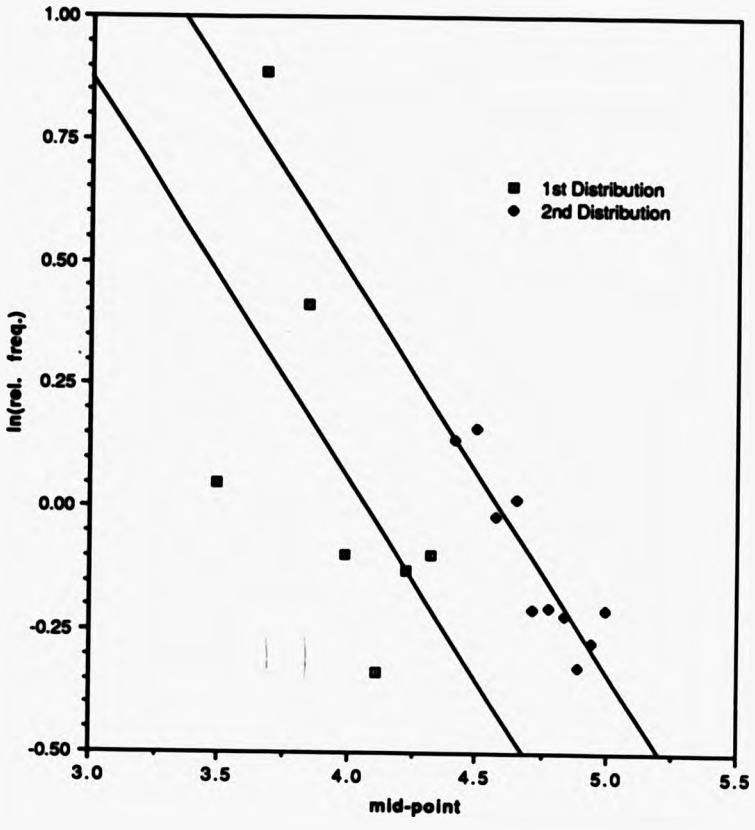

Figure 4.1: Graphical method for estimation of mixture distributions

	Time After innoculation			
	21 Days		23 Days	
Parameter	Estimate	s.e.	Estimate	s.e.
p	0.44	(0.052)		
μ_1	4.07	(0.038)	4.11	(0.044)
σ_1	0.24	(0.024)	0.30	(0.030)
μ_2	4.71	(0.024)	4.56	(0.025)
σ_2	0.21	(0.022)	0.24	(0.020)

Table 4.4: Estimates - Blood Cell Data (constrained Model)

mixing paramaters are (0.36,0.56) and (0.07,0.26) for 21 and 23 days after innoculation respectively. However, in line with the discussion in section 2.4, a preferable method is to fit a model constraining the mixing parameters to be equal in the two time periods and examine the change in deviance. Using the formulation described above, this is easily achieved by setting a different option in the GLIM macros described in Appendix A.2. The parameter estimates for this model are shown in Table 4.4. with deviance 25.12 on 25 degrees of freedom. Thus referring the difference in deviances between the two models, $25.12 - 15.87 = 9.25$, to a χ^2_1 distribution, we can clearly reject the hypothesis of equal mixing parameters ($p < 0.01$).

Weight (kgs)	UK	Asian	Weight (kgs)	UK	Asian
≤ 0.75	7	1	2.5–2.75	157	26
0.75–1.0	14	2	2.75–3.0	118	17
1.0–1.25	20	1	3.0–3.25	77	8
1.25–1.5	26	1	3.25–3.5	41	1
1.5–1.75	58	6	3.5–3.75	14	1
1.75–2.0	56	10	3.75–4.0	6	0
2.0–2.25	96	15	> 4.0	1	0
2.25–2.5	156	19			

Table 4.5: UK and Asian birthweight

Example 4.4 Analysis of Birthweight

The data in table 4.5, collected at the Department of Clinical Epidemiology, London Hospital Medical College, shows the frequency distribution of birthweight for multiple births to both UK born and Asian born mothers.

The deviances for fitted single component distributions are displayed in Table 4.6. In each of the models fitted to these data, the birthweights were truncated at 0kgs and 6kgs respectively. Examination of the deviance values shows that the births to Asian mothers, in particular, are fitted reasonably well by a single component distribution, although this is not the case for

	Distribution	
	Normal	Logistic
UK mothers	55.17	48.52
Asian mothers	15.32	12.62

Table 4.6: Deviances for single component fits-Birthweight data

births to English mothers. The assumption of an underlying mixture distribution for birthweight has already been mentioned in Section 4.3. However, for this data set, examination of the histograms does not reveal an obvious mixture distribution, although Fowlkes plots, as shown in Figures 4.2 and 4.3, suggest the possibility of a two component Normal mixture distribution. For purposes of comparison, two component Normal and Logistic mixture models were fitted to the data- the parameter estimates for both models are shown in table 4.7. The deviances for these two models are 16.779 and 15.805, for the Logistic and Normal distributions respectively, both on 18 degrees of freedom. Although there seems little to choose between the models in terms of goodness of fit, examination of the parameter estimates shows the assumption of a mixture distribution for births to Asian mothers is less clear. In particular, an approximate 95% confidence for the mixing parameter in this case includes the value zero.

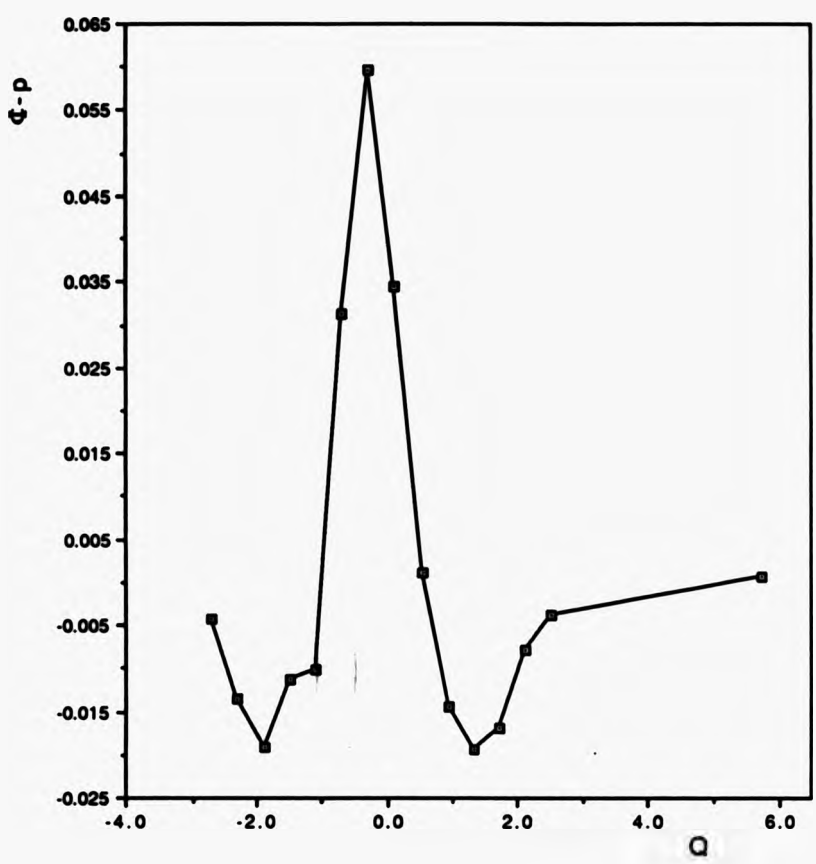

Figure 4.2: Fowlkes plots for Birthweight Data-UK Mothers

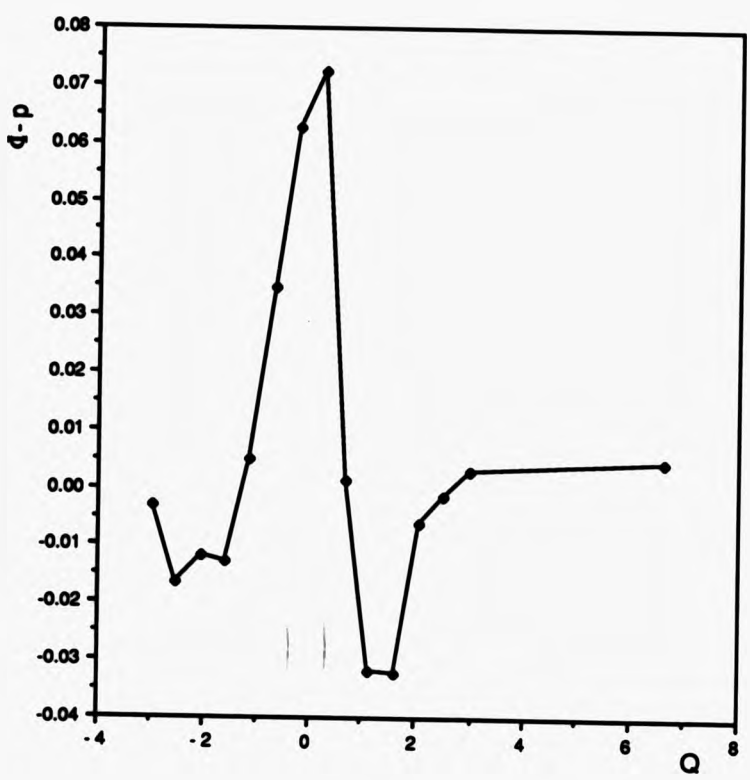

Figure 4.3: Fowlkes plots for Birthweight Data-Asian Mothers

	U.K.		Asian	
	NORMAL DISTRIBUTION			
Parameter	Estimate	s.e.	Estimate	s.e.
p	0.13	(0.0617)	0.43	(0.221)
μ_1	1.42	(0.257)	2.06	(0.254)
σ_1	0.44	(0.122)	0.62	(0.092)
μ_2	2.58	(0.054)	2.61	(0.078)
σ_2	0.48	(0.030)	0.32	(0.078)
	LOGISTIC DISTRIBUTION			
Parameter	Estimate	s.e.	Estimate	s.e.
p	0.17	(0.063)	0.42	(0.315)
μ_1	1.59	(0.184)	2.01	(0.346)
σ_1	0.30	(0.060)	0.32	(0.065)
μ_2	2.56	(0.039)	2.65	(0.092)
σ_2	0.27	(0.016)	0.18	(0.062)

Table 4.7: Parameter Estimates for Birthweight data

Example 4.5 Wind Shear Data

Kanji, [41], describes the analysis of wind shear data and fits a Laplace/Normal mixture distribution using the method of minimum chi-square. The data set is quite extensive and only the first eight cases are analysed as an illustration. These are listed in Appendix D. A graph illustrating the distributional form of the data, in fact case numbers 1, 2, 3 and 4, is given in Figure 4.4. Note that, in order to avoid problems of zero fitted values, the original data was re-grouped so that the extreme intervals contained a frequency of at least 5 observations. The underlying distribution of the data is assumed to be of the form given in Equation 4.12,

$$f(x;p,\mu\sigma) = \frac{p}{\sigma\sqrt{2}}e^{(-\frac{\sqrt{2}}{\sigma}|x-\mu|)} + \frac{(1-p)}{\sigma\sqrt{2\pi}}e^{(-\frac{1}{2\sigma^2}(x-\mu)^2)} \qquad (4.12)$$

This model can be fitted using the composite link formulation described in Section 4.3.1 , although a few amendments are required since the parameters of the two component distributions are assumed to be equal. Moreover, the data itself is classified by two factors, the Band width and the gradient separation time interval, H_s. In [41], it was suggested that the mixing parameter might be consistent across band widths for the different case numbers. Using the composite link formulation, we can test this assumption for both the Band width and gradient separation time interval.

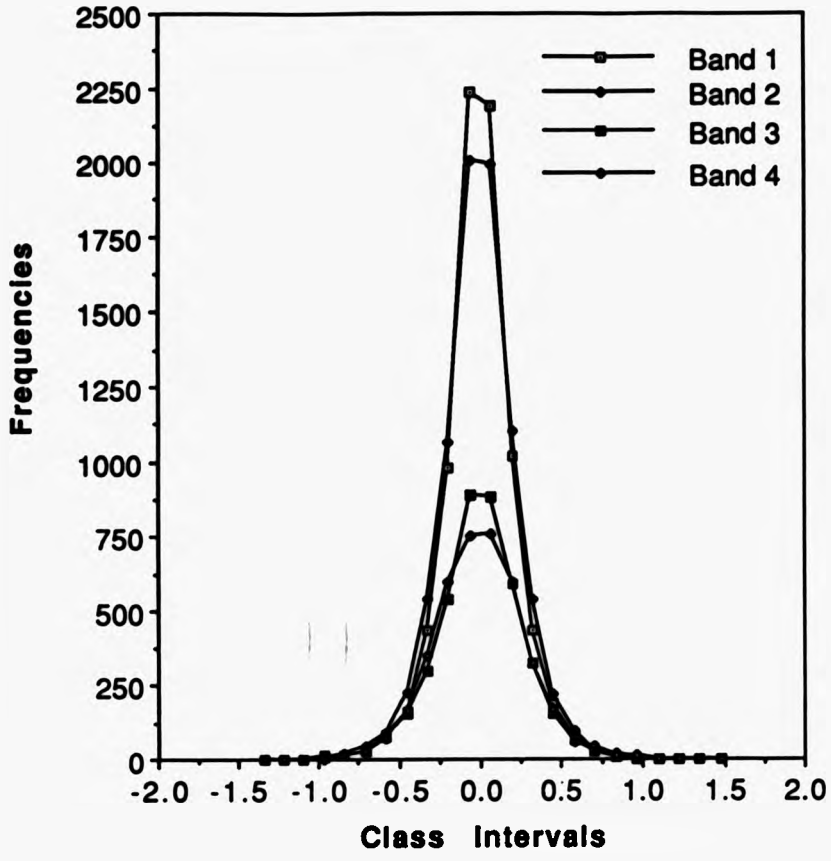

Figure 4.4: Observed frequencies for Wind Shear data, Cases 1,2,3 and 4

Band Width	Gradient	Parameter μ	σ	p
1	1	0.0003	0.2111	0.8581 (0.0405)
2	1	0.0002	0.2339	0.6806 (0.0390)
3	1	0.0022	0.2629	0.5203 (0.0527)
4	1	-0.0003	0.2767	0.3065 (0.0515)
1	2	0.0008	0.1954	0.8383 (0.0463)
2	2	0.0005	0.2160	0.6870 (0.0446)
3	2	-0.0023	0.2405	0.4210 (0.0589)
4	2	0.0012	0.2496	0.2516 (0.0574)

Table 4.8: Parameter Estimates for Wind Shear Data - full model

The parameter estimates after fitting the distribution in Equation 4.12 to the complete data set are listed in Table 4.8. The deviance for this model is 101.07 on 78 degrees of freedom and is a good fit to the data. We can use the composite link approach to test whether other models may fit the data equally well. In this context we are only interested in the mixing parameters, although, it should be noted that we could equally well consider the other parameters in the model. The results of these fits are illustrated in Table 4.9. These results clearly bear out the assumption that the mixing

Model	Common	H_s	Band	$H_s \times$ Band
Deviance	252.71	252.01	102.79	101.07
DF	85	84	82	78

Table 4.9: Deviances for models fitted to Wind Shear data

parameters are consistent across band widths for the different case numbers. The deviance difference for this test is $102.79 - 101.07 = 1.72$ on 4 degrees of freedom, clearly not significant. However, it is clear that the gradient separation time plays little part in explaining the variability between the distributional forms for the different case numbers. The estimates of the mixing paramteters assuming a common value across cases are 0.8495 (0.0305), 0.6834 (0.0293), 0.4671 (0.0393) and 0.2822 (0.0384) for Bands 1, 2, 3 and 4 respectively.

