Diagnostics for Generalized Linear Models

Sonia Benghiat

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

June 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-64046-9

Canada

Abstract

Diagnostics for Generalized Linear Models

Sonia Benghiat

The analysis of residuals can capture departures from a parametrized model. In this thesis we look at how the generalized linear model has become one of the most important developments in statistics in the last thirty years, and on the adequacy of regression model diagnostics that are meaningful and significant in a generalized linear model context. Some asymptotic properties are discussed and numerical examples are provided to illustrate the techniques for binomial, Poisson, and gamma distributed random variables.

Résumé

Des diagnostiques pour les modèles linéaires généralisés

Sonia Benghiat

L'analyse des résidus est un outil fort puissant qui nous permet de vérifier la validité d'un modèle paramètrique. Dans ce mémoire, je donne un aperçu de l'importance que les modèles linéaires généralisés ont eu sur le déroulement des statistiques dans les trentes dernières années. J'analyse la facilité que nous procurent de tels modèles lorsqu'il s'agit des diagnostiques de régressions. J'examine également les lois asymptotiques concernant ces modèles. Finalement, je présente des exemples pour des variables aléatoires binomiales, Poisson, et gamma.

Acknowledgements

This thesis could not have been possible without the patience and the boundless support from my husband. To him I owe a debt of gratitude. My parents, my brother and my sister continuously reminded me of the importance of completing my masters degree and to them I am thankful for their persistent encouragements. I hold a great respect for my supervisor Prof. Y. Chaubey. He very patiently guided the advancements of this thesis. To him I express my sincerest gratitude. I would also like to thank Prof. J. Garrido who willingly provided me with some useful material for the realization of this thesis. I thank Prof. A. Canty for kindly accepting to advise me on the choice of my software application. I thank the graduate secretaries and the professors from the Mathematics and Statistics department, and my classmates, not least, for their insightful help and for offering a pleasant learning environment altogether.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Linear Model

Most of the generalized linear model concepts stem from the theory of the normal linear model. Before introducing the generalized linear model, it is useful to set the scene by providing a brief review of the normal linear model in this first chapter, and hence to understand and see the parallels between the two types of models.

The normal-theory linear model is given by

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{1.1}$$

where $\mathbf{y}$ is an $n \times 1$ observation vector, $\mathbf{X}$ is a $n \times p$ known design matrix, β is a $p \times 1$ vector of unknown parameters, called regression parameters and ϵ is an $n \times 1$ vector of unobserved random variables with zero mean and constant variance σ^2, which are independently and normally distributed. The model (1.1) is alternatively described by the mean-vector and variance-covariance matrix of the observations $\mathbf{y}$ as

$$E(\mathbf{y}) = \mathbf{X}\beta, \qquad Var(\mathbf{y}) = \sigma^2\mathbf{I}.$$

The linearity of the model is understood in terms of the regression parameters β. For estimation of the parameters, the maximum likelihood method can be used when the errors are normal. Likewise, the principle of least squares provides the same estimates of the regression parameters. However, it does not require any distributional assumption. It is described below.

Least Squares Estimation of Parameters β

The least squares method estimates the regression parameters β by minimizing the sum of squares:

$$\mathbf{Q}(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

$$= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta. \tag{1.2}$$

Since

$$\frac{\partial \mathbf{Q}}{\partial \beta} = 0 \Leftrightarrow -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta = 0,$$

the least square estimator $\hat{\beta}$ for β is given by the so-called normal equations

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

This yields

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{1.3}$$

assuming that $\mathbf{X}$ is of full column rank. It is easily verified that $\hat{\beta}$ is unbiased for β and

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{1.4}$$

In addition to being unbiased, the least square estimator (LSE) $\hat{\beta}$, has the following properties:

(1) have minimum variance among all unbiased linear estimators (Gauss-Markov theorem),

(2) consistent, and

(3) sufficient.

Projection Matrix and Residuals

The building blocks for detecting influential observations in a given data are generated by the *projection matrix*, **M**, and *residuals*, **e** which are defined in what follows. Consider the model (1.1) with corresponding fitted values ($\hat{\mathbf{y}}$) and residual vector ($\mathbf{e}$) defined by:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}, \tag{1.5}$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}. \tag{1.6}$$

The projection matrix $\mathbf{M} = (m_{ij})$ is defined by:

$$\mathbf{M} = \mathbf{I} - \mathbf{H},$$
$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is called the "hat matrix". The projection matrix is most useful in the analysis of residuals as it spans the residual space, *i.e.*,

$$\mathbf{e} = \mathbf{M}\mathbf{y}. \tag{1.7}$$

The residuals **e** measure the difference between the observed and the fitted values, with the following properties:

- $E(\mathbf{e}) = \mathbf{0}$,

- $Var(\mathbf{e}) = \sigma^2\,(\mathbf{I} - \mathbf{H})$.

An unbiased estimator of σ^2 based on the residual **e** is given by

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-p} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}, \tag{1.8}$$

whereby (1.8) is denoted by MSE, the *mean square due to error*. Therefore,

$$\widehat{Var(\mathbf{e})} = MSE\,(\mathbf{I} - \mathbf{H}) \tag{1.9}$$

is an unbiased estimator of $Var(\mathbf{e})$.

Theorem 1.1 *The following are important properties related with the projection matrix* $\mathbf{M}$*:*

1. $\mathbf{H}$ *and* $\mathbf{M} = (\mathbf{I} - \mathbf{H})$ *are symmetric and idempotent,*

2. *rank* $\mathbf{M} = rank(I - H) = tr(M) = tr\,(\mathbf{I} - \mathbf{H}) = n - p$,

3. $\mathbf{MX} = (\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$

PROOF: (see Seber [24], Appendix A)

1. Symmetry is obvious as $\mathbf{H}' = [\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}]' = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} = \mathbf{H}$ and the idempotence is easily verified as

$$\mathbf{H}\cdot\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} = \mathbf{H},$$

and

$$(\mathbf{I}-\mathbf{H})\cdot(\mathbf{I}-\mathbf{H}) = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}\cdot\mathbf{H} = \mathbf{I} - \mathbf{H}.$$

2. Since $(\mathbf{I}-\mathbf{H})$ is idempotent, $\text{rank}(\mathbf{I}-\mathbf{H}) = \text{tr}\,(\mathbf{I}-\mathbf{H})$. Furthermore, since

$$\begin{aligned}\text{tr}\,(\mathbf{I}-\mathbf{H}) = \text{tr}\,\mathbf{I} - \text{tr}\,\mathbf{H} = n - \text{tr}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ = n - \text{tr}\,\mathbf{I}_{p\times p} \\ = n - p;\end{aligned}$$

then $\text{rank}(\mathbf{I}-\mathbf{H}) = \text{tr}\,(\mathbf{I}-\mathbf{H}) = n-p$.

3.

$$\begin{aligned}(\mathbf{I}-\mathbf{H})\mathbf{X} = [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{X}' &= \mathbf{X}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{X}-\mathbf{X} \\ &= \mathbf{0}.\end{aligned}$$

□

It can be further deduced that

$$E(\hat{\mathbf{y}}) = E(\mathbf{X}\hat{\beta}) = \mathbf{X}\beta, \tag{1.10}$$

and

$$\begin{aligned}Var(\hat{\mathbf{y}}) &= Var(\mathbf{X}\hat{\beta}) \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2\,\mathbf{H}. \end{aligned}\tag{1.11}$$

1.1.1 Validity of Assumptions

In fitting a linear regression model, the residuals $\mathbf{e}$ can be used to justify the assumptions about the random errors ϵ. Since $\mathbf{e}$ is linear in $\mathbf{y}$, $\mathbf{e}$ is a random variable following a normal distribution, and hence the assumption of normality can be used to draw inferences about the linear model. Thus, an analysis which combines the residuals and the fitted values will examine whether there are any departures from the linear model with normal errors. The model departures to be examined are categorized as :

- non-linearity,
- non-constant variance,
- non-independence,
- non-normality,
- outliers,
- omission of independent covariates.

Graphical methods (see Draper and Smith [7], Chapter 4), involving the residuals provide useful tools for detecting such model departures. They are described below:

1. Plots of residuals against independent variables will detect potential outliers, non-constant variance, non-linearity of an independent variable or the need for more independent variables,

2. Plots of residuals against the fitted values will detect non-constancy of variance,

3. Plots of residuals against time (if possible) will detect non-independence amongst errors or if the time effect has been omitted from the model,

4. Box-plots, normal probability plots, Half-normal plots, histograms and stem-and-leaf plots will check for normality and outliers, and

5. Plots of residuals against other significant independent variables (if possible) will detect whether such variables are to be included in the model.

Formal tests build statistics involving residuals which are used to test the validity of the following normal linear regression model assumptions:

- randomness;
- homoscedasticity;
- normality; and
- outliers.

F-test for Adequacy of the Regression Model

Consider the linear regression model (1.1) whereby the errors ϵ_i are assumed to be i.i.d.. The adequacy of the model is interpreted in the form of the significance of the independent variables $\{\mathbf{x}_i\}$, $i = 1, \ldots, p-1$. The following hypotheses are tested:

$$\begin{aligned} H_o &: \ \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_a &: \ \text{not all} \ \ \beta_j = 0; \ j = 1, \ldots, p-1. \end{aligned}$$

It can be shown that the likelihood ratio test for H_o *vs.* H_a if H_o is true yields the following F-statistic:

$$F = \frac{MSR}{MSE} \sim F_{p-1,n-p}. \tag{1.12}$$

where

$$MSE = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p} = \frac{\mathbf{e}'\mathbf{e}}{n-p} \tag{1.13}$$

and

$$MSR = \frac{\mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{1}'\mathbf{1})\mathbf{y}}{p-1}. \tag{1.14}$$

The critical region is given by

$$\{F : F \geq F_{\alpha,p-1,n-p}\}, \tag{1.15}$$

where, for any $\nu_1, \nu_2 \in N^-$, $F_{\alpha;\nu_1,\nu_2}$ is defined by

$$P[F_{\nu_1,\nu_2} \geq F_{\alpha;\nu_1,\nu_2}] = \alpha, \tag{1.16}$$

with the random variable F_{ν_1,ν_2} having an F-distribution with ν_1, ν_2 degrees of freedom. The critical region given in (1.15) is justified by the following facts:

(i) $(n-p)\frac{MSE}{\sigma^2} \sim \chi^2_{n-p}$,

(ii) $(p-1)\frac{MSR}{\sigma^2} \sim \chi'^2_{p-1}(\lambda)$, where $\lambda = \beta'\mathbf{X}'(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}\beta$, $\chi'^2_\nu(\lambda)$ denotes the non-central chi-square random variable with ν degrees of freedom and non-centrality parameter (ncp) λ.

(iii) MSE and MSR are independent,

(iv) $E(MSR) = \sigma^2 + \beta'\mathbf{X}'(\mathbf{H} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}\beta/(p-1) \geq \sigma^2 = E(MSE)$.

The assertions (i)-(iii) are consequences of *Cochran's Theorem* (see Searle [23], Chapter 3), essentially by using the following theorem:

Theorem 1.2 *Let* $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. *Then,*

(1) $\mathbf{z}'\mathbf{A}\mathbf{z}$ *has a* χ^2*-distribution with rank(*$\mathbf{A}$*)= degrees of freedom, iff.* $\mathbf{A}$ *is idempotent;*

(2) $\mathbf{z}'\mathbf{Az}$ *and* $\mathbf{z}'\mathbf{Bz}$ *are independent iff.* $\mathbf{AB} = \mathbf{0}$.

(3) The assertion in (1) changes to a non-central chi-square with $ncp = \mu'\mu$ *in case* $\mathbf{z} \sim N(\mu, \mathbf{I})$.

$\mathbf{z}'\mathbf{Az}$ can be written as

$$(n-p)\frac{MSE}{\sigma^2} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{\sigma^2} = \frac{\epsilon'(\mathbf{I}-\mathbf{H})\epsilon}{\sigma^2} = \mathbf{z}'\mathbf{Az},$$

where $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{A} = (\mathbf{I} - \mathbf{H})$.

Since $\mathbf{A}$ is idempotent with rank $n - p$ (Theorem 1.1), it follows that

$$(n-p)\frac{MSE}{\sigma^2} \sim \chi^2_{n-p}$$

and, similarly

$$(p-1)\frac{MSR}{\sigma^2} = \frac{\mathbf{y}'(\mathbf{H} - \frac{1}{n}\mathbf{1}'\mathbf{1})\mathbf{y}}{\sigma^2}$$

has a non-central chi-square distribution with degrees of freedom $=$ trace $(H - \frac{1}{n}11') = p - 1$ and non-centrality parameter

$$\lambda = \beta'\mathbf{X}'(\mathbf{H} - \frac{1}{n}\mathbf{11}')\mathbf{X}\beta / \sigma^2.$$

Since $\mathbf{HX} = \mathbf{X}$, the non-centrality parameter simplifies to

$$\lambda = \beta'\mathbf{X}'(\mathbf{I} - \frac{1}{n}\mathbf{11}')\mathbf{X}\beta / \sigma^2,$$

which is ≥ 0 and equal zero iff. H_o holds.

Independence easily follows since

$$(\mathbf{I}-\mathbf{H})(\mathbf{H} - \frac{1}{n}\mathbf{11}') = -(\mathbf{I}-\mathbf{H})\frac{1}{n}\mathbf{11}' = \mathbf{0}.$$

The assertion in (iv) is a strict inequality if at least one of the $\beta_j \neq 0$.

1.1.2 Other Diagnostics

Some diagnostic tools are used to detect influential and outlying observations in a given regression model. The ***Studentized residual*** is very informative in examining residuals under a normal model since it is standardized and it introduces the idea of case deletion, where the fit for all observations is compared to the fit with the deleted case. Also,

$$\mathbf{Var}(\mathbf{e}) = \sigma^2 \mathbf{M}.$$

then

$$\widehat{\mathbf{Var}(\mathbf{e})} = \hat{\sigma}^2 \mathbf{M} = \ MSE \cdot \mathbf{M}.$$

Furthermore,

$$\frac{e_i}{\sigma\sqrt{m_{ii}}} \sim \ N(0,1),$$

hence, the studentized residual e_i^* is defined as

$$e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{m_{ii}}}. \tag{1.17}$$

where

$$m_{ii} = 1 - h_{ii} \ ; \quad 0 \leq m_{ii} \ \leq 1.$$

The diagonal elements m_{ii} of the ***projection matrix*** depict those observations with high-leverage (i.e. highly influential observations) since they are related to the distance between $\mathbf{x}_i$ and $\bar{\mathbf{x}}$. Given that $\mathbf{X}$ is of full rank, then

$$\sum_i^n h_{ii} = p \quad \Rightarrow \quad \sum_i^n m_{ii} = n - p.$$

Hence, the average of diagonal elements m_{ii} is $1 - p/n$ and high-leverage observations should have small values for m_{ii} as compared to $1 - p/n$. As a rule of thumb, from Hoaglin and Welsch ([11]), if $m_{ii} \ \leq \ 1 - 2p/n$, then the *i*th observation is a high-leverage point. Thus, **M** is a useful diagnostic tool for detecting influential observations.