Chapter 5

Analysis of Multivariate Data

5.1 Normal Distribution models

In this section we show how the iterative weighted least squares algorithm can be generalised to include situations in which the observations are not independent, although it is assumed that the random component of the model is Normally distributed.

In general we assume that observations $y_1, \ldots, y_n$ are Normally distributed with $E[\underline{Y}] = \underline{\mu}$ and $\text{var}(\underline{Y}) = \sigma^2\Sigma$, i.e $\underline{Y} \sim \mathcal{N}(\mu, \sigma^2\Sigma)$. We also

assume , as usual, that the mean of y_i, μ_i is related to the linear predictor, $\eta_i = x_i^T \underline{\beta}$ by the link function g, i.e.for $i = 1, \ldots, n$, $\eta_i = g(\mu_i)$.

Following the same arguments as in section 2.3, it can be shown that the maximum likelihood estimates of the parameter vector, $\underline{\beta}$ may be found iteratively using,

$$\underline{\hat{\beta}}_{r+1} = \left(X^T F_r (\sigma^2 \Sigma_r)^{-1} F_r X\right)^{-1} X^T F_r (\sigma^2 \Sigma_r)^{-1} F_r \underline{z}_r \qquad (5.1)$$

where $F = \text{diag}[\delta_i^{-1}]$, $\delta_i = \frac{\partial \eta_i}{\partial \mu_i}$ and $z_i = \eta_i + (y_i - \mu_i)\delta_i$ for $i = 1 \ldots, n$ so that $\underline{z}$ the usual iterative dependent variable. Note that, if $\Sigma = I_n$, this is the weighted least squares routine for independent data with iterative weight matrix given by $F = \text{diag}[(\sigma^2 \delta_i^2)^{-1}]$.

Now, since Σ is a covariance matrix, and hence positive definite, so must be Σ^{-1}.This means we can find its Cholesky decomposition, i.e. a triangular matrix U such that $\Sigma^{-1} = U^T U$. Substituting for Σ^{-1} in equation 5.1 gives

$$\underline{\hat{\beta}}_{r+1} = \left((U_r F_r X)^T V_r (U_r F_r X)\right)^{-1} (U_r F_r X) V_r U_r F_r \underline{z}_r \qquad (5.2)$$

where $V = \text{diag}[1/\sigma^2]$.

Thus, $\underline{\hat{\beta}}$ can be found using weighted least squares with working independent variable $X^* = UFX$ and working dependent variable, $Z^* = UF\underline{z}$. This formulation is analagous to a composite link formulation except that, in this case, the link has essentially arisen from a transformation of the data.

Consider a linear transformation of the form $\underline{W} = U\underline{Y}$, such that var($\underline{W}$) = $\sigma^2 I_n$. Then, a $\underline{W} \sim \mathcal{N}(U\underline{\mu}, \sigma^2 I_n)$, so that a model with dependent variable, $\underline{w}$, would be fitted with linear composite link , U, in the usual way. Note that, in general, the matrix U will contain unknown parameters so that the model is a parametric composite link function, with the parameters estimated by a two-stage algorithm.

An important feature of this formulation is that, in many useful cases, the matrix U has a relatively simple form . However, although the method described above will, in principle, work for any positive definite covariance matrix, its applicability depends on two issues. Firstly, the ease with which the square root matrix, U can be calculated and, secondly, the calculation of the working design matrix, UFX. Since, this algorithm was first implemented in GLIM by Scallan, [62], and these issues highlighted, various authors have discussed the practical difficulties involved in its implementation. For example, when the design matrix, X, contains factors, Candy, [11], has suggested a method using logical operators rather than $(0, 1)$ values to represent factor levels. O'Brien, [50], has suggested swithching between the statistical package, used for data and model definition, and a subroutine library, for the numerical calculations and matrix manipulations, by utilising a 'pause' facility.

5.1.1 Applications

In this section we illustrate the application of the model discussed above in some special, but important, cases which may be easily programmed in GLIM without resort to the techniques discussed at the end of Section 5.1.

Example 5.1 Autoregressive Processes

Consider the first order autoregressive process defined by,

$$Y_i - \mu_i = \rho(Y_{i-1} - \mu_{i-1}) + \epsilon_i \tag{5.3}$$

where $|\rho| < 1$, $\mathrm{E}[Y_i] = \mu_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, for $i = 1, \ldots, n$ and, as a starting condition, $Y_0 \sim \mathcal{N}(\mu_0, \sigma^2/(1-\rho^2))$. For this model

$$\mathrm{cov}(Y_i, Y_j) = \sigma^2 \frac{\rho^{|i-j|}}{(1-\rho^2)}$$

It is easy to show that the matrix U defined by

$$U = \begin{bmatrix} \sqrt{(1-\rho^2)} & 0 & \ldots & 0 \\ -\rho & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & -\rho & 1 \end{bmatrix}$$

satisfies, $U^T U = \Sigma^{-1}$, and thus forms an appropriate transformation or link matrix for data from a first-order autoregressive process. In particular, the

application of the matrix to any vector is effectively a differencing operation which is computationally easy to perform.

Note that an approximate maximum likelihood estimate of ρ suggested by Box and Jenkins , [7] , is given by

$$\hat{\rho} = \frac{\sum_{i=2}^{n}(y_i - \hat{\mu}_i)(y_{i-1} - \hat{\mu}_{i-1})/(n-1)}{\sum_{i=2}^{n-1}(y_i - \hat{\mu}_i)^2/(n-2)} \tag{5.4}$$

which is the estimate given by minimising the quantity $(\underline{y} - \underline{\hat{\mu}})^T \Sigma^{-1} (\underline{y} - \underline{\hat{\mu}})$ with respect to ρ. In general, the variance matrix of $\underline{\hat{\beta}}$ and A will have to be adjusted to take into account the estimation of ρ using the procedure outlined in Section 3.1.1.

The joint estimates of ρ and $\underline{\beta}$ can be found using a two-stage procedure. Thus, for a fixed value of *rho*, we estimate $\underline{\beta}$, and then update the value of ρ using the current fitted values according to Equation 5.4.

As an example, consider the (log) logistic curve defined by

$$\mathrm{E}[\ln(y_i)] = \mu_i = \ln(A) - \ln(1 + \exp(-\eta_i))$$

,where $\eta_i = \beta_0 + \beta_1 t_i$ for $i = 1, \ldots, n$, fitted to data following a first-order autoregressive process. The variance stabilising logarithmic transformation was used by Nelder [47] and may be further justified, in this case, by the method of data collection which involved the estimation of leaf area from

Day	1	2	3	4	5	5	7	8	9
Area	0.745	1.047	1.695	2.428	3.664	4.022	5.447	6.993	8.221
Day	10	11	12	13	14	15	16	17	18
Area	8.829	10.080	12.971	14.808	17.341	19.705	22.597	24.537	25.869
Day	19	20	21	22	23	24	25	26	27
Area	27.816	29.595	30.451	30.817	32.472	32.999	33.555	34.682	34.682
Day	28	29	30	31	32	33	34	35	36
Area	35.041	35.356	35.919	36.058	36.454	36.849	37.200	37.200	37.200

Table 5.1: Bean Plant Leaf Area

the product of measured leaf length and breadth. This model has also been considered by Glaseby, [31], although he employed a direct Newton-Raphson approach rather than the IRLS model described above. This transform both sides approach has been studied in the independence case by Carroll and Ruppert, [13], although the extension to the case of non-independence needs further investigation, particularly in terms of the effect on the transform of the dependence structure of the observations. The data in Table 5.1 shows the area of bean plant leaf over a 36 day period.

In [62], a logistic curve was fitted to the data which gave an a value of $\hat{\rho} = 0.8177$. However, Ross (private communication), has suggested that a

Parameter	Estimate	s.e.
ρ	0.5358	0.0198
β_0	-4.1321	0.5111
β_1	0.1424	0.0182
A	39.728	1.9312
γ	6.256	4.0230

Table 5.2: Parameter estimates - Generalised Logistic curve

generalised logistic curve of the form

$$E[\log(Y)] = \mu = \ln(A) = \gamma \ln[1 + \exp(-\eta)/\gamma]$$

where $\gamma > 0$ would provide a better fit to the data. The parameter estimates for this model are shown in Table 5.2. The standard errors have been adjusted to take account of the estimation of ρ using the technique outlined in [62].

Example 5.2 Voting Behaviour

Forcina, [24], has suggested a model for voting behaviour which utilises the structure of the covariance matrix in an efficient way. Briefly, we assume $\underset{\sim}{Y} \sim \mathcal{N}(\underset{\sim}{\mu}, \sigma^2 V)$, where

$$V = \text{diag}[I + dD_1, \ldots, I + dD_r]$$

d is an unknown parameter and D_i is an n_i by n_i matrix of known constants. By writing $D_i = P_i^T H_i P_i$, where H_i is a diagonal matrix of the eigenvalues of D_i and $P = \text{diag}[P_1, \ldots, P_r]$, then $P\underline{y} \sim \mathcal{N}(P\underline{\mu}, \sigma^2 \text{diag}\,(W))$ where $W^{-1} = \text{diag}[I + dH_1, \ldots, I + dH_r]$

5.2 Analysis of non-Normal models

5.2.1 Introduction

Dependent observations in which the error distribution is non-Normal can arise naturally in a variety of situations. In Example 5.2.1 a commonly used experimental situation in plant physiology is described in which the resulting Binomial observations have a particular covariance structure. Following Cox, [17], this is an example of a process driven model in which the conditional distribution of observation Y_t is specified as a function of past observations, $Y_{t-1}, \ldots, Y_1$. In the alternative parameter driven models, dependence between successive observations is introduced through a latent process.

Example 5.3 Process Driven Model

As an example, consider the following experiment resulting from a trial of seed germination rates. A given number, N, seeds are sown and, on each

day, the number germinating, of those so far not germinated, counted.

Let Y_i = number germinating on day i, for $i = 1,\ldots,r$ then

$$Y_i \mid Y_{i-1},\ldots,Y_1 \sim B\left(N - \sum_{j=1}^{i-1} Y_j, p_i\right)$$

where p_i is the probability that a seed germinates on day i. Of interest is the relationship between p_i and possible explanatory variables such as growth inhibitors. The joint distribution of $Y_1,\ldots,Y_r$ is given by,

$$\begin{aligned} f(Y_1,\ldots,Y_r) &= f(Y_r \mid Y_{r-1},\ldots,Y_1).f(Y_{r-1} \mid Y_{r-2},\ldots,Y_1)\ldots f(Y_2 \mid Y_1).f(Y_1) \\ &= \prod_{i=1}^{r} {}^{(N-\sum_{j=1}^{i-1} Y_j)}C_{Y_i}\, p_i^{Y_i}(1-p_i)^{N-\sum_{j=1}^{i-1} Y_j} \end{aligned}$$

which is may be regarded as the product of Binomial observations conditional on the preceeding total of germinated seeds. This means, in particular, that we can find parameter estimates by treating the $Y_j \mid Y_{j-1},\ldots,Y_1$ $j = 1,\ldots,r$ as independent Binomial variables.

The underlying covariance structure of the distribution can best be illustrated by re-arranging the distribution to give,

$$f(Y_1,\ldots,Y_r) = \frac{N!}{\prod_{i=1}^{r} y_i!(N-\sum_{j=1}^{r} Y_j)!} \prod_{i=1}^{r} \left[p_i \prod_{j=1}^{i-1}(1-p_j)\right]^{y_i} \left[\prod_{i=1}^{r}(1-p_i)\right]^{N-\sum_{j=1}^{r} Y_j}$$

which is effectively the density function of a multinomial distribution. Using this result, the moment generating function is easily seen to be,

$$M([t]) = \left(\sum_{i=1}^{r} p_i e^{t_i} \prod_{j=1}^{i-1}(1-p_j) + \prod_{i=1}^{r}(1-p_i)\right)^N$$

from which moments are easily shown to be

$$\begin{aligned}
\mathrm{E}[Y_k] &= N p_k \prod_{j=1}^{k-1}(1-p_j) \\
\mathrm{var}(Y_k) &= N p_k \prod_{j=1}^{k-1}(1-p_j)(1-p_k\prod_{j=1}^{k-1}(1-p_j)) \\
\mathrm{cov}(Y_k, Y_l) &= -N p_k p_l \prod_{j=1}^{k-1}(1-p_j)\prod_{j=1}^{l-1}(1-p_j)
\end{aligned}$$

An example of the application of this model to the analysis of grouped survival data is given by Candy, [12].

Example 5.4 Parameter Driven Model

Zeger, [78], proposed the following model for time series of counts. The approach used was to assume a parameter driven model in which, conditional on a latent process ϵ_t, Y_t is an independent sequence of counts with mean and variance given by,

$$m_t = E[Y_t \mid \epsilon_t] = \exp(\eta_t)\epsilon_t, \qquad w_t = \mathrm{var}(y_t \mid \epsilon_t) = m_t$$

It was assumed that ϵ_t is a stationary process with $E[\epsilon_t] = 1$ and $\mathrm{cov}(\epsilon_t, \epsilon_{t+\tau}) = \sigma^2\rho_\epsilon(\tau)$. Then the full unconditional moments of Y_t are,

$$\begin{aligned}
\mu_t = E[Y_t] &= \exp(\underline{x}_t^T\underline{\beta}) \ , \\
v_t = \mathrm{var}(Y_t) &= \mu_t + \sigma^2\mu_t^2 \ , \\
\rho(t,\tau) = \mathrm{corr}(Y_t, Y_{t+\tau}) &= \frac{\rho_\epsilon(\tau)}{[\{1+(\sigma^2\mu_t)^{-1}\}\{1+(\sigma^2\mu_{t+\tau})^{-1}\}]^{\frac{1}{2}}}
\end{aligned}$$

The analysis of this model is discussed further below.

5.2.2 Analysis

Several methods of analysis for such models have been suggested in the literature. In this chapter we will primarily be concerned with the extension of the quasi-likelihood model to the multivariate case discussed by McCullagh, [43]. Briefly, we assume that the nx1 vector of random variables $\underline{Y}$ has mean $\underline{\mu}$ and covariance matrix $\phi V(\underline{\mu})$. As usual, the mean is related to the linear predictor by the link function $\underline{\eta} = g(\underline{\mu})$. Then the log quasi-likelihood, considered as a function of $\underline{\mu}$, is given by the system of partial differential equations,

$$\frac{\partial K(\underline{\mu};\underline{y})}{\partial \underline{\mu}} = V^{-}(\underline{\mu})(\underline{y} - \underline{\mu}) \tag{5.5}$$

where V^{-} is some generalised inverse. However, the use of such models does not appear to have been widespread. One reason for this is that the existence of the K function for a particular variance function is not guaranteed. However, as several authors have noted, for example Nelder [48], such equations are likely to be optimal estimating equations.

In Gilchrist and Scallan, [30], a special case of this formulation was

proposed. We assume the covariance matrix $V(\mu)$ can be written as

$$V(\mu) = V^{\frac{1}{2}}\Sigma V^{\frac{1}{2}} \quad , \tag{5.6}$$

where $V^{\frac{1}{2}} = \text{diag}[v(\mu_i)^{\frac{1}{2}}]$ and Σ is a correlation matrix, so that we are essentially modelling the mean/variance relationship and the correlation structure separately.

This formulation lends itself naturally to an extension of the estimation procedure outlined in Section 5.1 with few amendments. Again writing $U^TU = \Sigma^{-1}$, the maximum quasi-likelihood estimate of $\underline{\beta}$ is found using iterative design matrix

$$X^* = UV^{-\frac{1}{2}}FX \quad \text{and dependent variable,} \tag{5.7}$$

$$Z^* = UV^{-\frac{1}{2}}F_{\underline{z}} \tag{5.8}$$

where, as before, $z_i = \eta_i + (y_i - \mu_i)\delta_i$. Note that this formulation invloves the weighting and transformation of the dependent variable within cycles of the fitting algorithm, although, in GLIM, this does not present a problem. Mcros for fitting these models are given in Appendix C.

This is exactly the approach followed by Zeger in the analysis of the latent process model for time series of counts. The true unconditional covariance matrix of the observations was approximated by a covariance matrix

of the form given by Equation 5.6 and led to what was called an iterative weighted and filtered least squares procedure.

In general, the correlation matrix, Σ, will contain unknown parameters to be estimated. For the first order autoregressive process, with

$$\text{corr}(Y_i, Y_j) = \frac{\rho^{|i-j|}}{(1-\rho^2)}$$

a natural estimate, generalising the estimate in the Normal case, is given by

$$\hat{\rho} = \frac{\sum_{i=2}^{n} r_i r_{i-1}}{\sum_{i=2}^{n-1} r_i^2} \tag{5.9}$$

where $r_i = (y_i - \hat{\mu}_i)/v_{ii}^{1/2}$. As an estimate of ϕ, McCullagh suggests using

$$\hat{\phi} = (\underline{y} - \underline{\hat{\mu}})V^{-}(\underline{y} - \underline{\hat{\mu}})/(N-p) = X^2/(N-p) \tag{5.10}$$

where X^2 is a generalised form of Pearson's statistic. Alternatively, moment estimators may be used as in [78].

Example 5.5 Bean Plants

Consider again the data of Example 5.1 which showed the growth of bean plant leaves against time. We will assume that the mean of Y_i can be represented by the Generalised logistic curve of the form,

$$E[Y_i] = \mu_i = \frac{A}{(1+\exp(-\eta_i)/\gamma)^{\gamma}}$$

Parameter	Estimate	s.e.
ρ	0.376	
β_0	-4.258	0.3386
β_1	0.1470	0.0102
A	39.29	1.506
γ	5.407	2.140

Table 5.3: Parameter estimates - Gompertz curve

and that var($\underline{Y}$) is as specified in Equation 5.6 with $v(\mu_i) = \mu_i^2$ and $\Sigma_{ij} = \rho^{|i-j|}/(1-\rho^2)$. The parameter estimates for this model are given in Table 5.3, but note that standard errors have not been adjusted for the estimation of ρ. The estimates from this model compare favourably with the estimates from the model analysed in Section 5.1.1, although it is noticeable that the estimate of the autocorrelation coefficient is somewhat smaller for the untransformed data. The power paramater in both cases is poorly estimated, although, the estimation procedure for the untransformed model was more stable.

5.3 A Generalised Multivariate Logistic Distribution

In this section we introduce a specific model for the analysis of observations which arise from experiments involving repeated measures on the same experimental unit. A generalisation of the multivariate logistic distribution is introduced and its properties discussed, and a technique developed for fitting such distributions in a regression context by making use of an auxiliary multinomial model.

The practical reasons for introducing repeated measurements in experimental situations are many, but include the following,

- between subject variation may be so great that meaningful comparisons can only be made with difficulty between different treatments applied to different individuals.
- repeated measurements may be made in order to make efficient use of scarce or expensive experimental material.
- the change in a response over time may be of direct interest, for example in growth studies.

The analysis of repeated measurement experiments differ in the way they

describe the covariance structure between measurements in a time sequence, $Y_t : t = 1, \ldots, n$. The simplest approach assumes that,

$$\text{cov}(Y_t, Y_s) = \begin{cases} \phi & t = s \\ \phi\rho & t \neq s \end{cases} \tag{5.11}$$

This is known as the "uniform correlation" or "split-plot" model and is the approach we will concentrate on in this chapter. One justification for it is the usual random effects model which assumes a stochastic "individual" effect as well as the usual random experimental error. If we assume that the individual effect has mean 0 and variance σ^2 and the experimental error term has mean 0 and variance τ^2 then, in the notation of Equation 5.11, $\phi = \sigma^2 + \tau^2$ and $\rho = \sigma^2/(\sigma^2 + \tau^2)$.

In section 5.3.1 we introduce the multivariate logistic distribution which possesses the covariance structure described by equation 5.11. In section 5.3.4 we show how this distribution can be modelled in a statistical package such as GLIM by making use of an auxiliary multinomial model. This is a generalisation of the auxiliary binomial model used by Bennett and Whitehead, [5], used to fit the univariate logistic distribution. GLIM macros for fitting the distribution are listed in Appendix B.1, and examples of their use may be found in [63].

The model proposed in this section is similar in its results to one proposed by Crowder, [19]. The approach used there was to regard responses as being conditionally independent given the value of a random individual effect; the response having a Weibull distribution and the individual effect a gamma distribution. The full unconditional distribution has similar properties to the generalised multivariate logistic discussed in this chapter. In fact, a form of this distribution was independently suggseted by a referee in [19]. The advantage of our formulation is the use of the auxiliary multinomial model to fit the distribution, which makes it much easier to explore the fit of different models to data.

5.3.1 Distribution Function and Properties

The multivariate logistic distribution and its properties have been described in detail by Johnson and Kotz, [39]. The subject of this chapter is a generalised version of that distribution arrived at by introducing an extra shape parameter to the distribution.

Suppose the random vector, $\underline{Y}$, has distribution function defined by,

$$F_{\underline{Y}}(\underline{y}) = \frac{1}{[1 + \sum_{k=1}^{r} \exp(-\eta_k)]^{\gamma}} \qquad (5.12)$$

where $\gamma > 0$, $\underline{-\infty} < \underline{y} < \underline{\infty}$, $\eta_k = \phi_k y_k + \theta_k + \ln(\gamma)$ and $\phi_k > 0$ for

$k=1,\ldots,r$, then the vector $\underline{Y}$ will be said to follow a generalised multivariate logistic distribution .The case $\gamma = 1$ is the multivariate logistic described in [39] . In applications such a random vector will typically represent an $r \times 1$ vector of responses measured on the same individual or experimental unit.

A similar model was proposed by Cook and Johnson, [16], who described a family of distributions which included a form of generalised multivarite logistic distribution, amongst others, as a special case.

The formulation of the model given by Equation 5.12 assumes that the scale parameters, $\underline{\phi}$, are homogeneous over individuals but not necessarily over responses. This might reflect different experimental conditions or treatment combinations and will be discussed further in section 5.3.4.

In order to find the covariance structure of $\underline{Y}$, we can find the moment generating function, $M(\underline{t})$, of the standardised variable $x_k = \phi_k y_k + \theta_k, k = 1,\ldots,r$. This function is given by,

$$M(\underline{t}) = \mathrm{E}\left[e^{\underline{t}^T \underline{x}}\right] \tag{5.13}$$

where the density of $\underline{x}$ is given by

$$f_{\underline{X}}(\underline{x}) = \frac{C_0 \prod_{k=1}^{r} \frac{e^{-x_k}}{\gamma}}{\left[1 + \sum_{k=1}^{r} \frac{e^{-x_k}}{\gamma}\right]^{r+\gamma}} \tag{5.14}$$

where $C_0 = \frac{\Gamma(r+\gamma)}{\Gamma(\gamma)}$.

By substituting, $u_k = e^{-x_k}/\gamma$ in the expectation defined in Equation 5.13, it is easily seen that the required integral is given by,

$$M(\underline{t}) = \frac{C_0}{\gamma^{\sum t_k}} \int_0^\infty \cdots \int_0^\infty \frac{\prod_{k=1}^r u_k^{-t_k}}{[1 + \sum_{k=1}^r u_k]^{r+\gamma}} d\underline{u} \tag{5.15}$$

Utilisation of a standard result for beta functions, namely

$$\int_0^\infty t^{m-1} \frac{1}{[a+t]^{m+n}} dt = \frac{1}{a^n} \mathrm{B}(m,n) \tag{5.16}$$

gives a straightforward recursive procedure for the evaluation of the integral in Equation 5.15. For example, for $k = 1$, we have $m = 1 - t_1$, $n = r - 1 + \gamma + t_1$ and $a = 1 + \sum_{i=2}^r u_i$.

Repeated application of this result yields the moment generating function as,

$$M(\underline{t}) = \frac{1}{\gamma^{\sum_{k=1}^r t_k} \Gamma(\gamma)} \Gamma\left(\gamma + \sum_{k=1}^r t_k\right) \prod_{i=k}^r \Gamma(1 - t_k) \tag{5.17}$$

or, more usefully, the cumulant generating function defined by,

$$\begin{aligned} K(\underline{t}) &= \ln M(\underline{t}) \\ &= \sum_{k=1}^r \ln(\Gamma(1 - t_k)) && (5.18) \\ &\quad + \ln\left(\Gamma(\gamma + \sum_{k=1}^r t_k)\right) && (5.19) \\ &\quad - \sum_{k=1}^r t_k \ln(\gamma) - \ln(\Gamma(\gamma)) \end{aligned}$$

The joint cumulant of order $\underline{p} = (p_1, \ldots, p_r)$ is then obtained from differentation of $\mathcal{K}(\underline{t})$ (p_1 times with respect to t_1 etc.) at $\underline{t} = 0$ to give

$$\mathcal{K}_{\underline{p}} = \begin{cases} \begin{cases} \Psi(\gamma) - \Psi(1) - \ln(\gamma) & p_i = 1,\ p_j = 0\ \forall j \neq i \\ \Psi^{p_i-1}(\gamma) + (-1)^{p_i}\Psi^{p_i-1}(1) & p_i > 1,\ p_j = 0\ \forall j \neq i \end{cases} \\ \Psi^{p_\cdot-1}(\gamma), \text{ otherwise} \end{cases} \tag{5.20}$$

where $p_\cdot = \sum_{i=1}^{r} p_i$ and $\Psi^{(n-1)}(x) = \delta^n \ln(\Gamma(x))/\delta x^n$ is the polygamma function.