Another type of ill-fitting point which arises in model-fitting is an outlier. It does not necessarily imply an influential observation in a given model. In fact, an outlier may be outweighed by neighboring X-valued points. Still, the effect that an outlying point exerts on the fit needs to be measured. The smaller the number of observations involved in a model, the greater the effect of the outlier on the model. This can be done through the diagnostic tool of *Cook's distance* which measures the effect of deleting an outlier from the data:

$$c_\ell = (\Delta_\ell\hat{\beta})'X'X(\Delta_\ell\hat{\beta}). \tag{1.18}$$

where $\Delta_\ell\hat{\beta} = \hat{\beta} - \hat{\beta}_{-\ell}$, $\hat{\beta}_{-\ell}$ denoting the usual LSE of β with the ℓth observation deleted from the data.

It gives the distance between the usual least squares estimator and the least squares estimator obtained after the ℓth observation has been deleted and provides a measure for the change in least squares estimates $\hat{\beta}$ for the deletion of the ℓth observation. It can be shown that

$$\Delta_\ell\hat{\beta} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\ell(y_\ell - \hat{y}_\ell)}{m_{\ell\ell}}; \tag{1.19}$$

hence, it can be written that

$$c_\ell = \frac{(y_\ell - \hat{y}_\ell)^2 h_{\ell\ell}}{(1 - h_{\ell\ell})^2}. \tag{1.20}$$

The residual sum of squares (RSS) will also change as a result of an observation deletion. This is measured by:

$$\begin{aligned}\Delta_\ell RSS &= RSS - RSS_{-\ell}\\ &= \frac{(y_\ell - \hat{y}_\ell)^2}{m_{\ell\ell}};\end{aligned} \tag{1.21}$$

where $RSS_{-\ell}$ represents the RSS with the ℓth case deleted. Another approach is to measure the perturbation of the fit by letting $\epsilon_i \sim N(0, \sigma^2/v_i)$. Consider

$$v_i = \begin{cases} v, & i = \ell, \\ 1 & \text{else} \end{cases}$$

where $0 \leq v \leq 1$ is a weight factor defining the matrix $\mathbf{V} = diag(v_i)$. The resulting weighted LSE of β is denoted by $\hat{\beta}(v)$.

At $v = 1$: $\hat{\beta}(1) = \hat{\beta}$, the usual least squares estimate, and

at $v = 0$: $\hat{\beta}(0) = \hat{\beta}_{-\ell}$, the least squares estimate when the ℓth point is deleted from the data.

The normal equations are changed and consequently, so are the least squares estimates. $\hat{\beta}(v)$ can be expressed as

$$\begin{aligned} \hat{\beta}(v) &= (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y} \\ &= \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\ell(1-v)(y_\ell - \hat{y}_\ell)}{[1-(1-v)h_{\ell\ell}]}. \end{aligned} \tag{1.22}$$

The perturbation effect is measured by differentiating (1.22) with respect to v:

$$\hat{\beta}'(v) = \frac{\partial\hat{\beta}\ (v)}{\partial v} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\ell\ (y_\ell - \hat{y}_\ell)}{[1-(1-v)\ h_{\ell\ell}]^2}. \tag{1.23}$$

1.1.3 Remedial Measures

If the normality assumptions made on the least squares estimates for linear models are not met in practice, then some remedial measures need to be taken. Throughout the extensive literature available on this topic, one of the most prominent solutions is to use a transformation on the data which may keep the normal linear regression form. However, the implications involved with a selected transformation may not necessarily be easy to interpret. Some of the standard remedial measures taken in case of various model departures are described below.

- *Non-linearity*

  Non-linear Least Squares Estimation:

  When a model has normally distributed errors with constant variance, but is non-linear in the independent variables, then the property of additive errors may enable a linear model through a transformation of the independent variables. The most common transformations are the following:

$$\begin{array}{lcl} x' = \log x, & & x' = \sqrt{x}, \\ x' = x^2, & \text{or} & x' = \exp x, \\ x' = \frac{1}{x}, & & x' = \exp(-x). \end{array}$$

  Such models are ***intrinsically*** linear ([7], Chapter 5). If these transformations are not possible, then alternative non-linear models may have to be considered:

$$\mathbf{y} = g(\beta, \mathbf{x}) + \epsilon,$$

  where $\mathbf{x}$ represents a vector of predictor variables, $g(\beta, \mathbf{x})$ is not linear in β. The least squares estimator of $\hat{\beta}$ for β is obtained through differentiation of the p normal equations which are not linear, unlike in the case for ordinary least squares. Hence, these normal equations are more complicated to solve. Consequently, numerical methods are usually required to be used to obtain solutions.

- *Heteroscedasticity and/or non-independent errors*

  Weighted Least Squares Estimators:

When the observations are independent yet have unequal variances, an ordinary least squares regression may yield unbiased estimates, but it will not have minimum variance. Then the observations need to be transformed in terms of weights, $w_i > 0$, $Var(y_i) = \sigma^2/w_i$ such that

$$Var(\mathbf{y}) = \sigma^2\mathbf{W}^{-1} = \sigma^2 \cdot \mathrm{diag}(1/w_1, \dots, 1/w_n).$$

Large weights w_i imply small variances and have more impact in a regression model.

Examples of weight components:

1. if the ith response is the result of an average of n_i equally variable observations, then $Var(y_i) = \sigma^2/n_i$ where $w_i = n_i$;
2. if y_i results from a total of n_i observations, then $Var(y_i) = n_i\sigma^2$ where $w_i = 1/n_i$;
3. if $Var(y_i) \propto \mathbf{x}_i$, then $Var(y_i) = \sigma^2\ x_i$ where $w_i = 1/x_i$.

Then, introducing the weight matrix, $\mathbf{W}$, the modified estimator of β is given by

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

Variance Stabilizing Transformations:

When the variances of the observations are not constant, it is possible to transform (see Rao [22], Chapter 6) the observations to make the variance constant. For this method to work, the form of the heteroscedasticity must be known, which is often not the case. Hence, in practice, one seeks transformations in a larger family and looks for an optimal member in this family, which closely follows the assumptions of the model. One such transformation, known as the Box-Cox transformation, is discussed later.

- *Non-normality of errors*

Non-parametric Techniques:

The roughness penalty approach using cubic splines is a method for relaxing the model assumptions in the normal-theory linear case. It addresses two equally important problems in curve estimation: that of finding a good fit to the data used and that of quantifying the rapid fluctuation of a curve. Consider a model

$$y = g(t) + \epsilon,$$

which is specified without placing any restrictions on the curve g. Hence, if there are no distributional assumptions made, then the normality of errors assumption is relaxed. Methods associated with the above model come under the general auspices of the topic of Non-parametric Regression and the literature on this topic is extensive (see Green and Silverman [10]).

- *Non-normality and Heteroscedasticity*

Box-Cox Transformations:

$$y^{\lambda} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

for a positive response variable $y > 0$. This transformation may bring symmetry to a skewed response and reduce the heavy tails of a distribution while still retaining the simplicity of the normal linear model. When it does not provide a good fit to the data, alternative approaches have to be explored. One such approach is to use the *generalized linear model (GLM)*, where the response is assumed to belong to the exponential family. The assumptions made here are based on the concept that the response depends on the predictors through a linear form. Thus, the linear models are generalized through

1. a *link* function which relates the expectation of the response to the linear predictor, and through
2. an exponential family distribution for the errors.

This model will be described in detail in Chapter 2 and is the highlight of this thesis.

1.2 Outline of Thesis

The next chapter introduces the GLM, with all the relevant notations. It gives the properties of estimators and computational details for estimating the parameters for common exponential families. Tests for goodness-of-fit and inclusion/exclusion of variables are also included. The basic properties of residuals in the normal theory linear models are used for extending the regression diagnostics to the generalized linear models in Chapter 3. This extension is made possible through transformed residuals, which is explained in detail in that chapter. The final chapter presents numerical illustrations of the techniques discussed in Chapter 3 and gives a hands-on experience with real data through computer programs developed using the S-Plus software application.

Chapter 2

The Generalized Linear Model

2.1 Historical Aspects

The term "generalized linear model" was first introduced by Nelder and Wedderburn in 1972. The generalized linear model has been one of the most important developments in the field of statistics in the last thirty years. Much used in applications to the social sciences and medicine, these models also play an important role in the analysis of survival data. As their name suggest, these models generalize the normal-theory linear models such that the usual linear regression component is used to describe a wider class of probability distributions, specifically the exponential family distributions. Although generalized linear models have had an important impact on statistics, most introductory statistics textbooks however, still only present normal linear models.

It was seen in Chapter 1 that an adequate linear regression model should include a y-scale which ensures the combination of constancy of variance, approximate normality of the errors, and additivity of the systematic effects. However, this scale does not always respect all three criteria. For example, if some discrete data is found to have errors with an approximate Poisson distribution, the systematic effects may be multiplicative, in which case log-linear models are usually employed. The following choices of scaling are obtained by transforming y to :

- $y^{1/2}$ to ensure approximate constancy of variance,
- $y^{2/3}$ to ensure approximate symmetry or normality, or
- $\log y$ to ensure additivity of systematic effects.

Generally, none of these scaling possibilities combine all three criteria for an adequate linear regression analysis. Alternatively, a generalized linear model encompasses exponentially distributed errors and a variance function which depends on the mean in some known way, so that there is no need to scale y for normality of errors or for constancy of variance. In fact, the scaling problem is reduced to ensuring that the systematic effects are additive. It may be considered to be an extension to the normal-theory linear model with some added modifications where the mean μ of an exponential family with response variable y is linearly related to the predictors $\mathbf{x_1}, \ldots, \mathbf{x_p}$, by a link function, $g(\mu)$. This is described in detail in the sections that follow.

2.2 Mean and Variance Functions in an Exponential Family

An observation y follows an exponential family distribution if its probability density function is given by

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

where $a, b,$ and c are some known functions, θ is the ***location parameter*** and ϕ is the ***dispersion parameter***. This is denoted by

$$y \sim \mathcal{E}(\theta,\ \phi;\ a, b, c).$$

When the dispersion parameter ϕ is known, θ is the ***canonical parameter***. The mean and variance of y are given by $b'(\theta)$ and $a(\phi)b''(\theta)$. Thus it can be written that

$$E(y) = \mu = b'(\theta), \tag{2.2}$$

$$Var(y) = a(\phi)V(\mu), \tag{2.3}$$

where

$$V(\mu) = b''(\theta)$$

is called the ***variance function***. For example, in the case of the normal distribution, $\theta = \mu, V(\mu) = 1$ and $a(\phi) = \sigma^2$. These may be derived from

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0, \tag{2.4}$$

$$E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + E\left(\frac{\partial \ell}{\partial \theta}\right)^2 = 0, \tag{2.5}$$

respectively, where ℓ is the log-likelihood function. Note that

$$\ell = \log\left\{\exp\left[(y\theta - b(\theta))/a(\phi) + c(y,\phi)\right]\right\} = (y\theta - b(\theta))/a(\phi) + c(y,\phi);$$

hence

$$\frac{\partial \ell}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}.$$

Thus, equation (2.4) yields

$$\frac{\mu - b'(\theta)}{a(\phi)} = 0, \tag{2.6}$$

which implies that

$$E(y) = \mu = b'(\theta).$$

Also

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)},$$

hence equation (2.5) yields

$$E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + E\left(\frac{\partial \ell}{\partial \theta}\right)^2 = \frac{-b''(\theta)}{a(\phi)} + \frac{Var(y)}{a^2(\phi)} = 0, \tag{2.7}$$

which gives

$$Var(y) = a(\phi)b''(\theta).$$

2.3 Description of the Generalized Linear Model

The observations belonging to a statistical model can be summarized in terms of a systematic component and a random component. In the generalized linear model

(GLM) discussed by McCullagh and Nelder [17], the random component is inherent in the exponential family distribution of the observation, while the systematic component assumes a linear structure in the predictor variables for a function of the mean. This function is known as the *link function*. When the parameter θ is modeled as a linear function of the predictors, then the link function is known as the *canonical link*. Therefore, for a given set of observations $\{y_i\}_{i=1}^n$, where y_i is considered to be associated with predictor values $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})'$, the GLM is expressed as:

$$y_i \sim \mathcal{E}(\theta_i, \phi; a, b, c) \quad - \quad \textit{random component},$$

where θ_i is assumed to depend on $\mathbf{x_i}$ through the relation

$$\eta_i = g(\mu(\theta_i)) = \mathbf{x}_i'\beta \quad - \quad \textit{systematic component}.$$

If g is the canonical link, then, the link function is specified by

$$g(\mu) = \theta. \tag{2.8}$$

In practice, a given data set may be distributed according to some unknown member of the exponential family and therefore, different link functions have to be evaluated. The link function serves to determine the scale on which linearity is assumed, and the form of the exponential family structures the variation in the data. If the parameters $\beta_1, \ldots, \beta_n$ are unrestricted, then $g(\mu)$ can take any value in R, hence the link function is determined to some extent by the domain of variation of μ. For example, if the response is a proportion, then the link function g must map the unit interval of the domain of variation onto the unrestricted range $(-\infty, \infty)$. In the case where the response is limited to being positive, g must map the positive interval onto R.

It is shown, as follows, that in the case of a canonical link, the sufficient statistic for the linear parameter β is given by $\mathbf{X'y}$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)'$ represents the design matrix of the p predictor variables and $\mathbf{y}$ represents the column vector of the n observations.