Using the above, it is easily seen that

$$\begin{aligned} \mathrm{E}[X_i] &= \Psi(\gamma) - \Psi(1) - \ln(\gamma) \\ \mathrm{var}(X_i) &= \Psi^1(1) + \Psi^1(\gamma) \\ \mathrm{cov}(X_i, X_j) &= \Psi^1(\gamma) \quad \text{for i,j=1,...,r} \end{aligned} \tag{5.21}$$

Thus observations are equicorrelated with coefficient given by

$$\rho(X_i, X_j) = \frac{\Psi^1(\gamma)}{\Psi^1(1) + \Psi^1(\gamma)} \tag{5.22}$$

the coefficient falling from 1 to 0 as $\gamma \to \infty$. Such equicorrelated structures are typical in basic repeated measurements models. Following, [19], these results suggest moment estimators for the general model with $y_i = (x_i - \theta_i)/\phi_i$ as follows. Suppose vectors $\underline{y}_1, \ldots, \underline{y}_n$ are iid following the generalised multivariate logistic distribution. Let $\overline{y}_i$ and v_i denote the sample mean and

variance of y_i respectively, and $\bar{r}$ the average of the sample correlations between each of the pairs y_i and y_j. Then moment estimators may be found from

$$\begin{aligned}
\Psi^1(\tilde{\gamma}) &= \frac{\Psi^1(1)\bar{r}}{1-\bar{r}} \\
\tilde{\phi}_i^2 &= \frac{\Psi^1(\tilde{\gamma}) + \Psi^1(1)}{v_i} \\
\tilde{\theta}_i &= \bar{y}_i\tilde{\phi}_i - \Psi(\tilde{\gamma}) + \Psi(1) - \ln(\tilde{\gamma})
\end{aligned}$$

5.3.2 The Likelihood

We will consider the case in which the observation vector may contain a mixture of both uncensored and right censored components. The auxiliary multinomial estimation technique to be described in section 5.3.4 cannot easily deal with the more commonly occuring situation in which observations are right censored , except in the univariate case.

Suppose that a subvector y_A $(a \times 1)$ of y is uncensored and the remaining components , y_B say, are left censored. Then the likelihood contribution for such an observation vector is given by,

$$\frac{\partial F(y)}{\partial y_A} = \gamma(\gamma+1)\ldots(\gamma+a-1)\frac{\prod_{j\in A}\phi_j \exp(-\eta_j)}{[1+\sum_{k=1}^{r}\exp(-\eta_k)]^{\gamma+a}} \qquad (5.23)$$

where η_k is as defined in Equation 5.12. Note that this formulation assumes a different scale parameter for each response. The estimation of the scale

parameters is discussed in section 5.3.3 below.

Let

$$p_j = \frac{\exp(-\eta_j)}{[1+\sum_{k=1}^{r}\exp(-\eta_k)]} \quad j = 1,\ldots,r \tag{5.24}$$

then it is easy to show that the likelihood given by equation 5.23 can be written as,

$$\frac{\partial F(\underline{y})}{\partial \underline{y}_A} = \frac{\Gamma(\gamma+a)}{\Gamma(\gamma)} \prod_{j=1}^{r} p_j^{c_j} \left(1 - \sum_{k=1}^{r} p_k\right)^{c_{r+1}} \tag{5.25}$$

where

$$c_j = \begin{cases} 1 & \text{if } j \in A \\ 0 & \text{otherwise} \end{cases} \qquad j = 1,\ldots,r$$

$$c_{r+1} = \gamma$$

Thus, the likelihood contribution from a multivariate logistic distribution can be thought of as being proportional to an observation , $\underline{c}$, from a random variable following a multinomial distribution with $r+1$ cells and parameter vector $\underline{p}$.

For each cell corresponding to an uncensored onservation we observe 1 "success", and for each cell corresponding to a left censored observation we observe 0 "successes". In the $(r+1)$th cell we observe γ successes. Note that, although γ is not usually an integer, this does not present any computational problems since, GLIM, in particular, allows non-integer valued observations for discrete error distributions.

5.3.3 Estimation of the Scale parameters

As mentioned in section 5.3.1, we normally assume that the scale parameters, $\underline{\phi}$, are homogeneous over individuals but not necessarily over responses.

For a general analysis we will assume that the r responses can be divided into G groups of size, $g_1, \ldots, g_G$ respectively. Thus, if $G = 1$ with $g_1 = r$, we assume a common scale parameter for the entire response vector, whereas, if $G = r$ with $g_i = 1$ for $i = 1, \ldots, r$, we assume a different scale parameter for each response. In between these extremes, the groupings might reflect natural or experimental conditions .

Let y_{ijk} be the kth response in the jth group for the ith individual, and $y_{A_{ij}}$, $a_{ij} \times 1$ be the set of uncensored responses in the jth group for the ith individual, where $i = 1, \ldots, n, k = 1, \ldots, g_j$ and $j = 1, \ldots, G$. Then, the log-likelihood for the scale parameters can be written as

$$\begin{aligned} l(\underline{\phi}) &= \sum_{i=1}^{n} [\ln(\gamma + a_i) - \ln(\gamma)] \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{G} \left[a_{ij} \ln(\phi_j) - \sum_{k \in A_{ij}} \eta_{ijk} \right] \\ &\quad - \sum_{i=1}^{n} (\gamma + a_i) \ln \left(1 + \sum_{j=1}^{G} \sum_{k=1}^{g_j} \exp(-\eta_{ijk}) \right) \end{aligned} \tag{5.26}$$

where $a_i = \sum_{j=1}^{G} a_{ij}$.The maximum likelihood estimate of ϕ_l, $l = 1, \ldots, G$

is found as the root of the equation

$$\frac{\partial l(\phi)}{\partial \phi_l} = \sum_{i=1}^{n}\left[\frac{a_{il}}{\phi_l} - \sum_{k\in A_{il}} y_{ilk}\right] + \sum_{i=1}^{n}\frac{(\gamma+a_i)\sum_{k=1}^{g_l} y_{ilk}\exp(-\eta_{ilk})}{1+\sum_{j=1}^{l}\sum_{k=1}^{g_j}\exp(-\eta_{ijk})} \tag{5.27}$$

$$= \sum_{i=1}^{n}\left[\frac{a_{il}}{\phi_l} - \sum_{k\in A_{il}} y_{ilk}\right] + \sum_{i=1}^{n}(\gamma+a_i)\sum_{k=1}^{g_l} y_{ilk}p_{ilk} \tag{5.28}$$

where p_{ilk} is as defined in section 5.3.2. A simple re-arrangement gives the estimate for ϕ_l as,

$$\hat{\phi}_l = \frac{\sum_{i=1}^{n} a_{il}}{\sum_{i=1}^{n}\left[\sum_{k\in A_{il}} y_{ilk} - (\gamma+a_i)\sum_{k=1}^{g_l} y_{ilk}p_{ilk}\right]} \tag{5.29}$$

Obviously, the right hand side of this equation involves unknown parameter values through the p_{ilk}. Because of this, the estimation process will, in general, be a two-stage one; that is the location parameters θ_{ijk} will be estimated for fixed values of the scale parameters, and updated estimates of the scale parameters then calculated from these estimates. Experience has shown that this algorithm has good convergence properties, although, since the two sets of parameters are estiamted separately, some adjustment needs to be made to calculate the asymptotic standard errors of the estimates from the information matrix.

5.3.4 The Estimation Technique

The Univariate Case

At this stage, in order to motivate the technique for the analysis of the multivariate distribution, it will be useful to briefly outline the technique employed by Bennett and Whitehead, [5], to estimate the parameters of the univariate logistic distribution. The distribution function of the logistic distribution can be written as

$$F(y;\theta,\phi) = [1 + \exp(-(\phi y + \theta))]^{-1} \tag{5.30}$$

and the density function satisfies

$$f(y;\theta,\phi) = \phi[1 - F(y;\theta,\phi)]F(y;\theta,\phi)$$

Assuming that we have n observations, of which U are uncensored, R are right censored and L left censored, the likelihood for $\underline{\theta}$ and ϕ based on the data is,

$$L(\underline{\theta},\phi) = \prod_{L} F(y_i;\theta_i,\phi) \prod_{U} f(y_i;\theta_i,\phi_i) \prod_{R} [1 - F(y_i;\theta_i,\phi)]$$

writing $p_i = F(y_i;\theta_i,\phi)$, the likelihood becomes,

$$L(\underline{\theta},\phi) = \phi^{u} \prod_{L} p_i \prod_{U} p_i(1 - p_i) \prod_{R} (1 - p_i) \tag{5.31}$$

The analogy with the Binomial distribution is clear. For each uncensored observation there are two Bernouili trials, parameter p_i, yielding one success and one failure. For each left censored observation there is one trial yielding one success and for each right censored observation, one trial yielding one failure. Moreover, the relationship between the p_i and the explanatory variables can be expressed as

$$\ln\left(\frac{p_i}{(1-p_i)}\right) = \phi y_i + \theta_i \ , \ i=1\ldots,n)$$

where $\theta_i = x_i^T \underline{\beta}$. Thus, the model can be fitted using Binomial errors with logit link and treating ϕy_i as a fixed offset.The scale parameter, ϕ, is generally unknown and a two-stage algorithm is used to find the maximum likelihood estimates of both ϕ and $\underline{\beta}$,although Roger and Peacock, [60] discuss a method of estimating both sets of parametrs simultaneously.

Pettitt, [53] discussesd a generalisation of the univariate logistic distribution,analagous to the generalisation of the multivariate distribution introduced in section 5.3.1, which contains the logistic as a special case.

The Multivariate Case

In section 5.3.2 it was shown that the likelihood for the generalised miultivariate logistic distribution can be thought of as being proportional to

the likelihood for a multinomial distribution with appropriate probability vector. This is obviously a multivaraite version the of logitic/Binomial relationship discussed above.

Suppose we have independent observations $\underline{y}_i$, $i = 1\ldots,n$ where $\underline{y}_i = (y_{i1},\ldots,y_{ir})$ from the generalised multivariate distribution. Then the likelihood for the location parameters can be written as (omitting the constant term),

$$L(\underline{\theta}) = \prod_{k=1}^{n}\prod_{j=1}^{r} p_{ij}^{c_{ij}} \left(1 - \sum_{k=1}^{r} p_{ik}\right)^{c_{i,r+1}} \tag{5.32}$$

where

$$c_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ uncensored} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1,\ldots,r$$

$$c_{i,r+1} = \qquad\qquad \gamma \qquad\qquad i = 1,\ldots,n$$

and p_{ij} is as described in equation 5.24. Apart from the constant term, this is exaclty the likelihood we would have obtained if we had observed $\xi_1,\ldots,\xi_n$ from a multinomial distribution with parameters $p_1,\ldots,p_n$. Now, it is well known that multinomial data , and , in particular, contingency tables, can be modelled using the Poisson distribution by constraining certain marginal totals. In this case we can regard the data as forming an $n \times (r+1)$ contingency table with entries consisting of 0's or 1's according to whether the observations are left censored or right censored respectively, and the last

column having the value γ in each cell.

As in the univariate case, we will assume that the location parameters are linearly related to a set of explanatory variables and refer to this relationship as the MODEL. It is the parameters in the MODEL that we wish to estimate.

In order to estimate these parameters we need to constrain certain marginal totals. The terms relating to these constraints may be called OBS and RESP. OBS is an n-level factor, each level representing one observation, and RESP is and (r+1) level factor representing the dimensionality of the distribution.

The appropriate model formula which must be fitted to the contingency table is OBS+RESP*MODEL. The estimates of interest appear as interaction terms in RESP.MODEL. For a fuller discussion of this technique, which corresponds to a multivariate logit model, see Goldstein, [32] or McCullagh and Nelder, [44], pages 142-43.

In order to fit any particular explanatory variable, we need remember that the auxiliary contingency variable has one extra column, and that its value must be specified in this column. There are two cases to consider,

- If the explanatory variable is a factor its value is replicated in the (r+1)th column.
- If the explanatory variable is continuous it is assigned the value 0 in

the (r+1)th column.

5.3.5 Computational Aspects

As in the univariate case, the terms $\phi_j y_{ij} + \ln(\gamma)$ appear as offsets in the linear predictor, the values of the ϕ_j being updated after each fit until convergence. Starting values for $\underline{\phi}$ may be found either using the moment estimators given by equation 5.23, or ,as implemented in the GLIM macros, by equating the logistic variance to the sample variance of the data assuming independence. In most cases these simple estimates seem to be sufficient as starting values.

An obvious disadvantage of the technique is that, if the number of observations is large, the fitting of the OBS factor, which has n levels, becomes computationally expensive. One way of overcoming this is to note that the fitting of this factor simply ensures that the sum of the fitted values in each row of the table is equal to the sum of the observed values, in this case $\gamma + \sum_{j=1}^{r} c_{ij}$, for $i = 1, \ldots, n$. Using this relationship, we can find explicit equations for the maximum likelihood estimates in terms of the other parameters and offset and , thus, incorporate these values as part of the linear predictor during each cycle of the fit. This procedure may be thought of as a form of iterative proportional fitting routine for a subset of the parameters

in the model. The effect will be to underestimate the asymptotic standard errors of the estimates of interest, although these will need to be adjusted anyway to take account of the estimation of the scale parameters.

Example 5.6 Blood Histamine Levels

As a simple example of fitting this model, we will consider the results of an investigation into the blood histamine levels at different times for dogs in various treatment groups. The data in table 5.4 is taken from Cole and Grizzle, [15].

For this data we can define two MODEL factors, namely morphine/trimethaphan and intact/depleted. As described in Section 5.3.4, their values need to be replicated in the, in this case, 5th column of the auxiliary contingency table. The estimated parameter values and $-2\times$ log-likelihoods for various models fitted are summarised in table 5.5 . In each case the first row for each model entry represents the assumption of a common scale parameter and the second a different scale parameter for each response.

It is clear from the differences in log-likelihoods, and the parameter estimates themselves, that the assumption of a common scale parameter cannot be sustained. In particular, the estimated scale parameter for control, ϕ_1 is much larger than the others reflecting the lower variability for that mea-

		Blood Histamine (μg/ml)			
	Dog	Control	1 min	3 min	5 min
group I	1	0.04	0.20	0.10	0.08
(morphine intact)	2	0.02	0.06	0.02	0.02
	3	0.07	1.40	0.48	0.24
	4	0.17	0.57	0.35	0.24
group II	5	0.10	0.09	0.13	0.14
(morphine depleted)	6	0.12	0.11	0.10	0.11
	7	0.07	0.07	0.07	0.07
	8	0.05	0.07	0.06	0.07
group III	9	0.03	0.62	0.31	0.22
(trimethaphan intact)	10	0.03	1.05	0.73	0.60
	11	0.07	0.83	1.07	0.80
	12	0.09	3.13	2.06	1.23
group IV	13	0.10	0.09	0.09	0.08
(trimethaphan depleted)	14	0.08	0.09	0.09	0.10
	15	0.13	0.10	0.12	0.12
	16	0.06	0.05	0.05	0.05

Table 5.4: Blood Histamine levels

Model	Parameter			
	Gamma	Correlation	$-2l$	Scale Parameter
Mean	1.0	0.50	81.90	5.16
	0.95	0.93	26.22	(36.7,2.9,4.5,7.0)
+Intact/	0.75	0.61	52.79	6.75
Depleted	0.65	0.75	-6.33	(52.2,4.3,5.7,8.4)
+Morphine/	0.8	0.58	43.97	7.17
Trimethaphan	0.55	0.77	-14.56	(54.8,4.5,6.2,9.0)
+Interaction	0.8	0.58	24.41	8.62
	0.45	0.78	-38.63	(66.7,5.1,8.4,12.4)

Table 5.5: Fitted parameters for Generalised Logistic distribution, Histamine data

surment. It is also clear that there are significant treatment effects. From Table 5.5, we cane determine the deviance differences for the factors included in the model. For example, under the assumption of different scale parameters, the log-likelihood difference for the intact/depleted status compared to a model including only a mean is $26.22 + 6.33 = 32.55$ on 4 df. Similarly, including drug type and the interaction term produces successive log-likelihood differences of 8.23 and 24.07 respectively, each on 4 degrees of freedom. All these models can be fitted easily by simply changing the MODEL factor as described in above.