To see this, first note that $\mu = b'(\theta)$ and for the canonical link $g(\mu) = \theta$, then it follows that

$$g'(\mu) = \frac{d}{d\mu}g(\mu) = \frac{d\theta}{d\mu} = \left[\frac{d\mu}{d\theta}\right]^{-1} = \frac{1}{b''(\theta)},$$

hence by (2.3)

$$g'(\mu) = \frac{1}{V(\mu)}. \tag{2.9}$$

This fact is used in deriving the maximum likelihood estimator of β which will be consequently shown to depend on the observations $\mathbf{y}$ through $\mathbf{X}'\mathbf{y}$, proving the sufficiency. Here, the log-likelihood function is given by

$$\ell(\beta|\mathbf{y}) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right], \tag{2.10}$$

where $\theta_i = \mathbf{x}_i'\beta$. Now, the differentiation of the likelihood function in equation (2.10) gives

$$\frac{\partial\ell(\beta)}{\partial\beta} = \sum_{i=1}^{n}\frac{\partial}{\partial\theta_i}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]\frac{\partial\theta_i}{\partial\beta} = \sum_{i=1}^{n}\mathbf{x}_i \cdot \frac{(y_i - \mu_i)}{a(\phi)}.$$

Using equation (2.9) along with the above equation produces

$$\sum_{i=1}^{n}\mathbf{x}_i \cdot (y_i - \hat{\mu}_i) = 0,$$

which implies for canonical links that

$$\mathbf{X}'\mathbf{y} = \mathbf{X}' \cdot \mathbf{q}(\hat{\beta}),$$

for some nonlinear function q. This is attributed to the fact that $g(\mu) = \theta$ holds for canonical links only. Hence,

$$\mu = g^{-1}(\theta) \quad \Rightarrow \quad \hat{\mu}_i = g^{-1}(\theta_i) = g^{-1}(\mathbf{x}_i'\hat{\beta}).$$

Now, canonical links for the binomial, Poisson and gamma families are given respectively by the logit, log and inverse transformations. Consider the probability distribution of the proportion y based on a sequence of m identical Bernoulli trials with probability of success π, then

$$f(y;\, \theta, \phi) = \exp\left\{\frac{[y\theta - \log(1+e^{\theta})]}{1/m} + \log\binom{m}{my}\right\},$$

where $\theta = \log\frac{\pi}{1-\pi}$, hence the canonical link is given by the logit transformation and the generalized linear model is given by

$$\eta_i = g(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \sum_{r=1}^{p} x_{ir}\beta_r.$$

For the Poisson data with mean μ, the probability distribution function is denoted by:

$$f(y;\, \theta, \phi) = \exp\{(y\theta - e^{\theta}) \;-\; \log(y)\},$$

where $\theta = \log\mu$, then clearly the log transformation yields a canonical link. Similarly for the gamma data with density

$$f(y) = \frac{1}{k^{\alpha}\Gamma(\alpha)}e^{-y/k}y^{\alpha-1},$$

it may be reparametrized such that $\alpha = 1/\phi$ and $k = \; -\phi/\theta$, hence to get

$$f(y;\, \theta, \phi) = \exp\{\frac{y\theta + \log(-\theta)}{\phi} + c(y, \phi)\},$$

whereby

$$c(y, \phi) = [(1/\phi - 1)\log(y\phi) + \log(\phi) - \log\Gamma(1/\phi)].$$

Therefore, $\mu = k\alpha = -1/\theta$ and consequently, the canonical link is given by

$$g(\mu) = -\frac{1}{\mu}.$$

Table 2.1: *Dispersion Parameter, Canonical Link and Variance Function for Distributions of the Exponential Family*

DISTRIBUTION	Notation	$a(\phi)$	$\theta = g(\mu)$	Name	$V(\mu)$
Normal	$N(\mu, \sigma^2)$	σ^2	μ	identity	1
Poisson	$Poi(\mu)$	1	$\log(\mu)$	log	μ
Binomial	Bin(m, π)	$\frac{1}{m}$	$\log(\frac{\mu/m}{1-\mu/m})$	logit	$\frac{\mu}{m}(1-\frac{\mu}{m})$
Gamma	$Gam(\alpha, k)$	$1/\alpha$	$-\frac{1}{\mu}$	inverse	μ^2
Inverse Gaussian	$Inv(\mu, \sigma^2)$	σ^2	$-\frac{2}{\mu^2}$	1/mu^2	μ^3

Table 2.1 gives canonical links and other components for common distribution families with respect to the exponential family given by equation (2.1) [17]. The choice of a proper link function that will satisfy the criterion of the domain of variation μ is based on:

1. how the link function will easily interpret the parameters in the linear predictor;
2. how the link fits to the data; and
3. the existence of a simple sufficient statistic.

Possible link functions associated to some important members of the exponential family are cited in Table 2.2. In summary, generalized linear models make up a general class of probabilistic regression models with the assumptions that:

(1) the response probability distribution is a member of the exponential family of distributions;

(2) the response $y_i \quad i = 1, \ldots, n$ is a set of independent random variables;

(3) the explanatory variables are linearly combined to explain systematic variation in a function of the mean.

In a practical data situation, GLM fitting involves the following:

- choosing an error distribution that is relevant;
- identifying the independent variables to be included in the systematic component; and
- specifying the link function.

The next section presents the maximum likelihood method for estimating the regression parameters assuming that the above have been specified.

2.4 Maximum Likelihood Estimation for the GLM

If the probability specifications of an exponential family model are known by $f(y, \theta)$, then the best way to fit a generalized linear model is by Maximum Likelihood Estimation of the parameters β for the data observed (Silverman and Green [10]). With many desirable properties of maximum likelihood estimators such as consistency, efficiency, sufficiency and asymptotic normality, it is natural to consider such a method for GLMs. In general, the maximum likelihood equations which result from GLMs cannot be solved explicitly and hence recourse must be made to numerical methods. Three methods are described in this section: the Newton-Raphson method, the Fisher Scoring method, and the Iteratively Weighted Least Squares method. But first, the maximum likelihood equations are derived. Given the responses $y_1, \ldots, y_n$, where y_i is considered to be generated from a member of the exponential family $\mathcal{E}(\theta, \phi;\ a, b, c)$, the likelihood function is written as

$$\prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}. \tag{2.11}$$

Then the log-likelihood is given by

$$l(\beta;\ \phi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right] = \sum_{i=1}^{n} \ell_i, \tag{2.12}$$

whereby ℓ_i is the ith component to the log-likelihood and is therefore given by

$$\ell_i = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi). \tag{2.13}$$

The likelihood implicitly depends on the parameters $\beta_j, j = 1, \ldots, p$, firstly through the link function $g(\mu)$ and secondly through the linearity that it encompasses with respect to β_j values. The derivatives of the log-likelihood with respect to β_j, otherwise known as the score functions, are evaluated by the chain rule:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} = 0\ ;\ j = 1, \ldots, p. \tag{2.14}$$

It is easily seen that

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}; \tag{2.15}$$

$$\frac{d\theta_i}{d\mu_i} = (b''(\theta_i))^{-1} = V^{-1}(\mu_i), \tag{2.16}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \sum_{j=1}^{p} \beta_j x_{ij}}{\partial \beta_j} = x_{ij}. \tag{2.17}$$

Hence, the score functions reduce to

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} \ ; \ j = 1, \ldots, p. \tag{2.18}$$

In a vector form, the score equations are given by

$$(\mathbf{y} - \mu)'\mathbf{D}(\mu)\mathbf{X} = 0, \tag{2.19}$$

where

$$\mathbf{D}(\mu) = \mathrm{diag}(d_{ii}), \quad d_{ii} = 1/V(\mu_i)g'(\mu_i).$$

The maximum likelihood estimator of β is obtained by solving (2.19) using the linearity found in $g(\mu) = \mathbf{X}\beta$, where $g(\mu) = (g(\mu_1), \ldots, g(\mu_n))'$. Numerical methods to solve (2.19) are essentially iterative. Common to all these methods is the starting value of the estimate. With the ultimate aim of obtaining a "good" starting value of the estimate, the following technique is employed using the approximate linearized form of $g(y) = g(\mu) + (y - \mu)g'(\mu)$. The ***adjusted dependent variate***, z which depends on both y and μ is introduced.

$$\begin{aligned} z &= \eta + (y - \mu)\frac{d\eta}{d\mu} \\ &= g(\mu) + (y - \mu)g'(\mu). \end{aligned} \tag{2.20}$$

Given that the variance of z is $a(\phi)[g'(\mu)]^2V(\mu)$, an initial estimate of β may be obtained by Weighted Least Squares of $\mathbf{z}$ (with $\mu = y$) on $\mathbf{X}$, with variance-covariance matrix given by a diagonal matrix whose components are specified by

$$\begin{aligned} w_i &= \frac{1}{V(\mu_i)[g'(\mu_i)]^2} \\ &= \frac{1}{Var(z_i)}. \end{aligned} \tag{2.21}$$

Known as the ***working weights matrix***, this matrix is denoted by $\mathbf{W}$. In cases where repeated observations occur at a given design point, y_i is replaced by the average of the sample observations. Since the average also belongs to the same exponential family, with the variance replaced by $a(\phi)V(\mu_i)/n_i$, n_i being the number of observations on which the sample mean is based upon, the working weights matrix contains diagonal elements given by

$$\begin{aligned} w_i &= \frac{n_i}{V(\mu_i)[g'(\mu_i)]^2} \\ &= \frac{1}{Var(z_i)}. \end{aligned}$$

Clearly, the score equations can be expressed as

$$\sum_{i=1}^{n}(y_i - \mu_i)g'(\mu_i)w_i\, x_{ij} = 0, \tag{2.22}$$

which, when transformed to the adjusted variates yield the following

$$\sum_{i=1}^{n}(z_i - g(\mu_i))w_i x_{ij} = 0, \tag{2.23}$$

or equivalently, solving for the weighted least squares estimator from the model

$$E(\mathbf{z}) = \mathbf{X}\beta, \qquad Var(\mathbf{z}) = a(\phi)\text{diag}(1/w_1, \dots, 1/w_n).$$

Both **z** and **W** are used for maximum likelihood estimation through a *weighted least squares regression*. This process is iterative, since both **z** and **W** depend on the fitted values of current estimates available. Some scoring methods are needed to measure the iteration variations for a weighted least squares regression of a GLM, until convergence is reached.

2.4.1 The Newton-Raphson Method

The *Newton-Raphson method* presents a numerical approach to calculating the maximum likelihood estimate $\hat{\beta}$. This iterative process begins with a weighted least squares estimator obtained from the initial solution of (2.23). A Taylor-series expansion of $\ell(\hat{\beta})$ about $\ell(\hat{\beta}^{(i)})$ is used:

$$\begin{aligned} 0 = \frac{\partial \ell}{\partial \beta}\Big|_{\hat{\beta}} &\approx \frac{\partial \ell}{\partial \beta}\Big|_{\hat{\beta}^{(i)}} + \frac{\partial^2 \ell}{\partial \beta \partial \beta'}\Big|_{\hat{\beta}^{(i)}} (\hat{\beta} - \hat{\beta}^{(i)}), \qquad (2.24) \\ \hat{\beta} - \hat{\beta}^{(i)} &\approx \left[\left(- \frac{\partial^2 \ell}{\partial \beta \partial \beta'} \right)^{-1} \cdot \frac{\partial \ell}{\partial \beta} \right]_{\hat{\beta}^{(i)}} \\ &= \delta^{(i)}. \qquad (2.25) \end{aligned}$$

An updated estimate of $\hat{\beta}$ is then obtained:

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} + \delta^{(i)}. \qquad (2.26)$$

This is iteratively repeated until convergence is obtained [10].

2.4.2 Fisher's Scoring Method

If the negative second-derivative matrix, or the Hessian matrix, is not positive definite at every iteration (i.e. if it is not invertible), then the Newton-Raphson's algorithm is no longer valid. In this case, the Hessian matrix is replaced by its expectation, obtaining ***Fisher's scoring algorithm.*** This method is simple since the expected matrix is more likely to be positive definite as

$$-E\left[\frac{\partial^2 \ell}{\partial\beta\partial\beta'}\right] = E\left[\frac{\partial \ell}{\partial\beta}\frac{\partial \ell}{\partial\beta'}\right], \tag{2.27}$$

which is the expectation of a positive definite matrix. Thus, the iterative process for Fisher's scoring algorithm is given by:

$$\hat{\beta}^{(i-1)} = \hat{\beta}^{(i)} + \left\{-\left(E\left[\frac{\partial^2 \ell}{\partial\beta\partial\beta'}\right]\right)^{-1}\right\}\frac{\partial \ell}{\partial\beta}. \tag{2.28}$$

where $\delta^{(i)} = -\left(E\left[\frac{\partial^2 \ell}{\partial\beta\partial\beta'}\right]\right)^{-1}\frac{\partial \ell}{\partial\beta}$ is evaluated at the previous iteration. For evaluating the derivatives in (2.28), the linear predictor η_i is used where $\eta_i = x_i'\beta$:

$$\frac{\partial \ell}{\partial\eta_i} = \frac{\partial \ell_i}{\partial\theta_i}\frac{d\theta_i}{d\eta_i} = \frac{\partial \ell_i}{\partial\theta_i} \times (\frac{d\eta_i}{d\mu_i}\frac{d\mu_i}{d\theta_i})^{-1} \tag{2.29}$$

$$\Rightarrow \qquad \frac{\partial \ell}{\partial\eta_i} = (\frac{y_i-\mu_i}{a_i(\phi)}) \times \{g'(\mu_i)b''(\theta_i)\}^{-1}, \tag{2.30}$$

and

$$\begin{aligned} -E\left[\frac{\partial^2 \ell}{\partial\eta\partial\eta'}\right]_{ij} &= \frac{d\mu_i}{d\eta_i} \times \{a(\phi)g'(\mu_i)b''(\theta_i)\}^{-1} \\ &= g'(\mu_i) \times \{a\phi g'(\mu_i)b''(\theta_i)\}^{-1} \\ &= \{a(\phi)g'(\mu_i)^2 b''(\theta_i)\}^{-1}. \end{aligned}$$

Note that $-E\left[\frac{\partial^2 \ell}{\partial\eta\partial\eta'}\right]_{ij} = [a(\phi)]^{-1}w_{ij}$ for $i = j$, and it is $= 0$ for $i \neq j$. Consider $\mathbf{z}^{(0)}$ to be the initial n-vector with

$$z_i^{(0)} = (y_i - \hat{\mu}_i)g'(\mu_i).$$

Then it follows that

$$\frac{\partial \ell}{\partial\eta} = \frac{1}{a(\phi)}\mathbf{W}\mathbf{z}^{(0)}, \tag{2.31}$$

from (2.30). Furthermore,

$$-E\left(\frac{\partial^2 \ell}{\partial\eta\partial\eta'}\right) = \frac{1}{a(\phi)}\mathbf{W}. \tag{2.32}$$

Since $\eta = \mathbf{X}'\beta$, then by the chain rule

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \mathbf{X}'\frac{\partial \ell}{\partial \eta}\\ &= \frac{1}{a(\phi)}\mathbf{X}'\mathbf{W}\mathbf{z}^{(0)}\end{aligned} \tag{2.33}$$

and

$$-E\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right) = \mathbf{X}^{\mathbf{T}}\left[-E(\frac{\partial^2 \ell}{\partial\eta\partial\eta^T})\right]\mathbf{X}. \tag{2.34}$$

The Fisher's scoring algorithm yields the following sequence of updated estimates:

$$\hat{\beta}^{\mathbf{i-1}} = \hat{\beta}^{\mathbf{i}} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}. \tag{2.35}$$

The dispersion parameter ϕ is eliminated because $a(\phi)$ gets canceled in the multiplication, hence it is called a *nuisance* parameter (McCullagh and Nelder [17]).

2.4.3 Iteratively Weighted Least Squares (IWLS)

As indicated in Section 2.4, the introduction of the adjusted dependent variate $\mathbf{z}$ results in the following equation for the MLE $\hat{\beta}$ [see (2.23)];

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}.$$

However, the $\mathbf{z}$ and $\mathbf{W}$ depend on the unknown $\hat{\mu}$, hence this equation gives rise to the iterative process

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)}.$$

This is known as the method of ***iteratively weighted least squares, IWLS***. The starting value of the iteration is obtained by substituting $\hat{\mu}^0 = \mathbf{y}$. At each iteration i, a weighted least squares regression of the working response variate $\mathbf{z}^{(i)}$ on the design matrix $\mathbf{X}$ is obtained with the working weights matrix $\mathbf{W}^{(i)}$, where $\mathbf{z}^{(i)}$ and $\mathbf{W}^{(i)}$ are obtained by replacing μ with $\hat{\mu}^{(i)} = g^{-1}(X\hat{\beta}^{(i)})$. This algorithm can thus be summarized as follows :

- Start with a sufficient statistic from the data to get an initial fitted value vector $\hat{\mu}^{(0)}$.
- From this statistic, the link function g is used to derive an initial linear predictor $\hat{\eta}^{(0)}$.
- Calulate $(\frac{d\eta}{d\mu})_0$ and $V(\hat{\mu}^{(0)}) = (\frac{d\mu}{d\theta})_0$.

These statistics are used in creating the starting adjusted dependent variate and working weight matrix as follows:

$$\mathbf{z}^{(0)} = \hat{\eta}^{(0)} + (\mathbf{y} - \hat{\mu}^{(0)}) \left(\frac{d\eta}{d\mu}\right)_0,$$
$$(\mathbf{W}^{(0)})^{-1} = \left(\frac{d\eta}{d\mu}\right)_0^2 V^{(0)}.$$

A weighted least squares regression is carried out of $\mathbf{z}^{(0)}$ on $\mathbf{X}$ for the model $E(\mathbf{z}) = \mathbf{X}\beta$ with the working weights matrix, $\mathbf{W}^{(0)}$ to obtain a first maximum likelihood estimate:

$$\hat{\beta}^{(1)} = (\mathbf{X}'\mathbf{W}^{(0)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(0)}\mathbf{z}^{(0)},$$

which is then used to obtain updated values of $\hat{\eta}$ and $\hat{\mu}$:

$$\hat{\eta}^{(1)} = \mathbf{X}'\hat{\beta}^{(1)},$$

$$\hat{\mu}^{(1)} = g^{-1}(\hat{\eta}).$$

This process is repeated to update the regression estimates at each iteration via a scoring algorithm, until the variation from one iteration to the next is sufficiently small. The maximum likelihood estimation method through the *IWLS* procedure is an extension to the non-iterative least squares method of estimation for normal-theory linear models, with $\mathbf{W}^{1/2}\mathbf{X}$ as the design matrix and the adjusted dependent variate $\mathbf{W}^{1/2}\mathbf{z}$ as the response variable.

At convergence, if it occurs, $\mathbf{z}$ becomes $\mathbf{z} = \mathbf{X}\hat{\beta} + \mathbf{W}^{-1}(\mathbf{y} - \hat{\mu})$ so that the maximum likelihood estimate of β is:

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}. \tag{2.36}$$

If the working weights matrix $\mathbf{W} = \mathbf{I}$ (the identity matrix), then the maximum likelihood and least squares methods coincide. No iteration is required for the maximum likelihood estimation:

$$\begin{aligned} \hat{\beta}^{(i-1)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta^{(i)}. \\ \Rightarrow \qquad \beta^{(i-1)} &= \beta^{(i)} \end{aligned}$$

Hence, the *IWLS* method extends the least squares procedure beyond the linear model to the generalized linear model that includes the binomial, Poisson, normal, inverse normal, gamma, exponential, and multinomial distributions.

An interesting point to note is that the working weights matrix used in *IWLS*, $\mathbf{W}$, is updated at each iterative step of *IWLS* so that each element of $\mathbf{W}$, w_{ii} is updated too for each observation i. Hence, $\mathbf{W}$ depends entirely on the fit of the model, and not at all on the likelihood equations $\mathbf{X}'(\mathbf{y} - \hat{\mu}) = \mathbf{0}$ used to determine $\hat{\beta}$. In contrast, the weights determine the fit in the weighted least squares method.

The basic components of the generalized linear model, as an extension to the normal theory model, may be summarized in the following table:

2.5 The Goodness of Model Fit

As previously stated, the link function which is used to describe the systematic component is often unknown. Canonical links may simplify the mathematics, but they may not necessarily represent the best prediction. A natural question bound to arise in fitting a GLM is "how good is the link function used?", in comparison to some other potential link functions. In fact, the model fit is questioned. Other issues attributable to model fitting are based on assumptions such as the exponential family distribution of the observations, the constancy of the dispersion parameter and the independence of the observations, much like those seen in the normal-theory linear models, and the issue of identifying influential observations.

A common goal in postulating the systematic effect is to have only as many independent variables as necessary for a good fit. Consequently, measures which can determine the quality of the fit and statistical tests for keeping the variates in the model are sought for. In particular, the two most useful goodness-of-fit statistics are the *deviance measure* and the *Pearson statistic*. The deviance measure is motivated by the discrepancy between the maxima of the observed and the expected (under the model) log-likelihood functions. Conversely, the Pearson statistic measures the relative difference between the observed and the fitted values. Both of these statistics can be approximated by the χ^2 distribution with corresponding degrees of freedom. In either case, a large deviance or chi-square value implies poorly fitted observations with respect to the model.

2.5.1 The Deviance Function

The maximized likelihood for a given model may be considered to be an indicator of the goodness-of-fit. For example, the ratio of the maximized likelihoods under two models as a measure of the goodness of one model over the other may be such an indicator, or alternatively, taking the logarithm of this ratio. The deviance measure D is thus defined as twice the logarithm of the likelihood ratio. Subsequently, a related measure called the *scaled deviance* D^* is defined as such:

$$D^* = \frac{D}{\phi} = 2[\ell_{max} - \ell(\theta(\hat{\beta}))], \tag{2.37}$$

where $\ell_{max} = \max_{\theta} \ell(\theta, \phi)$ and $\ell(\theta(\hat{\beta})) = \max_{\theta;\, g(\mu)=X\beta}(\ell(\theta), \phi)$. Since a maximized $\ell\{\theta(\hat{\beta})\}$ implies a small $D*$, a good fit is indicated by small values of the deviance. The table below expresses the deviance function for the different members of the exponential family with their respective canonical links. Note that μ_i is the value of $\mu_i = E(y_i)$ for the model considered.

The unscaled version of the deviance is

$$\begin{aligned} D &= \phi D^* \\ &= 2\sum_{i=1}^{n}[\{y_i\theta_i - b(\theta_i)\} - \{y_i\hat{\theta}_i^{(M)} - b(\hat{\theta}_i^{(M)})\}] \qquad (2.38) \\ &= \sum_i d_i, \qquad (2.39) \end{aligned}$$

from which the deviance residuals r_D are obtained given by

$$r_{D_i} = \text{sign}\,(y_i - \hat{\mu}_i)\sqrt{d_i}. \tag{2.40}$$

In (2.38), the parameter $\hat{\theta}_i^{(M)} = MLE$ of θ_i under the fitted model .

Each d_i measure contributes to the deviance. The value of θ_i which maximizes the likelihood function, for each ith observation, is $\theta_i^{(M)}$ whereby $b'(\theta_i^{(M)}) = y_i$.

2.5.2 The Pearson Statistic

The Pearson statistic is defined using the weighted least squares approach, which provides the following chi-square goodness-of-fit:

$$\chi^2 = \min_{\beta} \sum_{i=1}^{n} w_i(z_i - x_i'\beta)^2. \tag{2.41}$$

This can be written as

$$\chi^2 = \sum_{i} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{2.42}$$

This measure is computationally simpler than the deviance measure but it is more useful for distributions closer to the Normal family, as it resembles the *RSS* under the normal-theory for other diagnostic purposes. However, when the probability density function of the observations is markedly asymmetric, the outliers may not be well detected by Pearson residuals. Conversely, the deviance residuals will detect outliers better in these situations.

2.5.3 Residuals and the Projection Matrix

The usefulness of residuals $r_i = y_i - \hat{y}_i$ where $\hat{y}_i$ is from the model fit as used for diagnostic purposes in normal-theory linear models, does not apply in GLMs. However, as $\sum r_i^2$ serves as a measure of goodness-of-fit in normal-theory models, it would be best if the two measures given here could be decomposed into components, which in turn could serve as modified residuals in GLMs. Using this concept, it can be seen that

$$\chi^2 = \sum_{i=1}^{n} r_{P_i}^2. \tag{2.43}$$

where

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}, \tag{2.44}$$

which are the weighted residuals or the Pearson residuals.

Similarly,

$$D = \sum_{i=1}^{n} r_{Di}^2.$$

where

$$r_{Di} = \pm\sqrt{2\{\ell(\hat{\theta}_i;\, y_i) - \ell(\mathbf{x}_i\hat{\beta};\, y_i)\}}. \tag{2.45}$$

These are the deviance residuals (see Pregibon [20]). Hence like in normal-theory models, both the Pearson and deviance residuals may be useful in developing diagnostic tools in GLMs. This will be discussed in Chapter 3.

For detecting influential observations and outliers, the use of the adjusted dependent variate z permits the use of the projection matrix

$$\mathbf{M_W} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2} \tag{2.46}$$

using the transformation $\mathbf{X} \to \mathbf{W}^{1/2}\mathbf{X} = \mathbf{X_W}$ and the least squares theory as introduced in Chapter 1. Hence

$$\mathbf{M_W} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X_W}(\mathbf{X}'_\mathbf{W}\mathbf{X_W})^{-1}\mathbf{X}'_\mathbf{W}, \tag{2.47}$$

shares the properties of a projection matrix. As mentioned in Chapter 1, the diagonal elements $m_{ii}^{(w)}$ can be used for diagnostic purposes. It is also interesting to note that

$$\mathbf{W}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{M}\mathbf{W}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}}). \tag{2.48}$$

This can be seen as follows [20]:

$$\begin{aligned}\mathbf{W}^{-1}(\mathbf{y}-\hat{\mu}) &= \mathbf{z}-\mathbf{X}\hat{\beta}\\ &= \mathbf{W}^{-1/2}\mathbf{M}\mathbf{W}^{1/2}(\mathbf{X}\hat{\beta}+\mathbf{W}^{-1}(\mathbf{y}-\hat{\mu}))\\ &= \mathbf{W}^{-1/2}\mathbf{M}\mathbf{W}^{-1/2}(\mathbf{y}-\hat{\mu}). \qquad (2.49)\end{aligned}$$

Consider multiplying the LHS and RHS by $\mathbf{W}^{1/2}$ to get

$$\mathbf{W}^{-1/2}(\mathbf{y}-\hat{\mu}) = \mathbf{M}\mathbf{W}^{-1/2}(\mathbf{y}-\hat{\mu}). \qquad (2.50)$$

This implies that

$$\mathbf{M}_W\,\chi = \chi, \qquad (2.51)$$

where χ denotes the vector of Pearson residuals for the canonical link. Hence to conclude, M_W spans the space of the Pearson residuals under the condition that the canonical link is used.

2.6 Alternative Models

For both normal linear models and GLMs, the form of the distribution and therefore the likelihood function is known. However, in practice this information may not be available. Then some features of the data need to be evaluated such as how the mean response μ relates to the independent covariates, how the variability of the response relates to μ, and whether the observations are all independent. *Quasi-likelihood* estimation is based on the idea of incomplete distribution specification. It is determined entirely by the mean and variance functions. Like the optimal property of linear least squares estimates, quasi-likelihood estimates have asymptotic optimality properties.

Definition 2.1 *Let* $\mathbf{y}$ *be the vector of responses of length* n*,*

(1) $E(\mathbf{y}) = \mu$,

(2) $V(\mathbf{y}) = a(\phi)V(\mu)$, $V(\mu) = \ diag[V(\mu_i)]$.

Consider g *to be the link function which relates the mean response* μ_i *to the systematic part of a GLM:*

$$g(\mu_i) = \mathbf{x}_i'\beta,$$

the variance is assumed to be a function of μ*:*

$$Var(y_i) = a(\phi)V(\mu_i); \qquad \phi > 0.$$

Only the form of the mean and variance functions are necessary for the quasi-likelihood function.

The quasi-likelihood function is defined by the quadratic form

$$Q(\mu; \mathbf{y}) = (\mathbf{y} - \mu)'V(\mathbf{y})^{-1}(\mathbf{y} - \mu)$$

where parameters β *relate to* μ *depending on* $\mathbf{X}$ *in a nonlinear model, written* $\mu = \mu(\beta)$ *. Then like the least squares function for normal linear models, the quasi-likelihood function estimates* β *which minimize the weighted sum of squares resulting in the following score like equations:*

$$\sum_i^n \frac{\partial \mu_i}{\partial \beta_j} \frac{(y_i - \mu_i)}{V(\mu_i)} = 0\,, \ j = 1 \ldots, p. \tag{2.52}$$

This form is useful when the y *components of the response vector have unequal variances.*

Several likelihood functions generated from the exponential family can be derived from this quasi-likelihood function when an appropriate variance function is assumed.