This final example serves to illustrate the theme of this thesis. A complicated model, i.e. the generalised multivariate logistic distribution , can be fitted to data using a statistical package and the fit of various models assessed interactively. Without expressing the distribution in terms of an auxiliary Poisson model, the fitting procedure would inevitably nvolve a fairly lengthy computer program and be difficult to implement in any interactive sense.

Chapter 6

A New Computing Environment ?

This thesis has presented techniques for fitting complicated models to data by showing that they can be represented in terms of a generalised linear model. This means that such models can then be fitted a package with the facilites for fitting generalised linear models, together with some facilites for data manipulation. Much use has been made of the GLIM package as a vehicle for implementing and fitting such models, although we note that other commonly used packages such as SAS or GENSTAT could equally well be used.

One facet of this approach whhich has not been highlighted is the inadequacy of the GLIM package for writing programs. Although, in comparison to many other statistical packages, it is extremely powerful and flexible, in comparison to other programming languages, it is archaic often intractable. A glance at some of the macros in the appendices will illustrate this point. As an example, consider the GLIM coding needed to accumulate the sum of the first ten integers, which would look something like,

```
$mac iter
$ca %i=%i+1 : %s=%i : %e=%e$
$end
$ca %i=%s=0:%e=10$whi %e iter$
```

The principal difficulty is that, in general, looping the involves the repeated invocation of macros. Because of this, and coupled with the lack of local variables, operations which involve nested subloops quickly become unintelligible and extremely messy to program. It is highly desirable that any future statistical programming language has much more user friendly and precise programming constructs.

In an attempt to address some of these issues, a prototype statistical modelling language, FUNIGIRLS, was developed based upon the functional

programming language ML, [75]. The host language ML was chosen primarily because of familiarity with its syntax and construct and is not intended that it should be used in any realistic implementation of the modelling language. However, the functional nature of ML does have a relevance as we believe that the functional approach provides a useful vehicle for the representation of the procdeures of fitting models to data. Indeed, we can think of the modelling process itself as a function mapping the data onto a set of results or statistics, of the form,

```
model fn : data -> results
```

,where results consists of whatever statistics or estimates are calculated from the model fitted to the data. This functional approach to model fitting has recently been considered by Chambers et. al, [14], who have implemented a system similar to FUNIGIRLS in an S type environment.

In GLIM terms, results can be thought of as consisting of the display of the deviance and degrees of freedom following a fit, together with other values, such as parameter estimates, which may be extracted. However, the model fitting procedures in most packages are not strictly functional in this sense because they involve the declaration of certain states, such as the error distribution or link function, pertaining to the current model.

This implied state has both advantages and disadvantages. An obvious advantage is that it is often desired to fit a sequence of similar models to the same set of observations, with only the design matrix changing between models. An ability to fix the state between the model fits certainly reduces the amount of work involved. However, a fixed state can also easily lead to mistakes being made since the user is often unaware of the current state in force when a model is fitted. Thus, in GLIM, a common mistake is to specify the link function before the error distribution which can lead to the default link being used by mistake.

The FUNIGIRLS implementation is entirely functional in the sense that the model fitting procedure consists entirely of a series of function calls, with computed results consisting almost entirely of values of function calls. The only exceptions to this rule are those functions which result in the display of quantities, such as the deviance, on the screen. Strictly speaking, such an operation is not the result of a function. Moreover, each function evaluation has a single data structure as its value, and normally an expression will have no-other side effects on the system. This means, for example, that the results from several model fits can co-exist side by side. Examples of such procedures are given below.

FUNIGIRLS emulates the operation of GLIM in that it provides facilities

for fitting generalised linear models using a similar syntax. There are restrictions on the syntax used since ML itself reserves some common operators and these cannot be overloaded for use in the FUNIGIRLS implementation. For example, the operator %+ is used in FUNIGIRLS for matrix addition and is defined as

```
fn  %+ : real matrix* real matrix -> real matrix
```

It is not possible to use the operator, + , because this is defined in ML as

```
fn + : real*real -> real
```

and because of ML's strong typing, would cause a syntax error if an attempt were made to use it for matrix addition. This convention has been circumvented to some extent by Harman and Danicic, [35],who have written a pre-processor which implements FUNIGIRLS function calls in a Pascal type language.

The operation of FUNIGIRLS is best illustrated by an example. The ML syntax is, hopefully, fairly self-explanatory, but readers are referred to [75] for further details.

Example 6.1 Linear Regression

This example illustrates how to read in some data perform a simple weighted

least squares fit. The example is a weighted linear regression of blood pressure on age for a set of 5 groups of women.The following code constructs the four vectors needed for the fit,

```
val mean = Init(5,1,["Intercept"],1.0);
val bp = Read(5,1,["Pressure"],[114.0,124.0,143.0,158.0,166.0]);
val age = Read(5,1,["AGE"],[35.0,45.0,55.0,65.0,75.0]);
val gr = Read(5,1,["WT"],[15.0,16.0,12.0,9.0,6.0]);
```

It should be clear that Read is function defined as,

```
fn Read: int*int*string list*list -> matrix
```

where the two integers are the dimensions and the list argument represents the actual elements of the matrix. In this respect, the type of the matrix is determined entirely by the type of the list. In the example above, the use of Read results in a matrix of reals. However, there is no reason why matrices of abstract data types cannot be defined. Thus we might consider matrices of "person event histories" in observational analyses.

In this example the response variable is blood presure, bp, with gr being a weight matrix and age an explanatory variable. In order to fit the model, the syntax is,

```
val res1 = W_Normal gr L_I bp mean;
```

```
val res2 = W_Normal gr L_I bp (mean^+age);
```

which results in two datatypes, res1 and res2, containing statistics calculated from the fits. Note again the use of the function $\hat{+}$ in the model formual because of the overloading of the + operator. The results from the fits can be displayed using the following

```
Display(res1,"de");Display(res2,"de");
```

which would display the deviances and estimates of the two fits.

This example simply illustrates the way that FUNIGIRLS can fit a given model. In order to illustrate the functionality of the implementation, consider the definition of the function W_Normal. This is defined as,

```
fn W_Normal: matrix*link*matrix*matrix -> results
```

Here, link is an abstract datatype consisting of two functions defining the relationships, $\mu = h(\eta)$ and $\partial\eta/\partial\mu$, while the three matrix arguments represent the weight, dependent variable and design matrix respectively. The abstract datatype results contains statistics calculated from the model. We can simplify the function by noting that both fits involve the same weight matrix and dependent variable. Thus, if we define

```
val myfit = W_Normal gr L_I bp;
```

the function myfit needs only the design matrix as a single argument and is thus defined as

```
fn myfit: matrix -> results
```

This procedure emulates the setting of the link function and declaration of the dependent variable in GLIM. In fact all the fitting functions available are built up from a low-level function **model_fit** , which is supplied with arguments defining the variance and deviance functions to define new functions for fitting particular distributions. At the very lowest level, the user has access to the Gram-Schmidt orthogonalisation routine used in the decomposition of the design matrix.

In this respect, users could effectively customise the package to perform whatever analyses they required. The existing basic functions can be combined in whatever way is required in order to create a new function to perform the required analyses. As an example, consider the construction of a simple function to perform simple linear regression. This might take the form,

```
fun Lin_Reg(X,Y)=
let val Beta=Inv(T(X) %* X) %* T(X) %* Y
in
```

```
        let val Mu=X %* Beta
        in
           let val Res=Y %- Mu
           in
              (Beta,Mu,Res,T(Res) %* Res)
           end
        end
    end
```

Application of this function to a design matrix and vector of observations results in a 4-tuple containing the parameter estimates, fitted values, residuals and residual sum of squares respectively.

Example 6.2 Using lists

We can illustrate the power of this functional approach further by utilising the built in map function of ML in order to apply a function to all elements of a list, for example a list of model formulae. The function may be built up as follows,

```
val List_Fit alist =
let fun Display_de result = Display(result,"de")
in
```

```
  map Display_de(map (myfit) alist)
end;
```

Thus, the application of this function to

```
List_Fit [mean,mean^+age];
```

will result in the display of the two fits on the screen. The easy implementation of list processing in ML means that it is straightforward to define functions to perform selection procedures such as stepwise regression from a list of candidate regressor variables. It would also seem an ideal vehicle for the implementation of graphical modelling procedures for contingency tables, although we not yet investigated this facility.

Other facilities which have currently been implemented in FUNIGIRLS include,

- matrix calculations and operations
- all current GLIM error/link combinations
- linear composite link functions
- quasi-likelihood models

These facilities were implemented in an extremely short period of time by a single programmer and the code required amounted to a little over 2000

lines, a fraction of that required by GLIM. The important point is that, since the code is embedded in ML as a set of functions, all the model fitting facilities are available, togethr with a high level programming language. Although the recursive nature of programming in ML is a little unusual to begin with, it is relatively straightforward to learn and, in keeping with the general philosophy of functional languages, encourages the writing of formally correct code.

In conclusion, the future of statistical computing and of packages and languages such as GLIM, is difficult to predict. Certainly, the rapid advances and reduction of costs of hardware will, in future, mean that statisticians will have at their disposal a formidable array of computing power. However, in our view, there still remain important developments to be made in facilities for utilising this power in an efficient manner.

It seems likely that, in the near future, the basic tools of the applied statistician will consist of a workstation with enhanced graphics capabilities. A windowing environment will be coupled with a pointing device to interact graphically with models that have been fitted and displayed on the screen. However, in terms of the statistician being able to describe precisely and accurately the model to be fitted to the data, there remains much work to be done. The functional approach described in this Chapter is an interesting

starting point.

Bibliography

[1] M. Aitkin. A simultaneous test procedure for contingency table models. *Appl. Statist.*, 28:233–42, 1979.

[2] M. Aitkin. Mixture applications of the EM algorithm in GLIM. In *COMPSTAT80*. Physica-Verlag, 1980.

[3] M. Aitkin. Modelling variance heterogeneity in normal regression in GLIM. *Appl. Statist.*, 36:332–339, 1988.

[4] F.J. Anscombe. Normal likelihood functions. *Ann. Inst. Statist. Math (Tokyo)*, 16:1–19, 1964.

[5] S. Bennett and J. Whitehead. Fitting logisitc and log-logistic regression models to censored data using GLIM. *GLIM Newsletter*, 4:12–19, 1981.

[6] C.G. Bhattacharya. A simple method of resolution of a distribution into Gaussian components. *Biometrics*, 23:115–135, 1967.

[7] G.E.P. Box and G.M. Jenkins. *Time Series Analysis, Forecasting and Control.* Holden Day, San Fransisco, 1970.

[8] R. Burns. Fitting a logit model to data with classification errors. *GLIM Newsletter*, 8:44–47, 1984.

[9] J. Burridge. A note on maximum likelihood estimation for regression using grouped data. *J.R.Statist.Soc. B*, 43:41–45, 1981.

[10] J. Burridge. Some unimodality properties of likelihoods derived from grouped data. *Biometrika*, 69:145–151, 1982.

[11] S.G. Candy. Using factors in composite link function models. *GLIM Newsletter*, 11, 1985.

[12] S.G. Candy. Fitting a parametric log-linear hazard function to grouped survival data. *GLIM Newsletter*, 13:28–31, 1986.

[13] R.J. Carroll and D. Ruppert. Power transformations when fitting theoretical models to data. *J.Amer.Statist.Ass.*, 79:321–328, 1984.

[14] J. Chambers. A computing environment for statisticians. *Bell Laboratories Research Report*, 40, 1987.

[15] J.W.L. Cole and J.E. Grizzle. Application of multivariate analysis of

variance to repeated measurement experiments. *Biometrics*, 22:810–828, 1966.

[16] R.D. Cook and M.E. Johnson. A family of distributions for modelling non-elliptically symmetric multivariate data. *J.R.Statist.Soc. B*, 43:210–218, 1981.

[17] D.R. Cox. Statistical analysis of time series, some recent developments. *Scand J. Statist*, 8:93–115, 1981.

[18] D.R. Cox and E.J. Snell. A general definition of residuals. *J.R.Statist.Soc. B*, 30:248–75, 1968.

[19] M. Crowder. A distributional model for repeated failure time measurements. *J.R.Statist.Soc. B*, 47:447–452, 1985.

[20] M. Davidian and R.J. Carroll. A note on extended quasi-likelihood. *J.R.Statist.Soc. B*, 50:74–82, 1988.

[21] A.P. Dempster. An overview of multivariate data analysis. *Jour. Mvt. Anal.*, 1:316–346, 1971.

[22] A. Ekholm and J.. Palmgren. A model for binary responses with misclassifications. In R.Gilchrist, editor, *GLIM82*, New York, 1982. Springer-Verlag.

[23] D. Firth. On the efficiency of quasi-likelihood estimation. *Biometrika*, 74:233–245, 1988.

[24] A. Forcina. Correlated observations with Normal errors. *GLIM Newsletter*, 12:31–32, 1986.