Since $V(\mu)$ is most often proportional to $Cov(\mathbf{y})$, it is safe to assume that $V(\mu) = Cov(\mathbf{y})$. Here, the proportionality of $Cov(\mathbf{y})$ to a matrix of known constants in normal linear models is extended to the proportionality to a matrix of known functions of the mean vector μ for nonlinear models. Then it follows from the least squares that

(1) the estimate $\bar{\beta}$ minimizes the quadratic form of $Q(\mu; \mathbf{y})$ over $\mu(\beta)$, and

(2) the weighted sum of squares estimate $\bar{\beta}$ will satisfy the quasi-score equations

$$\mathbf{U}(\mu; \mathbf{y}) = \left(\frac{\partial \mu}{\partial \beta_j}\right)' \frac{\mathbf{y} - \mu(\hat{\beta})}{V(\mu)} = 0.$$

This approach is the GLM counterpart of the least squares approach to the usual linear model with normality assumption. It makes a base for using the generalized linear model without adhering to a particular exponential family assumption.

Table 2.2: *Distribution Functions with their Associated Links*

	Family Member				
Link	Normal	Poisson	Binomial	Gamma	Inverse Gaussian
identity	X	X		X	
log		X		X	
logit			X		
probit			X		
cloglog			X		
inverse				X	
$1/\text{mu}^2$					X
sqrt					X

Table 2.3: *An Extension of the Normal-Theory Linear Model to the GLM*

Normal-Linear		GLM
$\mathbf{y}$ – dependent variate		$\mathbf{z}$ – adjusted dependent variate
$\hat{\mu}$ – linear predictor		$\hat{\eta}$ – linear predictor
$\mathbf{s}^2$ – the residual variance	*replaced by*	$\hat{\phi}\mathbf{V}(\hat{\mu})$
$\mathbf{X}$		$\mathbf{W}^{1/2}\mathbf{X}$
$\mathbf{H}$ – the hat(projection) matrix		$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$

Table 2.4: *Deviance Function for Exponential Family Distributions*

DISTRIBUTION	D
Normal	$\sum(y_i - \hat{\mu}_i)^2$
Poisson	$2\sum\{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$
Binomial	$2\sum\{[y_i \log(y_i/\hat{\pi}_i)] + (1 - y_i)\log[(1 - y_i)/(1 - \hat{\pi}_i)]\}$
Gamma	$2\sum[-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i]$
Inverse Gaussian	$\sum(y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$

Chapter 3

Residual Diagnostic Measures

3.1 Modified Residuals

Two types of residuals were introduced in Chapter 2, namely, the Pearson type (r_{P_i}) and the deviance-based (r_{D_i}). It is found that the deviance-based residuals provide better goodness-of-fit measures for GLMs than does the Pearson statistic, even though the latter is more nearly chi-squared distributed. The reasons for this are the almost normality of the deviance-based residuals and the convenience in their use for likelihood-based inference. In fact, deviance-based residuals are especially appropriate for identifying individual poorly fitted observations. Here, the dispersion parameter ϕ is considered to be known, in which case the exponential family is essentially given by the density function

$$f(y_i; \theta_i) = h(y_i) \exp\{y_i\theta_i - b(\theta_i)\},$$

where the scale parameter is omitted. The θ_i are assumed to follow the tentative model given by

$$\theta_i = g(\mathbf{x}_i'\beta),$$

where $g(\cdot)$ is a specified function, $\mathbf{x}_i$ is a vector of known covariables, and β is a vector of unknown parameters. The residuals discussed in this chapter, however, are useful in a more general setting than just for the exponential family distribution. The diagnostics are based on the asymptotic distribution of residuals. In GLM, two types of asymptotic situations arise:

(1) when $n \to \infty$, and

(2) when the index $m \to \infty$, which is equivalent to each y_i becoming approximately normal.

These situations are referred to as $n - asymptotics$ and $m - asymptotics$ respectively. In situation (2), m would correspond to the sample size for the binomial distribution, the means for Poisson, or the gamma shape parameters. Hence m can be thought of as a common factor multiplying the exponents in these aforementioned densities. The standard asymptotic results for estimation and hypothesis testing with respect to β apply if either m or n is large. However, asymptotic results pertaining to individual case diagnostics require large m, irrespective of n. The problem arises when n is large but m is not. This is a common occurring situation for residual distributions. Distinguishing between first- and second- order $m - asymptotics$ (i.e.: corresponds to the stochastic convergence of order $m^{-1/2}$ and m^{-1} respectively), the second-order asymptotic results are more useful when m is small than the first-order ones (see Pierce and Schafer [21]).

Consider residuals that are approximately normally distributed. In the following models, θ_i is treated as known, but in practice, it is replaced by

$$\hat{\theta}_i = g(x_i'\hat{\beta}). \tag{3.1}$$

Three types of residuals are considered:

(1) Linear

$$R_L(y, \theta) = (y - E_\theta(y))/SD_\theta(y), \tag{3.2}$$

where E = mean and SD = standard deviation,

(2) Transformed linear

$$R_T(y, \theta) = (t(y) - E_\theta[t(y)])/SD_\theta[t(y)], \tag{3.3}$$

where $t(\cdot)$ is a specified transformation depending on the particular distribution of y.

There are two ways to go about in choosing a transformation $t(\cdot)$. One way lets the first-order $m-$ *asymptotic* skewness of $t(y)$ be zero (i.e. symmetrizing) and hence approximate normality may be achieved. This is done using primarily the Anscombe residual.

(a) Anscombe Residual (see [2])

Starting with a function which will make the distribution of $A(y)$ as normal as possible, standardized with 0 mean and unit variance to the first order in μ, for the likelihood functions in GLMs, the function $A(\cdot)$ is given by:

$$A(\mu) = \int \frac{1}{V^{1/3}(\mu)} d\mu.$$

A 'symmetrizing transformation '(see Chaubey and Mudholkar [3]) on $t(\cdot)$ (for $t' \neq 0$) can be obtained by solving

$$t(\theta) = \int e^{-\frac{2}{3}\int \frac{f_1(\theta)}{f_2(\theta)} d\theta} d\theta, \tag{3.4}$$

whereby $f_1(\theta) = \xi_3(\theta) - 3\xi_1(\theta)\xi_2(\theta)$; $f_2(\theta) = \xi_4(\theta) - \xi_2^2(\theta)$; and $\xi_j(\theta) = E(X_n - \theta)^j$, $j = 1, 2, 3, 4$.

In the case of the binomial distribution with proportions π and m trials, the symmetrizing transformation is given by

$$t(\theta) = \int \frac{1}{\theta^{1/3}(1-\theta)^{1/3}} d\theta; \tag{3.5}$$

which can be solved numerically using the incomplete beta function, with no explicit solution.

For a Poisson distribution with mean μ, the transformation yields

$$t(\mu) = \int \frac{1}{\mu^{1/3}} d\mu = \frac{3}{2}\mu^{2/3}. \tag{3.6}$$

As for the gamma distribution with mean μ and shape parameter α, the transformation is known as the Wilson-Hilferty cube-root transformation

$$t(\mu) = \int (\frac{\alpha}{\mu^2})^{1/3} d\mu = 3\alpha^{1/3}\mu^{1/3}. \tag{3.7}$$

An alternative to the approximate normality objective is to choose a $t(\cdot)$ that will make the $m-$*asymptotic* variance of $t(y)$ constant in θ.

(b) <u>Variance Stabilizing Residual</u> (see [22])

If $\{t_n\}, n = 1, 2, \ldots$, is a sequence of statistics such that

$$\sqrt{n}(t_n - \theta) \rightarrow y \sim N(0, \sigma^2(\theta)),$$

i.e. $\sqrt{n}(t_n - \theta)$ has an asymptotic distribution,

then it follows that if g is a function with the first derivative existing and being continuous, $g'(\theta) \neq 0$, then

$$\begin{aligned} & \sqrt{n}[h(t_n) - h(\theta)] \to y \sim N(0, (h'(\theta)\sigma(\theta))^2) \\ \Rightarrow \quad & \frac{\sqrt{n}[h(t_n) - h(\theta)]}{h'(t_n)} \to y \sim N(0, \sigma^2(\theta)) \end{aligned}$$

and further, if $\sigma(\theta)$ is continuous, then

$$\frac{\sqrt{n}[h(t_n) - h(\theta)]}{h'(t_n)\sigma(t_n)} \approx N(0, 1).$$

By the Taylor series expansion,

$$h(t_n) - h(\theta) = (t_n - \theta)(h'(\theta) + \epsilon_n),$$

$$h(t_n) \doteq h(\theta) + (t_n - \theta)h'(\theta).$$

Now if h is a function such that $h'(\theta)\sigma(\theta) = c$ where c is independent of θ, then

$$\frac{dh}{d\theta} = \frac{c}{\sigma(\theta)} \Rightarrow h = c\int \frac{1}{\sigma(\theta)}\, d\theta.$$

Then the asymptotic variance of $h(t_n)$ is independent of θ:

$$Var[h(t_n)] = \sigma^2(\theta)h'^2(\theta) = c^2.$$

If y is a random variable with $B(m, \pi)$, then the variance-stabilizing transformation for the binomial distribution is

$$h(\pi) = \int \frac{1}{\pi^{1/2}(1-\pi)^{1/2}}\, d\pi = \arcsin\sqrt{\pi}, \tag{3.8}$$

and for the Poisson, $P(\mu)$,

$$h(\mu) = \int \frac{1}{\mu^{1/2}}\, d\mu = \sqrt{\mu}. \tag{3.9}$$

The variance-stabilizing transformation for the gamma distribution $G(\alpha, k)$, where $E(y) = \alpha k = \mu$, $Var(y) = \alpha k^2 = k\mu$ yields the following asymptotic mean and variance

$$E\left(\frac{y}{k}\right)^{1/2} = \left(\frac{\mu}{k}\right)^{1/2} = \alpha^{1/2}, \quad Var\left(\frac{y}{k}\right)^{1/2} = \frac{1}{2}k\alpha^{1/2} \tag{3.10}$$

Table (3.1) summarizes the Anscombe residuals with a $O(m^{-1/2})$ correction added to $t[E_\theta(y)]$ and the variance-stabilizing residuals (see [21]):

Table 3.1: *Anscombe and Variance-Stabilizing Residuals Expressed for the Binomial, Poisson and Gamma distributions*

	ANSCOMBE RESIDUAL	VARIANCE-STABILIZING
Binomial(m, π)	$\frac{t(y/m)-[t(\pi)+(\pi(1-\pi)^{-1/3}(2\pi-1)/6m}{(\pi(1-\pi)^{1/6}/\sqrt{m}}$	$\frac{\sin^{-1}[(y/m)^{1/2}]-\sin^{-1}(\pi^{1/2})}{1/(2\sqrt{m})}$
Poisson(μ)	$\frac{y^{2/3}-(\mu^{2/3}-\mu^{-1/3}/9)}{(2/3)\mu^{1/6}}$	$\frac{y^{1/2}-\mu^{1/2}}{1/2}$
Gamma(α, k)	$\frac{(y/k)^{1/3}-[\alpha^{1/3}-(\alpha^{-2/3})/9]}{(\alpha^{-1/6})/3}$	$\frac{(y/k)^{1/2}-\alpha^{1/2}}{(1/2k\alpha^{1/2})^{1/2}}$

(3) Deviance residual

$$R_D(y, \theta) = sgn(\hat{\theta} - \theta)\{2[\ell(\hat{\theta}, y) - \ell(\theta, y)]\}^{1/2}. \tag{3.11}$$

$\hat{\theta}$ is the MLE of θ based on y without restriction by model $\theta_i = g(\mathbf{x}_i'\beta)$. The deviance residual will measure the discrepancy between the maximized log-likelihood for the current model and the maximum possible log-likelihood for the data. Under a first-order m-asymptotic, the deviance has an approximate normal standard distribution. An adjustment to the deviance residual will remove the bias coming from the asymptotic term, $O(m^{1/2})$, and the *adjusted deviance* residual is formed, as described next.

(4) Adjusted deviance residual

$$R_{AD}(y;\theta) = R_D(y,\theta) + \rho_3(\theta)/6, \qquad (3.12)$$

$$\rho_3(\theta) = E_\theta[(y-\mu)/SD_\theta(y)]^3,$$

$$\mu = E_\theta(y).$$

The table which follows cites the expressions for deviance residuals and adjusted deviance residuals, for the three given densities.

Table 3.2: *Deviance and Adjusted Deviance Residuals for the Three Distributions*

		ADJUSTMENT TERM TO
	DEVIANCE RESIDUAL	ADJUSTED DEVIANCE, $\frac{1}{6}\rho_3(\theta)$
Binomial$(m.\pi)$	$\{2[y\log(\frac{y}{\hat{\pi}}) - (m-y)\log(\frac{(m-y)}{(m-\hat{\pi})})]\}^{1/2}$	$\frac{(1-2\pi)}{6\{m\pi(1-\pi)\}^{1/2}}$
Poisson(μ)	$\{2[-\log(\frac{y}{\hat{\mu}}) - (y-\hat{\mu})]\}^{1/2}$	$\frac{1}{6\sqrt{\mu}}$
Gamma$(\alpha;\mu/\alpha)$	$\{2\alpha[-\log(\frac{y}{\hat{\mu}}) + \frac{(y-\hat{\mu})}{\hat{\mu}}]\}^{1/2}$	$\frac{1}{3\sqrt{\alpha}}$

Taking the normal approximated tail probabilities, these residuals for different values of y lie between .0001 and .10 for the binomial and Poisson distributions and are equal to .05 and .01 for the gamma. Pierce and Schafer [21] compared the true tail probabilities for each respective density,

$$Pr(Y \leq y) \quad \text{or} \quad Pr(Y \geq y)$$

to approximations

$$\Phi[R(y+.5,\theta)] \quad \text{and} \quad 1-\Phi[R(y-.5,\theta)],$$

by considering different residuals R, where y is an integer. In all three density functions, Pierce and Schafer found that the Anscombe residual and the adjusted deviance residual are good for approximate normality, even when m is small. Furthermore, the adjusted deviance residual should be consistently the closest to the true tail probability throughout, for the different distributions due to its almost-normal characteristic.