[25] E.B. Fowlkes. Some methods for studying the mixture of two normal (log-normal) distributions. *J.Amer.Statist.Ass.*, 74:561–575, 1979.

[26] E.A. Gehan. A generalised wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52:203–223, 1965.

[27] R. Gilchrist. Calculation of residuals for all GLIM models. *GLIM Newsletter*, 4:26–28, 1981.

[28] R. Gilchrist. Adjusting residuals in composite link models. *GLIM Newsletter*, 9:41, 1985.

[29] R. Gilchrist, M. Green, and Scallan A.J. Testing the mean- variance relationship and the link function in generalised linear models. In *COMPSTAT82 (supplement)*, Wien, 1982. Physica-Verlag.

[30] R. Gilchrist and A.J. Scallan. Parametric link functions in generalized linear models. In *COMPSTAT84*, pages 203–208, Wien, 1984. Physica-Verlag.

[31] C.A. Glaseby. Correlated residuals in non-linear regression applied to growth data. *Appl. Statist.*, 28:251–259, 1979.

[32] H. Goldstein. Specifying a multivariate logit model using GLIM. *GLIM Newsletter*, 4:23–26, 1982.

[33] P. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J.R.Statist.Soc. B*, 46:149–162, 1984.

[34] A. Hald. Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Skand. Aktuarietidskr.*, 32:119–34, 1949.

[35] M. Harman and S. Danicic. Functional reasoning for procedural programs. *to appear*, 1990.

[36] V. Hasselblad. Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8:431–44, 1966.

[37] J.R. Hill and C.L. Tsai. Calculating the efficiency of maximum quasi-likelihood estimation. *Appl. Statist.*, 37:219–230, 1988.

[38] J. Hinde. Compound poisson regression models. In R. Gilchrist, editor, *GLIM82*, pages 109–121, New York, 1982. Springer.

[39] N.L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York, 1972.

[40] B. Jorgenson. Maximum likelihood estimation and large-sample inference for generalised linear and non-linear regression models. *Biometrika*, 70:19–28, 1983.

[41] G.K. Kanji. A mixture model for wind shear data. *Journal Appl. Statist.*, 12:49–58, 1985.

[42] E.D. Landwehr, D. Pregibon, and A. Shoemaker. Graphical methods for assessing logistic regression diagnostics. *J.Amer.Statist.Ass.*, 79:61–83, 1984.

[43] P. McCullagh. Quasi-likelihood functions. *Annals of Statistics*, 11:59–67, 1983.

[44] P. McCullagh and J.A. Nelder. *Generalised linear models.* Chapman and Hall, London, 1983.

[45] C.J. McLachlan and P.N. Jones. Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 44:571–578, 1988.

[46] E. McLaren, G.M. Brittenham, and V. Hasselblad. Analysis of the volume of red blood cells: Application of the EM algorithm to grouped data from the doubly-truncated lognormal distrbution. *Biometrics*, 42:143–158, 1986.

[47] J.A. Nelder. The fitting of a generalisation of the logistic curve. *Biometrics*, 17:89–100, 1961.

[48] J.A. Nelder. Discussion to Jorgenson. *J.R.Statist.Soc. B*, 49:149–50, 1987.

[49] J.A. Nelder and R.W.M. Wedderburn. Generalised linear models. *J.R.Statist.Soc. A*, 135:370–384, 1972.

[50] O'Brien. Correlated observations with Normal errors: a rejoinder. *GLIM Newsletter*, 15:35–37, 1987.

[51] C. (ed) Payne. *The GLIM System Release 3.77.* Numerical Algorithms Group, Oxford, 1985.

[52] R.J. Pethybridge, J.R. Ashford, and J.G. Fryer. Some features of the distribution of birthweight of human infants. *Br. J. Prev. Soc. Med.*, 28:10–18, 1974.

[53] A.N. Pettitt. Fitting generalised logistic and log-logistic regression models to censored data using GLIM. *GLIM Newsletter*, 7, 1983.

[54] A.N. Pettitt and I. Bin Daud. Case weighted measures of influence for proportional hazards regression. *J.R.Statist.Soc. A*, 38:51–67, 1989.

[55] D.A. Pierce and D.W. Schafer. Residuals in generalised linear models. *J.Amer.Statist.Ass.*, 81:977–986, 1986.

[56] D. Pregibon. Goodness of link tests for generalised linear models. *Appl. Statist.*, 29:15–24, 1980.

[57] D. Pregibon. Logistic regression diagnostics. *Ann. Statist.*, 9:705–724, 1981.

[58] F.S.G. Richards. A method of maximum likelihood estimation. *J.R.Statist.Soc. B*, 23:469–475, 1961.

[59] J. Roger. Using factors when fitting the scale parameter to weibull, extreme value, logistic and log-logistic regression models with censored data. *GLIM Newsletter*, 11:30–37, 1985.

[60] J.H. Roger and S.E. Peacock. Fitting the scale as a *glim* parameter for Weibull, extreme-value, logistic and log-logistic regression models with censored data. *GLIM Newsletter*, 6:30–37, 1983.

[61] A.J. Scallan. Some aspects of parametric link functions. In R. Gilchrist, editor, *GLIM82*, New York, 1982. Springer-Verlag.

[62] A.J. Scallan. Fitting autoregressive processes in GLIM. *GLIM Newsletter*, 9:17–22, 1985.

[63] A.J. Scallan. A GLIM model for repeated measurements. *GLIM Newsletter*, 15:10–22, 1987.

[64] A.J. Scallan. Applications of quadratic forms in generalised linear models. *Submitted for publication*, 1990.

[65] A.J. Scallan, R. Gilchrist, and M. Green. Fitting parametric link functions in generalised linear models. *Comput. Statist. and Data Anal.*, 1:37–49, 1984.

[66] L.R. Shenton and K. Bowman. Higher moments of a maximum likelihood estimate. *J.R.Statist.Soc. B*, 25:305–317, 1963.

[67] G.K. Smyth. Generalized linear models with varying dispersion. *J.R.Statist.Soc. B*, 51:47–60, 1989.

[68] W.D. Stirling. Iteratively reweighted least squares for models with a linear part. *Appl. Statist.*, 33:7–17, 1984.

[69] B.E. Storer and J. Crowley. A diagnostic for Cox regression and general conditional likelihoods. *J.Amer.Statist.Ass.*, 80:139–147, 1985.

[70] T. Stukel. Implementation of an algorithm for fitting a class of generalised logistic models. In R. Gilchrist, B. Francis, and J. Whittaker, editors, *GLIM85*, New York, 1985. Springer-Verlag.

[71] R. Thompson and R.J. Baker. Composite link functions in generalised linear models. *Appl. Statist.*, 30:125–31, 1981.

[72] C. Vanderhoeft. Macros for calculating the covariance matrix of functions of parameter estimates. *GLIM Newsletter*, 11:21–24, 1985.

[73] R.W.M. Wedderburn. Quasi-likelihood functions, generalised linear models and the Gauss-Newton method. *Biometrika*, 61:81–89, 1974.

[74] J. Whittaker. GLIM syntax and simultaneous tests for graphical log linear models. In R. Gilchrist, editor, *GLIM82*, New York, 1982. Springer-Verlag.

[75] A. Wikstrom. *Functional Programming using Standard ML.* Prentice Hall, New Jersey, 1987.

[76] A.J. Wilcox and I.T. Russell. Birthweight and perinatal mortality : I on the frequqency distribution of birthweight. *Int. J. Epidemiol.*,

12:314–318, 1983.

[77] D. Williams. Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.*, 36:181–191, 1987.

[78] S.L. Zeger. A regression model for time series of counts. *Biometrika*, 75:621–629, 1988.

Appendix A

Macros for Grouped Data

A.1 Truncated Distributions

```
! Macros to fit truncated distribution to grouped data. Assume
! that upper limit is given by last element of UB.
! Arguments : FREQ - frequencies
!               UB - upper bounds of each interval
!             SVAL - starting values for mu and s.d.
!               %l - lower truncation point (%l < min(UB))
!
!  Macros also use variables %n, NVAR, MEAN, SD. All other  variables
!  end with an _ .
!
$mac fitm!
$del tfrq mean tub sd nvar i_ lim_$!
$ca %z1=%cu(sval==sval):%z1=(%z1/=2)$swi %z1 err1$!
$ca %z1=(%o/=1)&(%o/=2)$swi %z1 err2$!
$pri ' Enter range of cells to be included and lower bound  :-'
$data 3 lim_$din 1$!
$ca %z1=lim_(2)-lim_(1)+1:%l=lim_(3)$var %z1 tfrq tub$uni %z1$!
$ca i_=%gl(%z1,1)+lim_(1)-1:tfrq=freq(i_):tub=ub(i_)$us iter$!
$ca %z1=-%pe(1)/%pe(2):%z2=1/%pe(2)!
:sval(1)=%z1:sval(2)=%z2$!
$ca %z1=%pe(1):%z2=%pe(2):$ext %vc$ca %vc=%vc*%df/(%df-1)!
$ca %z3=(%vc(1)*%z2**2-2*%vc(2)*%z1*%z2+%vc(3)*%z1**2)/(%z2**4)!
:%z4=%vc(3)/(%z2**4):%z3=%sqrt(%z3):%z4=%sqrt(%z4)!
```

```
:%z1=-%pe(1)/%pe(2):%z2=1/%pe(2)$!
$pri : ' Mean = ' %z1  '  (' %z3 ' )'!
:' S.D. = ' %z2 '  (' %z4 ' )' : $!
$use tidy$$end!

$mac fv!
$ca %s=1+(%pl/=0) $swi %s init next!
$us ngam$! calculate new values for gamma, dh and mu
$us wvar$! calculate working variates
$end!

$mac dr $ca %dr=1$end!

$ma va $ca %va=%fv$end!

$mac di $calc %di=2*(%yv*%log(%yv/%fv)-(%yv-%fv))$end!

$mac init!
$ca %z1=%nu+1$var %z1 lp_ gm_ dh_ ub_!
$ca ind_=%gl(%nu,1)$ass ub_=%1,tub$!
$ca lp_=(ub_-sval(1))/sval(2)$!
$end!

$mac next!
$ext %pe$ca lp_=ub_*%pe(2)+%pe(1):%n=%pe(3)$!
$end!

$mac ngam!
$swi %o norm logi$!
:dif_=gm_(ind_+1)-gm_(ind_):%z2=gm_(%nu+1)-gm_(1)!
:%fv=%n*dif_/%z2:nvar=%fv/%n$!
$end!

$mac norm!
$ca gm_=%np(lp_):dh_=%exp(-lp_*lp_/2)/%sqrt(2*%pi)$!
$end!

$mac logi!
$ca gm_=1/(1+%exp(-lp_)):dh_=%exp(-lp_)/(1+%exp(-lp_))**2!
```

```
$end!

$mac wvar!
$ca lp_=lp_+dh_$arg mult lp_ %lp$us mult$$ca %lp=%lp+%n*nvar!
$arg mult dh_ mean$use mult$ca dh_=dh_*ub_$arg mult dh_ sd!
$use mult$!
$end!

$mac mult!
$ca %z1=%1(1)-%1(%nu+1)!
:%2(ind_)= %n*(dif_(ind_)+%z1+%z2*(%1(ind_+1)-%1(ind_)))/(%z2*%z2)!
$end!

$mac tidy $del lp_ dh_ ub_ gm_ ind_ dif_ $$end!

$mac iter!
$own fv dr va di$yvar tfrq$ca %n=%cu(tfrq)!
: %lp=mean=sd=nvar=0$fit mean+sd+nvar-1$!
$use tidy$ $end!

$mac err1 $pri '*** starting values must be supplied in SVAL ***'$skip 2$$end!

$mac err2 $pri '*** set value of %b: 1 - Normal , 2 - Logistic ***'$skip 2$$end
```

A.2 Mixture Distributions

```
!           MACROS TO FIT TWO COMPONENT MIXTURE DISTRIBUTIONS
!                          TO GROUPED DATA
! ARGUMENTS :-
!
! FREQ - vector containing observed frequencies of (m) sets.
! DFAC - vector specifying which obs belongs to which set,
!        i.e. DFAC(j)=k if obs j is in k'th set.
! SVAL - starting values for parameter estimates.
!        (P,MU1,s1,MU2,s2) where capitals denotes vector.
! UB   - upper bounds for categories
! OPT - denotes whether parameters are fixed or vary.
```

```
! LLIM - vector holding lower bound for each set.
! %D   -  1 = Normal, 2 = Logistic   (*** Must be set ***)
! %O   -  0 = Fixed mixture parameter(s), 2 = Estimate mixture parameter(s)
!         (*** Default %O = 1 ***)
!
! Variable names used in macros which should be avoided are :-
! %m, MIXF, MU1F, SD1F, MU2F, SD2F, MU1, MU2, SD1, SD2, MIXP, NVAL
! All other variable names end with an _ .
!
$mac fitm!
$ca one_=1$t t one_ t f dfac i nst_$!
$calc %m=%cu(nst_==nst_)$fac dfac %m$!
$var %m m1_ s1_ m2_ s2_ l1_ l2_ d1_ d2_ g1_ g2_
$ca %z1=%cu(%m*(opt==1)+(opt==0)):%z2=%cu(sval==sval)!
:%z1=(%z1/=%z2)$swi %z1 err1$!
$ca %z1=%if((%d/=1)&(%d/=2),1,0)$swi %z1 err2$!
$var 5 npm_$ca npm_=%if(opt==1,%m,1)!
$ca mixf=%if(opt(5)==1,dfac,1):mu1f=%if(opt(1)==1,dfac,1)!
:sd1f=%if(opt(2)==1,dfac,1):mu2f=%if(opt(3)==1,dfac,1)!
:sd2f=%if(opt(4)==1,dfac,1)!
$ca %z1=npm_(5):%z2=npm_(1):%z8=npm_(2):%z4=npm_(3):%z5=npm_(4)!
$fac mixf %z1 mu1f %z2 sd1f %z3 mu2f %z4 sd2f %z5$!
$own fv dr va di$yvar freq$tab the freq total for dfac into nt_$!
$ca %lp=mu1=mu2=sd1=sd2=mixp=nval=0$sca 1$cycle 20 10 0.00001!
$ca %z9=%o+1$swi %z9 mod1 mod2$use res$use tidy$!
$end!

$mac mod1
$fit nval.dfac+mu1.mu1f+sd1.sd1f+mu2.mu2f+sd2.sd2f-1$!
$end

$mac mod2
$fit nval.dfac+mu1.mu1f+sd1.sd1f+mu2.mu2f+sd2.sd2f+mixp.mixf-1$!
$end!

$mac fv!
$calc %s=1+%ne(%pl,0)$swi %s init next$!
!
! Calculates gamma and the derivative!
```

```
$use lvec$swi %d norm logi$!
!
! Calculate working variates for means and s.d.'s
$calc mu1=-dh1_/sd1_(sd1f):m1_=m1_+d1_!
:sd1=dh1_*(mu1_(mu1f)-ub)/(sd1_(sd1f)**2):s1_=s1_+d1_!
:mu2=-dh2_/sd2_(sd2f) :m2_=m2_+d2_
: sd2=dh2_*(mu2_(mu2f)-ub)/(sd2_(sd2f)**2):s2_=s2_+d2_!
$arg mult p_ mu1 m1_$use mult$arg mult * sd1 s1_$use mult!
$arg mult mp_ mu2 m2_$use mult$arg mult * sd2 s2_$use mult!
!
! Calculate working variate for p and also %fv.
!
$ca p1_=gm1_:p2_=gm2_$arg mult v1_ p1_ g1_$use mult!
$arg mult * p2_ g2_$use mult$ca mixp=(p1_-p2_)!
:%fv=(p_(mixf)*p1_+mp_(mixf)*p2_):nval=%fv/nt_(dfac)$!
!
! Calculate working linear predictors
!
$calc lp1_=-dh1_*ub/sd1_(sd1f):lp2_=-dh2_*ub/sd2_(sd2f)!
$var %m i_ i1_ i2_$ca i_=%gl(%m,1)!
:i1_=%if(opt(2)==1,%gl(%m,1),1):i2_=%if(opt(4)==1,%gl(%m,1),1)!
:l1_=-d1_*llim/sd1_(i1_) :l2_=-c2_*llim/sd2_(i2_)$del i_ i1_ i2_$!
$arg mult p_ lp1_ l1_$use mult$arg mult mp_ lp2_ l2_$use mult !
$calc %lp=lp1_+lp2_+nt_(dfac)*nval:%lp=%if((%o==0),%lp,%lp+p_(mixf)*mixp)
$end!


$mac dr $calc %dr=1 $end!

$mac va $calc %va=%fv $end!

$mac di $calc %di=2*(%yv*%log(%yv/%fv)-(%yv-%fv)) $end!

$mac init!
! Assumes starting value given for each parameter to be estimated.
$del  p_ mp_ mu1_ sd1_ mu2_ sd2_ $!
$ca %z1=npm_(5)$var %z1 pt_ p_ mp_ v1_$ca %z1=npm_(1)$var %z1 mu1_!
$ca %z1=npm_(2)$var %z1 sd1_$ca %z1=npm_(3)$var %z1 mu2_!
$ca %z1=npm_(4)$var %z1 sd2_$ca v1_=1:p_=pt_=0$ass sv1_=nt_,sval$!
```

```
$ca %z8=%z9=1$arg exte svl_$use exte$del svl_$ca %z9=(%o==1)!
$ca lp1_=(ub-mu1_(mu1f))/sd1_(sd1f)!
:lp2_=(ub-mu2_(mu2f))/sd2_(sd2f):%z9=(%o==1)!
$end!

$mac next!
$ext %pe $arg exte %pe$use exte$!
$ca lp1_=(ub-mu1_(mu1f))/sd1_(sd1f) :lp2_=(ub-mu2_(mu2f))/sd2_(sd2f)!
$end!

$mac exte!
$pri 'Current Estimates' %1 : $
!extracts parameter estimates from starting values or %pe!
$var %m i_$ca i_=%gl(%m,1) : nt_=%1(i_) : %z2=%m $del i_!
$ca %z1=npm_(1)$var %z1 i_$ca i_=%gl(%z1,1)!
:mu1_=%1(%z2+i_)$del i_$ca %z2=%z2+%z1!
$ca %z1=npm_(2)$var %z1 i_$ca i_=%gl(%z1,1)!
:sd1_=%1(%z2+i_)$del i_$ca %z2=%z2+%z1!
$ca %z1=npm_(3)$var %z1 i_$ca i_=%gl(%z1,1)!
:mu2_=%1(%z2+i_)$del i_$ca %z2=%z2+%z1!
$ca %z1=npm_(4)$var %z1 i_$ca i_=%gl(%z1,1)!
:sd2_=%1(%z2+i_):%z2=%z2+%z1$del i_$arg ext1 %1$swi %z9 ext1$!
$end!

$mac ext1!
$ca %z1=npm_(5)$var %z1 i_$ca i_=%gl(%z1,1)!
:p_=%1(%z2+i_):mp_=1-p_: %z8=%if((%cu(pt_-p_)**2<=0.00001)?(%o==0),1,0)!
:pt_=p_$del i_!
$end!

$mac lvec
$var %m i_ i1_ i2_$ca i_=%gl(%m,1)!
$ca i1_=%if(opt(1)==1,%gl(%m,1),1): i2_=%if(opt(2)==1,%gl(%m,1),1)!
$ca l1_=(llim(i_)-mu1_(i1_))/sd1_(i2_):m1_(i_)=-1/sd1_(i2_)
: s1_(i_)=(mu1_(i1_)-llim(i_))/sd1_(i2_)**2
$ca i1_=%if(opt(3)==1,%gl(%m,1),1): i2_=%if(opt(4)==1,%gl(%m,1),1)!
$ca l2_=(llim(i_)-mu2_(i1_))/sd2_(i2_):m2_(i_)=-1/sd2_(i2_)
: s2_(i_)=(mu2_(i1_)-llim(i_))/sd2_(i2_)**2$del i_ i1_ i2_$!
$end
```

```
$mac norm!
$ca gm1_=%np(lp1_):gm2_=%np(lp2_):g1_=%np(l1_):g2_=%np(l2_)!
:dh1_=%exp(-lp1_*lp1_/2)/%sqrt(2*%pi)!
:dh2_=%exp(-lp2_*lp2_/2)/%sqrt(2*%pi)$!
:d1_=%exp(-l1_*l1_/2)/%sqrt(2*%pi)!
:d2_=%exp(-l2_*l2_/2)/%sqrt(2*%pi)$!
$end!

$mac logi!
$ca gm1_=1/(1+%exp(-lp1_)):gm2_=1/(1+%exp(-lp2_))!
:g1_=1/(1+%exp(-l1_)):g2_=1/(1+%exp(-l2_))!
:dh1_=%exp(-lp1_)/(1+%exp(-lp1_))**2!
:dh2_=%exp(-lp2_)/(1+%exp(-lp2_))**2!
:d1_=%exp(-l1_)/(1+%exp(-l1_))**2:d2_=%exp(-l2_)/(1+%exp(-l2_))**2!
$end!

$mac mult!
! multiplies %2 by matrix containing values %1
$ca %z1=0:%z2=%m:%z4=0$arg mul1 %1 %2 %3$whi %z2 mul1$!
$ca %2=nt_(dfac)*%2$!
$end!

$mac mul1!
$ca %z1=%z1+1:%z2=%z2-1:%z3=nst_(%z1)-1:%z5=mixf(%z4+1):%z5=%1(%z5)!
$var %z3 i_$ca i_=%z4+%z3+2-%gl(%z3,1)!
:%2(i_)=%z5*(%2(i_)-%2(i_-1)):%2(%z4+1)=%z5*(%2(%z4+1)-%3(%m-%z2))!
:%z4=%z4+%z3+1$del i_$!
$end!

$mac tidy $del ind_ mp_ v1_ lp1_ lp2_ gm1_ gm2_ dh1_ dh2_ p1_!
p2_ npm_ one_  nst nt s1_ s2_ m1_ m2_ g1_ g2_ d1_ d2_ l1_ l2_ pt_ $!
$end!

$mac res!
$swi %d res1 res2$
$var %m ind$cal ind=%gl(%m,1)!
$pri 'Factor    Level   | ' *integer ind,8!
$pri '------------------|'
```

```
$pri 'Mean, first  distn| ' mu1_!
:    'S.D.,   "      "  | ' sd1_!
:    'Mean, second distn| ' mu2_!
:    'S.D.,   "      "  | ' sd2_ !
:    'Mixing parameter  | ' p_ ::
$end!

$mac res1 $pri '   Error distribution is NORMAL ' : $end
$mac res2 $pri '   Error distribution is LOGISTIC ' : $end

$mac err1 !
$pri '***  Not enough starting values for paramaterisation ***'!
$exit 2$ $end!

$mac err2!
$pri '*** Set value of %d : 1 - Normal, 2 - Logistic ***'!
$exit 2$end!
```

Appendix B

The Multivariate Logistic distribution

B.1 GLIM Macros

```
$MAC MLOG!
$C Use macro INIT to set initial table and variables!
$ARG INIT #ARGS$OUT$SWI %0 TIDY$SWI %0 INIT$CA %Z9=0$!
$CA %Z3=%6:%Z2=1:AJ_=ALJ_-1$VAR %4 G_$CA G_=1$WHI %Z2 GVAL$!
$CA %Z7=%CU(%LOG(G_))$DEL G_$!
$C Use macros to calculate di and trigamma functions!
$VAR 2 G_$CA G_(1)=%6$ARG DGAM G_$$USE DGAM!
$CA %Z8=G_(2)+0.5772-%LOG(%6):G_(1)=%6$ARG TGAM G_$USE TGAM$!
$CA %Z3=G_(2)/(1.6449+G_(2))$DEL G_$CA AJ_=ALJ_+%6$!
$C Fit model until convergence!
$ARG MPHI #ARGS$USE MPHI$CA %Z1=%4$WHI %Z1 DYVAR$!
$ARG ITER #ARGS $CA %Z2=1$WHI %Z2 ITER $OUT 2$!
$C Print results of fit
$USE MRES
$END

$MAC MRES
$PRI : 'Power parameter          ' %G!
: 'Correlation              ' %z3!
: 'Number of iterations     ' %z9!
: e5 'Deviance                ' %z4!
```

```
: 'Scale parameter(s)        ' ph2_!
$END!

$MAC ITER!
$C Calculate new estimate for PHI!
$CA PH1_=PH2_ : PH1_=%IF((PH1_<=0),-PH1_,PH1_)!
: %Z1=%4$WHI %Z1 NOFF$FIT DESI+RESP+$MODE$USE NPHI$!
$C Check for convergence!
$CA %Z2=%IF(((PH1_-PH2_)**2<=0.00005),0,1):%Z9=%Z9+1!
$out 2$use mres$out$
$END

$MAC INIT!
$CA %Z1=%3+%4$VAR %Z1 DES_ RES_ P_ FV_ $VAR %4 IN1_$CA IN1_=%GL(%4,1)!
:DES_=%GL(%4,%3) : RES_=%GL(%3,1)!
$T T %1 T W %2 F RES_ I YIJ_ : T YIJ_ T F %5 I YIL_$DEL YIJ_$!
$T T %2 T F RES_ I AIJ_ : T AIJ_ T F %5 I AIL_$DEL AIJ_$!
$T T %2 T F DES_ I ALJ_$!
$C Calculate initial estimates for PHI!
$CA SCA1=%5(RES_)$T T %1 V F SCA1 I PH1_$CA %Z1=%CU(PH1_==PH1_)$!
$DEL SCA1$VAR %Z1 PH2_$CA PH1_=%PI/%SQRT(3+PH1_)$!
$C calculate size of table!
$CA %NU=%4+(%3+1)$UNI %NU$
! Set up index and margin variables
$VAR %3 IND_ : %NU DY_ OFF_ $CA IND_=%GL(%3,1) : %Z2=%3+1$!
$FAC DESI %4 RESP %Z2$CA DESI=%GL(%4,%3+1) : RESP=%3+2-%GL(%3+1,1)!
! Calculate dummy Y-variable
$ARG DYVAR #ARGS$CA %Z1=%4$WHI %Z1 DYVAR$!
$ARG NOFF #ARGS$CA %Z1=%4$WHI %Z1 NOFF$
! Declare model
$OWN FV DR VA DI$YVAR DY_ $OFF OFF_$!
$CA %LP=0$FIT DESI+RESP+$MODE$!
$END

$MAC DYVAR!
! Calculates dummy Y-variable
$CA DY_((%4-%Z1)*(%3+1)+IND_)=%2((%4-%Z1)*%3+IND_)!
:DY_((%4-%Z1+1)*(%3+1))=%6 : %Z1=%Z1-1$!
$END!
```

```
$MAC NOFF
! Calculates working offset
$CA OFF_((%4-%Z1)*(%3+1)+IND_)=%1((%4-%Z1)*%3+IND_)*PH1_(%5(IND_))!
:%Z1=%Z1-1$!
$END!

$MAC NPHI!
! Calculates new scale parameters and deviance
$EXT %PE$CA %Z1=%PE(1) : PE_=%Z1+%PE(IN1_) : PE_(1)=%Z1$!
$CA %Z1=%4$ARG PROBS $ARGS$WHI %Z1 PROBS$!
$T T P_ T F DES_ I SUM_$!
$CA SUM_=SUM_+1 : YP_=P_*%1 : AJS_=AJ_/SUM_ : PT=P_/SUM_(DES_)!
:YPA_=AJS_(DES_)*YP_$!
$T T YPA_ T F RES_ I YPT_ : T YPT_ T F %5 I YPL_$!
$CA PH2_=AIL_/(YIL_-YPL_)$!
$CA %Z5=%CU(AIL_*%LOG(PH1_)) : %Z6=%CU(%2*%LOG(P_))$!
:%Z4=%CU(AJ_*%LOG(SUM_)) : %Z4=-2*(%Z7+%Z5+%Z6-%Z4)$!
$DEL SUM_ YP_ YPJ_ YPL_ PT_$!
$END!

$MAC ARGS %1 %2 %3 %4 %5 %6 $END!

$MAC PROBS!
! Calculates fitted values etc.
$CA P_((%4-%Z1)*%3+IND_)=PE_(%4-%Z1+1)-%LP((%4-%Z1)*(%3+1)+IND_)!
:FV_((%4-%Z1)*%3+IND_)=(P_((%4-%Z1)*%3+IND_)+OFF_((%4-%Z1)*(%3+1)+IND_)!
                        +%Z8)/PH1_(%5(IND_))!
:P_((%4-%Z1)*%3+IND_)=%EXP(P_((%4-%Z1)*%3+IND_)) : %Z1=%Z1-1$!
$END!

$MAC GVAL!
! Calculates constant function of gamma in likelihood
$CA G_=%IF((AJ_>=0),G_*%Z3,G_) : %Z3=%Z3+1!
:AJ_=%IF((AJ_<=0),0,AJ_-1) : %Z2=%CU(AJ_)$!
$END!

$MAC TIDY $DEL DES_ IN1_ ALJ_ YIL_ AIL_ RESP DESI OFF_ DY_
AJ_ PH1_ PH2_ FV_ AJS_ $END!
```

```
$C Macros to fit POISSON model with negative log-link

$MAC FV $CA %FV=%EXP(-%LP) $END!
!
$MAC DR $CA %DR=-1/%FV $END!
!
$MAC VA $CA %VA=%FV $END!
!
$MAC DI $CA %DI=2*(%YV*%LOG(%YV/%FV)-(%YV-%FV)) $END!

$MAC DGAM
! Calculates digamma function
$CA %Z2=1-%A1$SWI %Z2 ERR1 $CA %Z2=%IF((%1(1)<=0),1,0)$SWI %Z2 ERR2$!
$DATA 6 CNS_$READ
1.0E-5 8.5 8.3333E-2 8.3333E-3 3.9683E-3 -5.7722E-1
$CA %1(2)=0.0 : %Z1=%1(1) : %Z2=%IF((%Z1<=CNS_(1)),1,0)$!
$ARG DGA1 %1$SWI %Z2 DGA1$!
$CA %Z2=%IF((%Z1>=CNS_(2)),0,1)$ARG DGA2 %1$WHI %Z2 DGA2$!
$CA %Z3=1.0/%Z1 : %1(2)=%1(2)+%LOG(%Z1)-0.5*%Z3$!
:%Z3=%Z3*%Z3 : %1(2)=%1(2)-%Z3*(CNS_(3)-%Z3*(CNS_(4)-%Z3*CNS_(5)))$!
$DEL CNS_!
$END

$MAC DGA1 $CA %1(2)=CNS_(6)-1.0/%Z1$DEL CNS_$ $EXIT 2$END!

$MAC DGA2
$CA %1(2)=%1(2)-1.0/%Z1 : %Z1=%Z1+1 : %Z2=%IF((%Z1>=CNS_(2)),0,1)!
$END

$MAC TGAM!
$CA %Z2=1-%A1$SWI %Z2 ERR1$CA %Z2=%IF((%1(1)<=0),1,0)$SWI %Z2 ERR2$!
$DATA 4 CNS_$READ
1.6667E-1 -3.3333E-2 2.3810E-2 -3.3333E-2
$CA %1(2)=0 : %Z1=%1(1) : %Z2=%IF((%Z1<=0.0001),1,0)!
$ARG TGA1 %1$SWI %Z2 TGA1$CA %Z2=%IF((%Z1>=5),0,1)$ARG TGA2 %1$WHI %Z2 TGA2$!
$CA %Z3=1/(%Z1*%Z1):%1(2)=%1(2)+0.5*%Z3  !
+(1+%Z3*(CNS_(1)+%Z3*(CNS_(2)+%Z3*(CNS_(3)+%Z3*CNS_(4)))))/%Z1!
$DEL CNS_!
```

```
$END

$MAC TGA1!
$CA %1(2)=1/(%Z1*%Z1)$DEL CNS_$EXIT 2$!
$END

$MAC TGA2!
$CA %1(2)=%1(2)+1/(%Z1*%Z1):%Z1=%Z1+1:%Z2=%IF((%Z1>=5),0,1)!
$END!

$MAC ERR1
$PRI : ' This macro requires one argument - a vector of length >=2'
: $exit 2$END!

$MAC ERR2
$PRI :' Invalid argument for this macro, CHECK %1(1) > 0' : $EXIT 2$!
$END
```