3.2 Influential Observations

Deletion or perturbation of observations from a given model helps detect those individual observations which may exert influence on the various components of the fitted model. The following approach is described in Pregibon [20]. To see the effect of perturbing an individual observation is to see the effect of its deletion. Pregibon pursues this idea by considering the likelihood

$$\ell_v(\beta; y) = \sum_{i=1}^{n} v_i \ell(\beta; y_i), \tag{3.13}$$

where considering $v_i = 1, \forall i$ yields the usual likelihood, whereas $v_i = 1 \forall i$ except $i = \ell$ amounts to deleting the ℓth observation. Thus, a matrix composed of diagonal elements v_i may be defined by

$$v_i = \begin{cases} v & i = \ell, \\ 1 & \text{otherwise} \end{cases}$$

for $0 \leq v \leq 1$.

Then the likelihood estimate $\hat{\beta}$ becomes a function of $\mathbf{V}$ and is denoted by $\hat{\beta}(v)$. The likelihood equations are

$$\mathbf{X'V(y - \hat{\mu}) = 0}. \tag{3.14}$$

Then Fisher's scoring algorithm for the modified likelihood leads to a new sequence of estimates:

$$\beta^{i+1}(v) = \beta^{i}(v) + (\mathbf{X}'\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}(\mathbf{y} - \hat{\mu}). \tag{3.15}$$

$\hat{\beta} = \hat{\beta}(1)$ is the maximum likelihood estimate from *IWLS*. An alternative to considering the maximum likelihood $\hat{\beta}(v)$ is to start from the usual maximum likelihood estimate $\hat{\beta} = \hat{\beta}(1)$ obtained through *IWLS* and to finish this sequence after one additional step:

$$\hat{\beta}^{1}(v) = (\mathbf{X}'\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}\mathbf{z}. \tag{3.16}$$

As $v \to 0$, the *ℓth* point has less leverage in the fit. The ℓth point is influential if a small value for v yields a large $\hat{\beta}^{1}(v)$:

$$\hat{\beta}^{1}(v) = \hat{\beta} - \frac{(\mathbf{X}'\mathbf{W}\mathbf{V})^{-1}\mathbf{x}_{\ell}(y_{\ell} - \hat{\mu}_{\ell})(1 - v)}{(1 - (1 - v)h_{\ell\ell})} \tag{3.17}$$

$$\Rightarrow \qquad \frac{\partial}{\partial v}\hat{\beta}^{1}(v) = \hat{\beta}'^{1}(v)$$

$$= \frac{(\mathbf{X}^{T}\mathbf{W}\mathbf{V})^{-1}\mathbf{x}_{\ell}(y_{\ell} - \hat{\mu}_{\ell})}{(1 - (1 - v)h_{\ell\ell})^{2}} \tag{3.18}$$

measures the impact that an ℓth observation exerts on the vector of coefficients in a GLM regression. Plotting the standardized change in coefficients $\Delta_{\ell}\hat{\beta}^{1}_{j}/s.e.(\hat{\beta}_{j})$ against ℓ detects any influential observations in the selected coefficient, $\hat{\beta}_{j}$.

Cook's statistic c_{ℓ}, measures the impact of an observation on all the coefficients $\hat{\beta}$. One convenient way of interpreting c_{ℓ} in a GLM context is by the confidence region displacement for β due to deleting an ℓth observation, namely,

$$c_{\ell} = -2\{\ell(\mathbf{X}\hat{\beta};\mathbf{y}) - \ell(\mathbf{X}\hat{\beta}_{(\ell)};\mathbf{y})\} \tag{3.19}$$

$$\Rightarrow \qquad c_{\ell} = (\hat{\beta} - \hat{\beta}_{(\ell)})'\mathbf{X}'\mathbf{W}\mathbf{X}(\hat{\beta} - \hat{\beta}_{(\ell)})$$

A large c_ℓ corresponds to a highly influential ℓth observation on the overall fit of the model. By applying a second-order Taylor series expansion to (3.19), the confidence region is generated by the limiting Normal distribution of $\hat{\beta}$.

The concept of observation deletions can be extended to perturbations by letting $v_\ell = 0$ so that $\beta = \hat{\beta}(0)$ measures the influence that the ℓth point exerts on the coefficient estimates $\hat{\beta}$ through c_ℓ. Then the confidence interval displacement is measured by the one-step approximation to $\hat{\beta}(0)$:

$$c^1_\ell = \frac{\chi^2_\ell h_{\ell\ell}}{(1 - h_{\ell\ell})^2} \tag{3.20}$$

where $\chi^2_\ell = r^2_{P_\ell}$ (2.43).

3.3 Testing the Goodness-of-Fit

Measuring the goodness-of-fit of a model can be done by calculating the effect of change in v on the diagnostic measures of the deviance function D and Pearson's statistic χ^2. In case of the deviance function, the maximum likelihood estimate should minimize D, much like the least squares estimate minimizes the residual sum of squares RSS in a normal-theory linear model. Subsequently, deletion of an observation decreases D , like it would decrease RSS in the normal-theory model.

Using the observation count matrix $\mathbf{V}$ in the loglikelihood function yields a deviance function expressed by:

$$D_v(\mathbf{X}\hat{\beta}(v); \mathbf{y}) = 2\sum_i^n v_i[\ell(\hat{\theta}_i; y_i) - \ell(x'_i\hat{\beta}(v); y_i)]. \tag{3.21}$$

A one-step estimate $\hat{\beta}^1(v)$, and a second-order Taylor series expansion of $D_v(\mathbf{X}\hat{\beta}^1(v); \mathbf{y})$

about $\hat{\beta}$ approximates the above quantity :

$$D_v(\mathbf{X}\hat{\beta}^1(v);\mathbf{y}) \doteq D(\mathbf{X}\hat{\beta};\mathbf{y}) - \left[(1-v)d_\ell^2 + \frac{\chi^2(1-v)^2 h_{\ell\ell}}{[1-(1-v)h_{\ell\ell}]}\right]. \qquad (3.22)$$

At $v = 1$: $D_v(\mathbf{X}\hat{\beta}^1(v);\mathbf{y}) = D(\mathbf{X}\hat{\beta};\mathbf{y})$ (maximum),

at $v = 0$: $D_v(\mathbf{X}\hat{\beta}^1(v);\mathbf{y})$ is at a minimum of $D(\mathbf{X}\hat{\beta};\mathbf{y}) - (d_\ell^2 + \bar{c}_\ell^1)$, where $\bar{c}_\ell^1$ is the change in the confidence interval displacement diagnostic $\bar{c}_\ell^1$.

The deviance decreases as $v \to 0$.

The rate of change of D due to perturbations is obtained by taking the derivative of (3.22) with respect to v.

The change in deviance due to deletion of the ℓth point is approximated by:

$$\begin{aligned} \Delta_\ell D &= D_1(\mathbf{X}\hat{\beta};\mathbf{y}) - D_0(\mathbf{X}\hat{\beta}^1(0);\mathbf{y}) \\ &\doteq d_\ell^2 + \frac{\chi_\ell^2 h_{\ell\ell}}{1-h_{\ell\ell}} \qquad (3.23) \end{aligned}$$

Since the individual components $\Delta_\ell D$ are asymptotically χ_1^2, then each χ_ℓ can be replaced by d_ℓ to get the approximately normal studentized residuals

$$d_\ell/\sqrt{m_{\ell\ell}} \qquad (3.24)$$

which are useful for index plotting. The presence of χ^2 components is a feature found in the one-step approximation, making it a useful diagnostic tool.

The Pearson's statistic is not a straight-forward measure to interpret since it doesn't extend from the normal-theory linear model as does the deviance function. As observations are deleted from a given model, the χ^2 measure does not necessarily decrease. However, like the RSS, the χ^2 is the result of the sum of squares of differences of the observed from the fitted values. The one-step approximation to the χ^2 due to the deletion of the ℓth observation is:

$$\begin{aligned} \Delta_\ell \chi^2 &= \chi^2 - \chi^2_{(\ell)} \\ &\doteq \frac{\chi^2_\ell}{m_{\ell\ell}}. \end{aligned} \tag{3.25}$$

In extreme cases, χ^2 will increase for some observation deletions.

The deviance function and Pearson's χ^2 goodness-of-fit statistics can be interpreted in two ways:

(1) when the ℓth point is not well fit by a given model, i.e. an outlier, then a model perturbation caused by v will be reflected in the single components of D and χ^2: d_ℓ and χ^2_ℓ respectively;

(2) when the ℓth point is an extreme point in the design matrix, i.e. an influential point, then all the individual components of D and χ^2 will change.

A change in either the deviance function or the Pearson's statistic can't distinguish whether the change comes from (1) or (2). An addtional diagnostic measure $h_{\ell j}$ can resolve this problem, where $h_{\ell j}$ is an off-diagonal element in the hat matrix **H** for the ℓth observation with respect to the jth observation, $|h_{\ell j}| \leq \sqrt{h_{\ell\ell}}\sqrt{h_{jj}}$. The $h_{\ell j}$'s in combination with the χ_ℓ and χ_j are useful for measuring how an ℓth point is influential on the remaining $(n-1)$ points.

There are other ways of measuring the goodness-of-fit such as by investigating the interactions between covariates, or by looking for non-linear effects by adding some terms to a model in the hopes of reducing the approximated deviance.

3.4 Testing Goodness-of-Link Functions

Once a model has been tested for potential outliers and influential observations and that they've been removed from the data, then the validity of the link function needs to be checked. Consider a generalized linear model to be fitted with a hypothesized link function $g_o(\mu)$ generated from a class of functions, of which the true and unknown link function $g_*(\mu)$ is also a member. All link functions belonging to a class of functions are indexed by one or more unknown parameters. Plotting for a range of fixed parameter values against the corresponding deviances is useful in deciding which range of parameter values are most consistent with the data. The adequacy of the hypothesized link function is examined by expanding and linearizing the link to optimize over the range of parameters. The deviances obtained from fixed parameter values are tested against best-fitting values. This is called the *goodness-of-link test.* If a class of link functions is generated by the the power family for one parameter λ, then it is defined either by

$$g(\mu;\lambda) = \frac{\mu^\lambda - 1}{\lambda}$$

with limiting value $g(\mu;\lambda) = \log\mu$ as $\lambda \to 0$
or by

$$g(\mu;\lambda) = \begin{cases} \mu^\lambda, & \text{if } \lambda \neq 0, \\ \log\mu, & \text{if } \lambda = 0. \end{cases}$$

The power family transforms the fitted values μ in a GLM case. Conversely, the Box-Cox transformation is a power function which transforms the data in a normal linear model.

If a model is fitted with a link function $g_o(\mu)$ when the true link is $g_*(\mu)$, then this can be represented by:

$$\begin{aligned}\text{Hypothesized link} &: g_o(\mu) = g(\mu; \lambda_o) \rightarrow D_o; \chi^2_o \\ \text{True link} &: g_*(\mu) = g(\mu; \lambda_*) \rightarrow D_*; \chi^2_*\end{aligned}$$

To optimize over λ_*, one approach is to linearize the power family through a first-order Taylor series expansion about $g_o(\mu)$. Based on the approximate relationship

$$g_*(\mu) \doteq g_o(\mu) + (\lambda_* - \lambda_o) g'_\lambda(\mu; \lambda_o), \tag{3.26}$$

the true link $g_*(\mu) = \mathbf{X}\beta$ is approximated by

$$g_o(\mu) = \mathbf{X}\beta_{p\times 1} + \mathbf{z}'\gamma_{q\times 1} \tag{3.27}$$

where $\mathbf{z}' = (g'_\lambda(\mu; \lambda_o))$ and $\gamma' = (-\lambda_* + \lambda_o)$.

The hypothesized link function is now modified by the addition of a covariate $\mathbf{z}'$ to the design matrix and its parameter estimate $\hat{\gamma}$ yields a first-order adjustment to λ_o. Hence the additional factor in the systematic linear component accounts for local differences between the hypothesized link and the modified one. These differences are measured by a reduction in the deviance. In turn, this reduction serves to test whether λ_o is suitable enough for λ_*:

$$\frac{D_o - D_*}{\hat{\phi}} \approx \frac{\chi^2_o - \chi^2_*}{\hat{\phi}} \sim \chi^2_p \qquad \text{approximately,} \tag{3.28}$$

$\hat{\phi} = D_*/(n-p-q)$ or $\chi^2_*/(n-p-q)$.

When $g_o(\mu)$ is assumed to have the identity link (i.e. the data is normally distributed), then the approximations made on the χ^2 distribution are exact:

$$\frac{D_o - D_*}{\hat{\phi}}/p \equiv \frac{SSE_o - SSE_*}{\hat{\sigma}^2}/p \sim F_{n-p-q}. \tag{3.29}$$

The process is repeated to form a new adjusted value for λ_* at each iteration until a possible convergence is reached and then the maximum likelihood estimate of λ_* is obtained. If the initial λ_o is sufficiently close to $\hat{\lambda}_*$, convergence is assured. Then the linearization of the power family will yield the true maximum likelihood estimate. The process follows a sequence

$$g(\mu; \lambda_{i-1}) = g(\mu; \lambda_i) + (\lambda_{i-1} - \lambda_i) g'_i, \quad i \geq 0, \tag{3.30}$$

which is implemented in the iterations for fitting a generalized linear model.

The link modification method has its limits such that it is restricted to a specified class of link functions g. The most which can be done is to improve an already reasonable fit in order to obtain the true link function. On the other hand, if the hypothesized link is inadequate, then the true link function belongs to another class of link functions altogether. This is attributable to a misspecfication of the systematic component of the model.

Consider a model initially fitted with link $g_o(\mu) = \mathbf{X}\beta$ to get estimates $\hat{\beta}$ and fitted values $\hat{\theta} = \mathbf{X}\hat{\beta}$. Thus $\hat{\mathbf{z}} = (g'_\lambda(\hat{\mu}; \lambda_o))$ can be obtained, and the model is refitted with the extended design matrix now including the covariates $\hat{\mathbf{z}} = \hat{\mathbf{z}}_\lambda$. In turn,

$$\hat{\gamma}' \quad = \quad (-\lambda_* + \lambda_o) \quad = \quad (\hat{\mathbf{z}}'\mathbf{W}\hat{\mathbf{z}})^{-1}\hat{\mathbf{z}}'\mathbf{W}(\mathbf{y} - \hat{\mu}). \tag{3.31}$$

The sum of squares corresponding to $\hat{\gamma}$ (to test if $\gamma = 0$) is

$$\hat{\gamma}'\hat{\mathbf{z}}'\mathbf{W}\hat{\mathbf{z}}\hat{\gamma} \quad \sim \quad \chi^2_1. \tag{3.32}$$

A parallel reduction in the degrees of freedom and in the deviance from the initial model to the extended one including $\hat{\mathbf{z}}$ is produced. This reduction is evaluated by an F-test to decide for the validity of the hypothesized link function.