Appendix C

Macros for Logistic Curves

```
! Macros to fit generalised log-logistic curve to data
! following AR(1) process
! Arguments : Y  - vector of (log)observations
!             %A - initial value for asymptote
!             %G - value of power parameter
!             %O - switch variable 0 - %g fixed, >0 - %g estimated
!             X  - explanatory variabel (time)
! Macro invoked by $use FITM
!
$MAC MULT
$C Multiplies vector argument %1 by matrix U!
$CA %1(in2_)=%1(in2_)-%R*%1(in2_-1) : %1(1)=%SQRT(1-%R*%R)*%1(1)!
$END

$MAC FV
$ca %Z1=(%PL/=0)+1$SWI %Z1 INIT MEXT$
$C Calculate means!
$CA MU=%LOG(%A)-%g*%LOG(1+%EXP(-LP)/%g)
$C Calculate extra explanatory variable for asymptote
   and working linear predictor.
$USE DR$$CA M1=DER/%A!
:m2=der*(%exp(-lp)/(%g+%exp(-lp))-%log(1+%exp(-lp)/%g))!
: %LP=(LP+%A*M1+(%o>0)*%g*m2)/DER$ARG MULT %LP$US MULT$
$C Calculate working explanatory variables.
$CA X1=1/DER : X2=X/DER : X3=M1/DER : x4=m2/der$!
```

```
$ARG MULT X1$US MULT$ARG MULT X2$US MULT$ARG MULT X3$US MULT$!
$arg mult x4$us mult!
$C Calculate working fitted values
$ca %FV=MU$ARG MULT %FV$US MULT$
$END

$MAC DR $CA %DR=1 : DER=(%g+%EXP(-LP))/(%g+%EXP(-LP)) $END

$MAC VA $CA %VA=1 $END

$MAC DI $CA %DI=(%YV-%FV)*(%YV-%FV) $END

$MAC INIT $CA LP=-%LOG(%g*(%EXP((%LOG(%A)-MU)/%g)-1))!
 $END!

$MAC MEXT $EXT %PE $CA %A=%PE(3) : LP=%PE(1)+%PE(2)*X$!
$$ca %z1=(%o>0)$swi %z1 mex1$ $END!

$mac mex1 $ca %g=%pe(4)$ $end

$mac newr $ca %z2=%cu((y(in1_)-mu(in1_))*(y(in1_-1)-mu(in1_-1))) !
: %z3=%cu((y-mu)*(y-mu))-(y(1)-mu(1))**2-(y(%nu)-mu(%nu))**2!
:%r=%z2*(%nu-2)/(%z3*(%nu-1)) $end!

$MAC SETU $CA Y1=Y$YVAR Y1$CA MU=Y$OWN FV DR VA DI!
$ca %z1=%nu-1$var %z1 in1_ in2_$ca  in1_=%gl(%z1,1)+1 : in2_=%nu+2-in1_$
: %LP=X1=X2=X3=X4=0$END!

$MAC FITM $ca %z1=(%g<=0)$swi %z1 err1 !
$use setu$ca %z9=1:%z8=%r$whi %z9 iter$end!

$mac iter $ca %z1=1+(%o>0)$CA Y1=Y$ARG MULT Y1$US MULT$!
$swi %z1 mod1 mod2$$
$pri : : '  Value of rho = ' %r $di e$!
$use newr !
$ca %z9=%if(((%z8-%r)**2<1.0e-6),0,1):%z8=%r$END!

$mac mod1  $fit X1+X2+X3-1 $end
$mac mod2  $fit x1+x2+x3+x4-1 $end
```

```
$mac err1 $pri :  '*** %G MUST BE NON-NEGATIVE *** ': $exit 2$end!

! Macros to fit generalised log-logistic curve to data
! following AR(1) process with arbitrary variance function
! Arguments : Y  - vector of (log)observations
!             %A - initial value for asymptote
!             %B - variance function V(Y)=MU**%B
!             %G - value of power parameter
!             %O - switch variable 0 - %g fixed, >0 - %g estimated
!             X  - explanatory variabel (time)
! Macro invoked by $use FITM
!
$MAC MULT
$C Multiplies vector argument %1 by matrix U!
$CA %1(in2_)=%1(in2_)-%R*%1(in2_-1) : %1(1)=%SQRT(1-%R*%R)*%1(1)!
$END

$MAC FV
$ca %Z1=(%PL/=0)+1$SWI %Z1 INIT MEXT$
$C Calculate working fitted values
$ca elp=%exp(-lp) : MU=%A/(1+elp/%g)**%g
$C Calculate extra explanatory variable for asymptote
   and working linear predictor.
$USE DR$ca mva=mu**(%b/2)
$CA M1=DER/(1+elp)**%g
:m2=der*mu*(elp/(%g*(1+elp/%g))-%log(1+elp/%g))
: %LP=(LP+%A*M1+(%o>0)*%g*m2)/(DER*MVA)
$ARG MULT %LP$US MULT$
$C Calculate working explanatory variables.
$CA X1=1/(DER*MVA) : X2=X/(DER*MVA) : X3=M1/(DER*MVA): x4=m2/(der*mva)!
$ARG MULT X1$US MULT$ARG MULT X2$US MULT$ARG MULT X3$US MULT$
$arg mult x4$us mult$!
$CA %YV=Y/MVA$ARG MULT %YV$US MULT$: %FV=MU/MVA$ARG MULT %FV$US MULT$
$END

$MAC DR $CA %DR=1 : DER=1/(mu*%g*(1-(mu/%a)**(1/%g)))$ $END

$MAC VA $CA %VA=1 $END
```

```
$MAC DI $CA %DI=(%YV-%FV)*(%YV-%FV) $END

$MAC INIT
$CA LP=-%log(%g*((%a/mu)**(1/%g)-1))!
: %z1=%nu-1$var %z1 in1_ in2_$ca  in1_=%gl(%z1,1)+1 : in2_=%nu+2-in1_$ $END!

$MAC MEXT $EXT %PE $CA %A=%PE(3) : LP=%PE(1)+%PE(2)*X$!
$$ca %z1=(%o>0)$swi %z1 mex1$ $END!

$mac mex1 $ca %g=%pe(4)$ $end

$mac newr $ca r_=(y-mu)/mu**(%b/2)
: %z2=%cu(r_(in1_)*r_(in1_-1)) !
: %z3=%cu(r_*r_)-r_(1)-r_(%nu)!
:%r=%z2*(%nu-2)/(%z3*(%nu-1)) $end!
$MAC SETU $CA Y1=Y$YVAR Y1$CA MU=Y$OWN FV DR VA DI!
$ca %z1=%nu-1$var %z1 in1_ in2_$ca  in1_=%gl(%z1,1)+1 : in2_=%nu+2-in1_$
: %LP=X1=X2=X3=X4=0$END!

$MAC FITN $ca %z1=(%g<=0)$swi %z1 err1 !
$use setu$ca %z9=1:%z8=%r$whi %z9 iter$end!

$mac iter $ca %z1=1+(%o>0)$CA Y1=Y$ARG MULT Y1$US MULT$!
$swi %z1 mod1 mod2$$
$pri : : '  Value of rho = ' %r $di e$!
$use newr !
$ca %z9=%if(((%z8-%r)**2<1.0e-6),0,1):%z8=%r$END!
$mac mod1  $fit X1+X2+X3-1 $end
$mac mod2  $fit x1+x2+x3+x4-1 $end
$mac err1 $pri :  '*** %G MUST BE NON-NEGATIVE *** ': $exit 2$end
```

Appendix D

Wind Shear Data

This appendix contains a portion of the wind shear data analysed by Kanji, [41]. The data is classified by two factors as follows,

1. Band width. This refers to the 120 seconds of recorded flight before landing. Band widths 1 and 2 are each of 40 seconds duration, while Band widths 3 and 4 are of 20 seconds duration.

2. Gradient separation time interval, H_s. This represents the elapsed time between a change in the wind velocity gradient.

Class Range		Case Number							
Lower	Upper	1	2	3	4	5	6	7	8
-1.4157	-1.2870	0	0	1	0	0	0	1	0
-1.2870	-1.1583	0	1	0	0	0	1	0	0
-1.1583	-1.0296	0	3	1	2	0	0	0	1
-1.0296	-0.9009	4	6	10	2	0	4	3	1
-0.9009	-0.7722	12	19	15	15	1	10	6	6
-0.7722	-0.6435	24	43	27	29	26	22	13	14
-0.6435	-0.5148	86	92	75	72	41	67	51	41
-0.5148	-0.3861	162	227	153	167	118	140	113	97
-0.3861	-0.2574	433	538	296	351	316	386	222	242
-0.2574	-0.1287	979	1060	540	594	760	839	472	511
-0.1287	0.0	2237	2006	886	749	1843	1691	751	619
0.0	0.1287	2188	1998	882	755	1829	1667	727	627
0.1287	0.2574	1020	1101	589	595	802	878	468	465
0.2574	0.3861	437	538	326	324	324	383	253	246
0.3861	0.5148	186	217	156	168	118	154	95	109
0.5148	0.6435	77	96	63	60	44	59	42	47
0.6435	0.7722	26	33	27	43	18	26	12	16
0.7722	0.9009	9	19	7	13	1	5	5	8
0.9009	1.0296	3	7	1	11	1	2	1	4
1.0296	1.1583	0	0	2	3	0	0	1	1
1.1583	1.2870	1	1	1	0	0	0	0	0
1.2870	1.4157	1	0	0	0	1	0	0	0
1.4157	1.5444	0	0	1	0	0	0	0	0
Band		1	2	3	4	1	2	3	4
H_s (seconds)		0	0	0	0	2	2	2	2

Table D.1: Lobe Distributions. Wind Shear Data

TITLE Some Applications of Generalised Linear Models

AUTHOR Anthony Scallan

INSTITUTION and DATE Polytechnic of North London
July 1990
(CNAA)

DX

97502