For every parameter added to the power function, an extra covariate is added to the design matrix which is given by $-\frac{\partial g}{\partial \lambda} |_{\lambda=\lambda_o}$. The power family provides link generalizations for the normal distribution with identity link, for the Poisson with log link, for the gamma with reciprocal link and for the inverse gaussian with μ^{-2} link. For log-linear data, the power family is defined by the one-link parameter function

$$g(\mu;\lambda) = \frac{\mu^\lambda - 1}{\lambda}. \tag{3.33}$$

The log link is generated by the limit

$$\lim_{\lambda \to 0} g(\mu;\lambda) = \log(\mu) \tag{3.34}$$

As for binomial data, the power family does not apply. Another one-parameter link family is given instead by

$$g(\mu;\lambda) = \log\left[\left(\left(\frac{m}{m-\mu}\right)^\lambda - 1\right)/\lambda\right]. \tag{3.35}$$

It will generate the logit link at $\lambda = 1$:

$$\begin{aligned} g(\mu;1) &= \log\left[\left(\frac{m}{m-\mu}\right) - 1\right] \\ &= \log\left[\left(\frac{1}{1-\mu/m}\right) - 1\right] \\ &= \log\left[\frac{1-(1-\mu/m)}{1-\mu/m}\right] \\ &= \log\left(\frac{\mu/m}{1-\mu/m}\right) \\ &= \text{logit }(\mu/m). \end{aligned} \tag{3.36}$$

As $\lambda \to 0$, the complementary log-log link is generated:

$$\begin{aligned}\lim_{\lambda\to 0} g(\mu; \lambda) &= \log\left[\log\left(\frac{m}{m-\mu}\right)\right]\\ &= \log\left[\log\left(\frac{1}{1-\mu/m}\right)\right]\\ &= \log\left[-\log(1-\mu/m)\right]. \qquad (3.37)\end{aligned}$$

Another family of link functions applied to tolerance distributions (see Pregibon [19]) is given by:

$$g(\mu;\ \alpha, \delta) = \frac{(\pi)^{\alpha-\delta}-1}{\alpha-\delta} - \frac{(1-\pi)^{\alpha-\delta}}{\alpha+\delta}, \qquad (3.38)$$

$\pi = \mu/m$ is the responding proportion. It is a two-parameter link family with parameters α and δ (based on tolerance distributions). This family of functions generates the logit link as the limiting form of g:

$$\begin{aligned}\lim_{\alpha,\delta\to 0} g(\mu;\ \alpha, \delta) &= \log(\mu/m) - \log(1-\mu/m)\\ &= \log\left(\frac{\mu/m}{1-\mu/m}\right)\\ &= \text{logit}\ (\mu/m). \qquad (3.39)\end{aligned}$$

For this model, the series expansion is

$$\begin{aligned}g_*(\mu) = g_o(\mu) &+ \alpha_*\left(\frac{1}{2}(\log^2(\mu/m) - \log^2(1-\mu/m))\right)\\ &+ \delta_*\left(-\frac{1}{2}(\log^2(\mu/m) + \log^2(1-\mu/m))\right). \qquad (3.40)\end{aligned}$$

A fit using the logit link will give estimates $\hat{\beta}$ and fitted values $\hat{\mu} = \mathbf{X}\hat{\beta}$ from which $\hat{\mathbf{z}} = \mathbf{z}\mid_{\beta=\hat{\beta}}$ is obtained:

$$\hat{\mathbf{z}}' = \left\{\frac{1}{2}(\log^2(\hat{\mu}/m) - \log^2(1-\hat{\mu}/m)),\ -\frac{1}{2}(\log^2(\hat{\mu}/m) + \log^2(1-\hat{\mu}/m))\right\}. \qquad (3.41)$$

This implies

$$\hat{\gamma}' = -(\hat{\alpha}_*, \hat{\delta}_*). \tag{3.42}$$

The true link function is approximated by the extended model

$$g_o(\mu) = \underbrace{\mathbf{X}\beta}_{\text{hypothesized link}} + \underbrace{\mathbf{z}'\gamma}_{\text{additional factor}} . \tag{3.43}$$

The maximum likelihood estimate of γ is reached through the iterative process described earlier. A reduction in deviance results from adding on the additional factor to the systematic linear component. Finally, an F-test uses the change in deviance to assess whether the estimate of γ via (α_*, δ_*), hence of the link function itself, is adequate.

3.5 Software Applications

The software application GLIM ("Generalized Linear Interactive Modelling") was created in the early 1970's for generalized linear model computations, but because one had to have some in-depth knowledge of statistics to use this tool, the generalized linear models were not popularized. It took twenty years for generalized linear modelling procedures to become accessible to everyone through user-"friendly" software applications. In SAS, GLMs can be fitted through the Genmod procedure, and the GEE macro analyzes longitudinal data by using the Generalized Estimation Equation approach. In S-Plus, the StatMod library contains some functions for GLM statistical modelling. R, which is a non-commercial equivalent to S-Plus, can fit GLMs. It shares some libraries with S-Plus which are accessible from the website

- http://www.ci.tuwien.ac.at/R/mirrors.html .

LispStat is useful for GLMs and uses some R coding. Matlab uses a module called *glmlab* to fit GLMs. Another application is Genstat which is much like GLIM. Some websites offer articles and abstracts on GLMs. The following are only a few websites worth consulting for a start:

- http://www.ams.org/mathscinet/ and
- http://www.maths.uq.edu.au/ gks/research/glm/articles.html.

Chapter 4

Numerical Examples

4.1 Introduction

In this chapter, three sets of data are used for illustration of the techniques presented earlier for generalized linear models. The first set of data is assumed to come from the binomial family, the second one from the Poisson family and the third one from the gamma family. In each case, maximum likelihood fit of the model is provided along with the residual diagnostics. The parameter estimates were obtained through some computer programs created in S-Plus. These programs are provided in Appendix A: see A.1 for binomial data, A.2 for Poisson, and A.3 for gamma data.

4.2 Binomial Data

A study of a herbicide effect on the proportion of birth abnormalities was conducted over a time span of one year (see Aitken, Anderson, and Francis, 1989, "Statistical Modelling in GLIM"). The data was collected on a monthly basis. The birth abnormality proportions are determined by dividing the observed number of birth abnormalities by the total number of births for a given month.

Table 4.1: *Number of birth abnormalities out of total births per month for herbicide effect*

MONTH	ABNORM.	TOTAL	HERB	MONTH	ABNORM.	TOTAL	HERB
Jan.	10	222	0	July	20	208	788
Feb.	17	221	0	Aug.	17	219	0
Mar.	18	188	0	Sep.	9	198	304
Apr.	11	183	0	Oct.	15	216	560
May	16	197	1454	Nov.	16	244	0
June	24	218	3280	Dec.	15	218	0

Based on the assumption that the data is binomially distributed and that the logit link is used to fit this model, a combination of graphical and analytical techniques are used to test for any high-leverage or outlying observations. The maximum likelihood estimates for this logistic regression model are calculated using an S-Plus computer program that was created for this purpose. Other pertinent statistics(z- adjusted dependent variate, fitted values, variance) are also calculated in an iterative fashion through the S-Plus linear model function(see **lm**, A.1). The output is presented in the following page in table format. Testing the goodness-of-fit for the current logit model with one explanatory variable accounting for birth abnormalities, the test statistic (2.37) gives $D^* = 8.31 < 18.3 = \chi^2_{0.95,df=10}$ which implies that this logit model is well fitted by the binomially distributed data at a 5% level of significance. Further, a one-step function based on Pregibon's work [20] which modifies the loglikelihood function was also developed in S-Plus to determine the effect that each observation exerts on the regression coefficients through model perturbations to the extent of case deletions. I called this function '**w4onestep**'(see Appendix B). A small change in coefficients for ℓth observation means that the observation is non-influential in the model fit.

Model	LOGIT	
Parameters	β_0	β_1
Estimates	-2.620648	0.0001629353

Data	Fitted Values	Adjusted Dependent Variable	Variance
10	15.05633	-2.980909	14.03519
17	14.98851	-2.476682	13.97197
18	12.75041	-2.178973	11.88566
11	12.41130	-2.742632	11.56955
16	16.63094	-2.425176	15.22694
24	24.07666	-2.089799	21.41755
20	15.89181	-2.212360	14.67763
17	14.85287	-2.465570	13.84553
9	14.06209	-2.958618	13.06339
15	15.94563	-2.593435	14.76849
16	16.54840	-2.656198	15.42607
15	14.78505	-2.605052	13.78231

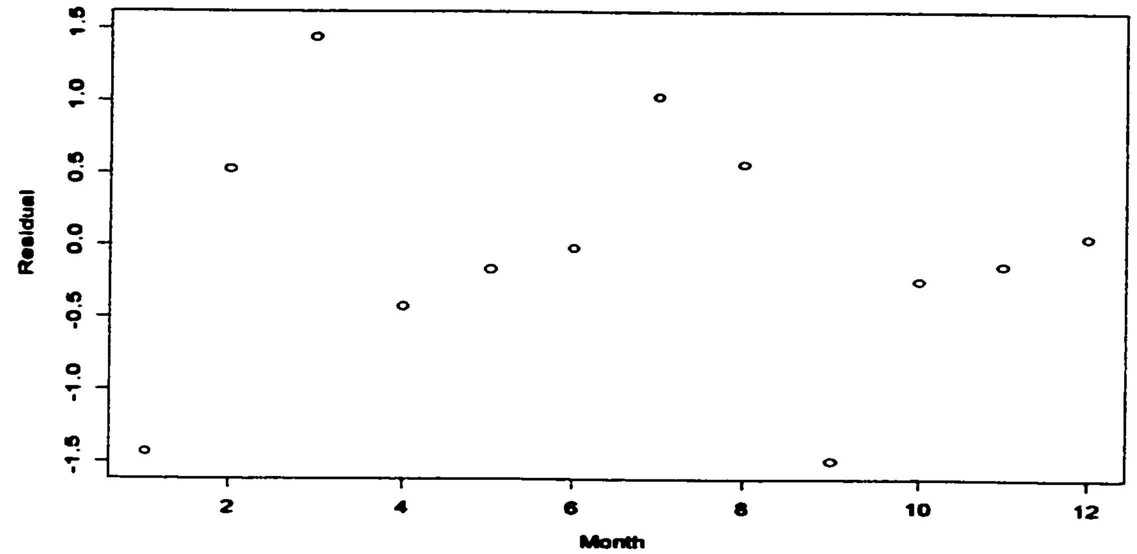

Figure 4.1: Deviance residuals for birth abnormalities due to herbicide spray exposure

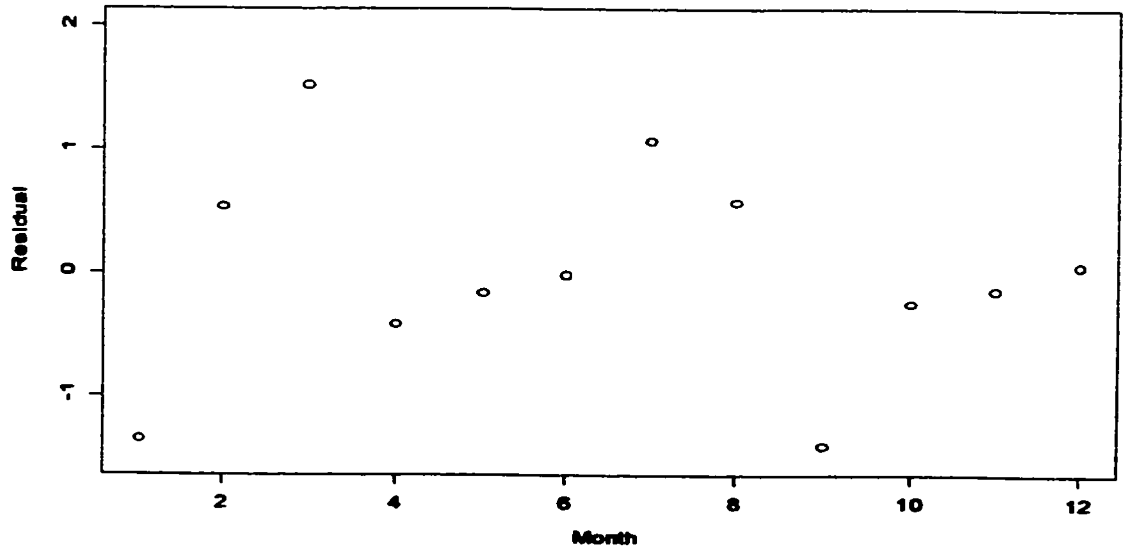

Figure 4.2: χ residuals for birth abnormalities due to herbicide spray exposure

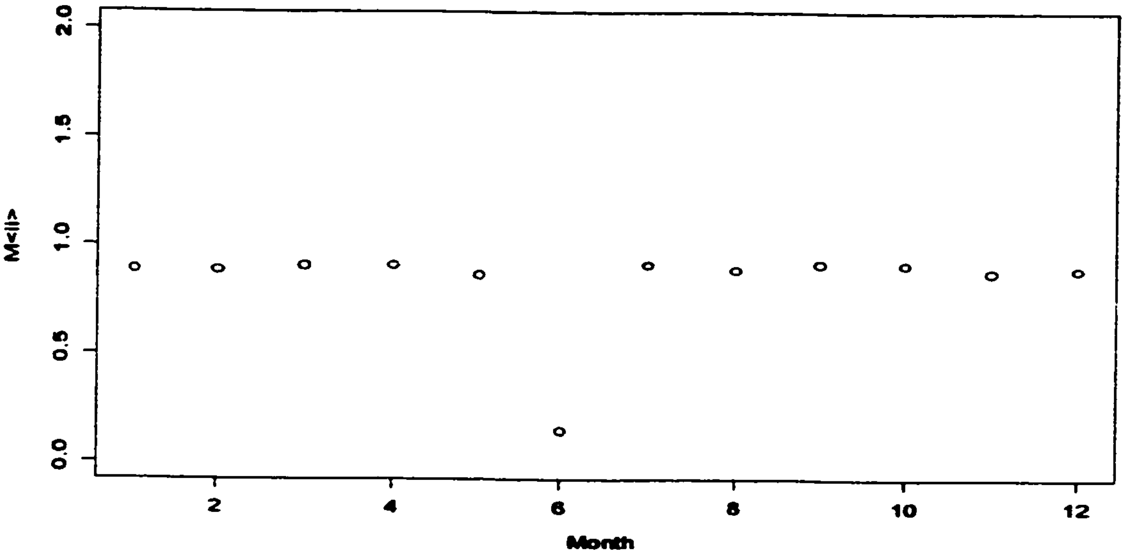

Figure 4.3: Projection matrix diagonal elements for birth abnormalities due to herbicide spray exposure

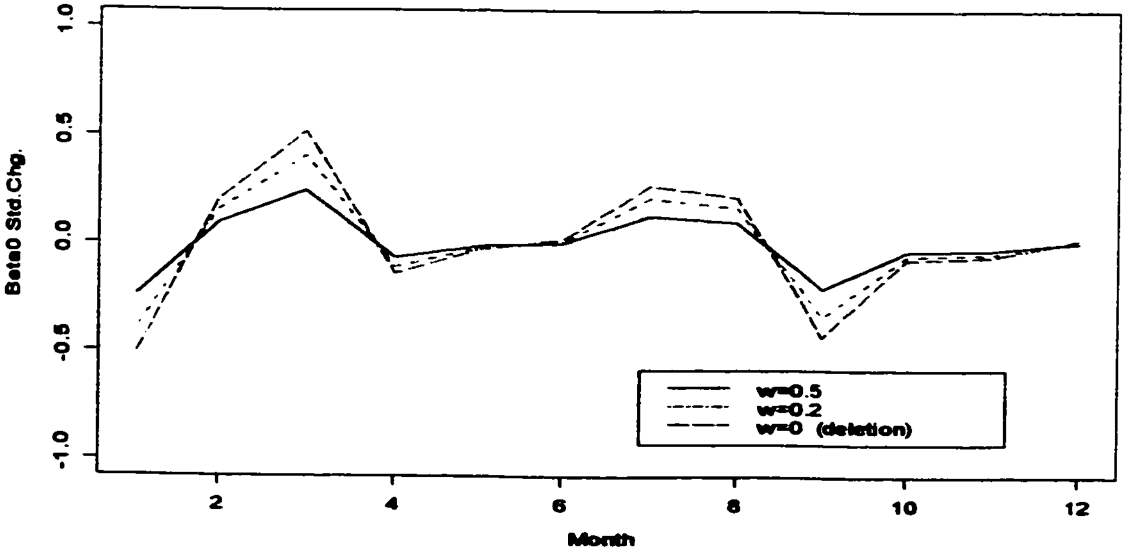

Figure 4.4: Standardized change in $\hat{\beta}_0$ for birth abnormalities due to herbicide spray exposure

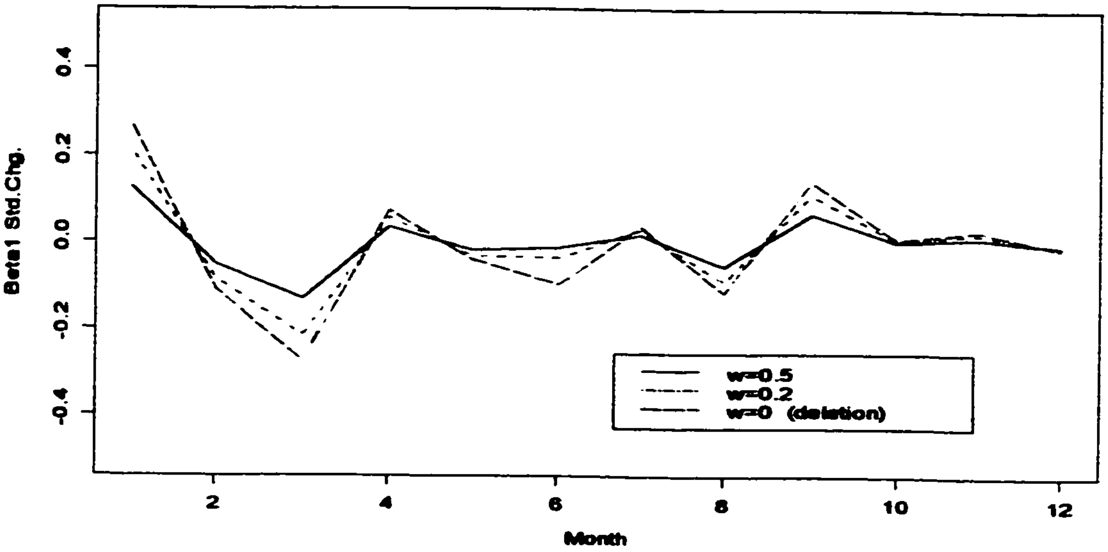

Figure 4.5: Standardized change in $\hat{\beta}_1$ for herbicide data

According to the deviance residual and the χ residual index plots, the month of March would indicate that the herbicide spray effect is significantly greater on birth abnormalities than for any other month of the year. The standardized change plots in both the intercept($\hat{\beta}_0$) and the herbicide spray exposure variable($\hat{\beta}_1$) would also agree that a perturbation or a deletion of the observation for the month of March (i.e.$w = 0.5, 0.2$ or $w = 0$ resp.) would cause a greater standardized change in the regression coefficients than for any other month. Hence, based on these diagnostics, it is likely that the month of March exerts an undue influence on the total number of birth abnormalities.

4.3 Poisson Data

The set of data given here classifies the defects found on furniture from a given manufacturing plant obtained from (see Aitken, Anderson, and Francis, [1]). The defects are thus classified as the type of defect, and the production shift. There were a total of $n = 309$ defects recorded in all, classified in one of four types: $A; B; C; D$. Each piece of furniture is also classified by one of three production shifts: 1,2,3. The contingency table below tabulates these defect counts by type of defect and production shift. The Poisson distribution model is fitted to the data with the log link. The computer program in Appendix A.2 calculates the MLEs for the GLM log-linear regression. The output is summarized in the following tables:

Table 4.2: *Contingency table for furniture defect*

	Type of Defect				
Shift	A	B	C	D	Total
1	15	21	45	13	94
2	26	31	34	5	96
3	33	17	49	20	119

Model	LOG-LINEAR					
Parameters	β_0	β_1	β_2	β_3	β_4	β_5
Estimates	3.114019	-0.06995859	0.5479652	-0.6664789	0.02105341	0.2358287

Data	Fitted Values	Adjusted Dependent Variable	Variance
15	22.51133	2.780350	22.51133
21	20.99029	3.044523	20.99029
45	38.93851	3.817652	38.93851
13	11.55987	2.572120	11.55987
26	22.99029	3.265984	22.99029
31	21.43689	3.511218	21.43689
34	39.76699	3.538018	39.76699
5	11.80583	1.892113	11.80583
33	28.49838	3.507808	28.49838
17	26.57282	2.919640	26.57282
49	49.29450	3.891838	49.29450
20	14.63430	3.050020	14.63430

This model is explained by four levels of defect types and three levels of production shifts. To assess the significance of this log-linear model, the statistics from equations (2.37) and (2.42) are compared to $\chi^2_{0.95,6} = 12.6$. Since $D^* = 20.34$ and $\chi^2 = 19.14$, it is concluded that the log-linear model does not provide a good fit to the Poisson distributed data at a 5% significance level. In fact, the goodness-of fit for this model is only significant at the 1% level. The index plots of the deviance residuals, the χ residuals and the diagonal elements of the projection matrix are based on the fitted log-linear model. Both the 6th and the 8th observations, which correspond to the Type B number of defects and Type D number of defects respectively, found in the second production shift, are not well fit by the model. In fact, the 8th observation has a very large m_{ii} value. The standardized change in coefficient plots for the intercept($\hat{\beta}_0$), the Type B defect variable($\hat{\beta}_1$), and the second production shift variable($\hat{\beta}_4$) agree that the 6th observation is causing instability in these coefficients, while the 8th observation is causing instability more so in the Type D defect variable($\hat{\beta}_3$) and the second production shift variable($\hat{\beta}_4$). Hence, the standardized change in coefficient plots are in-line with the residual and projection matrix index plots.

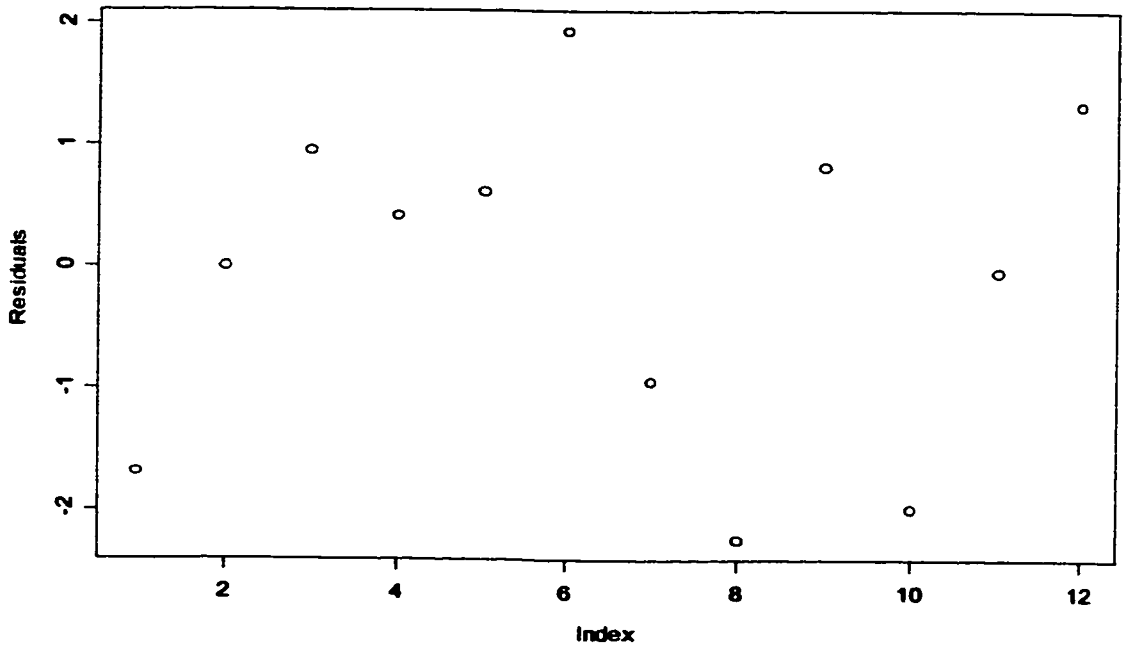

Figure 4.6: Deviance residuals for defects found on furniture produced in a certain manufacturing plant

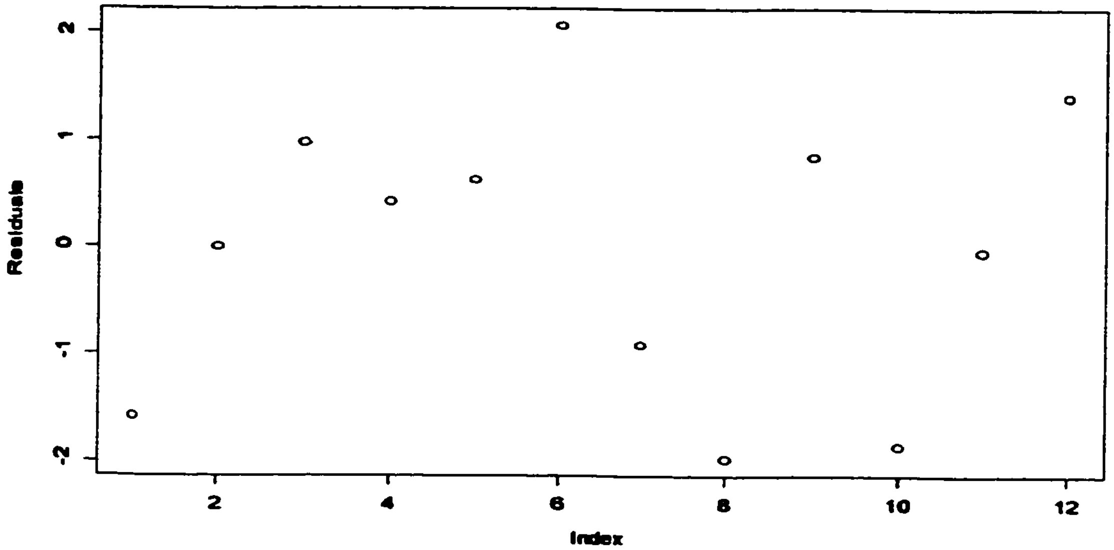

Figure 4.7: χ residuals for defects found on furniture produced in a certain manufacturing plant

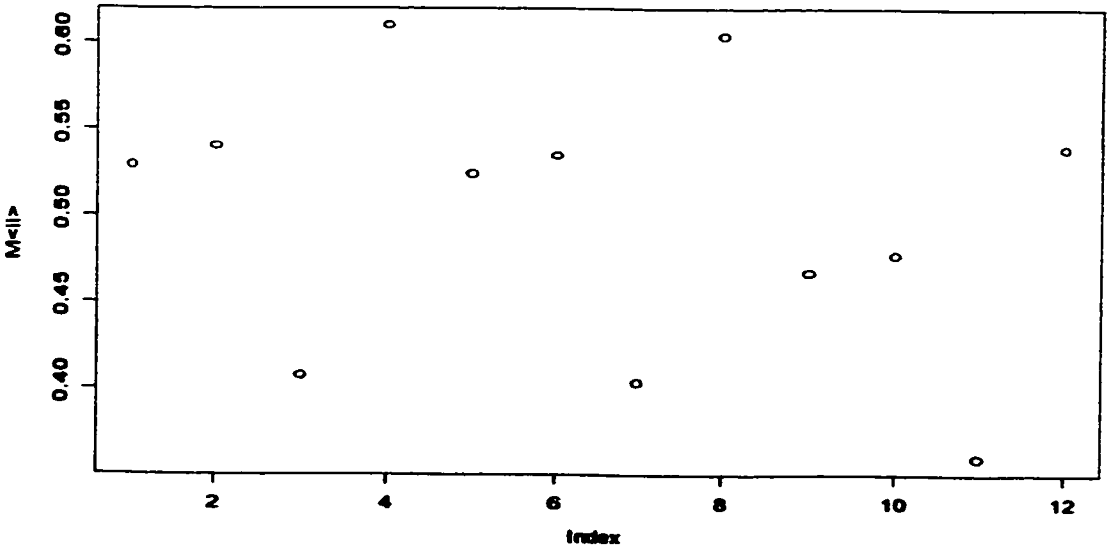

Figure 4.8: Projection matrix diagonal elements for defects found on furniture produced in a certain manufacturing plant

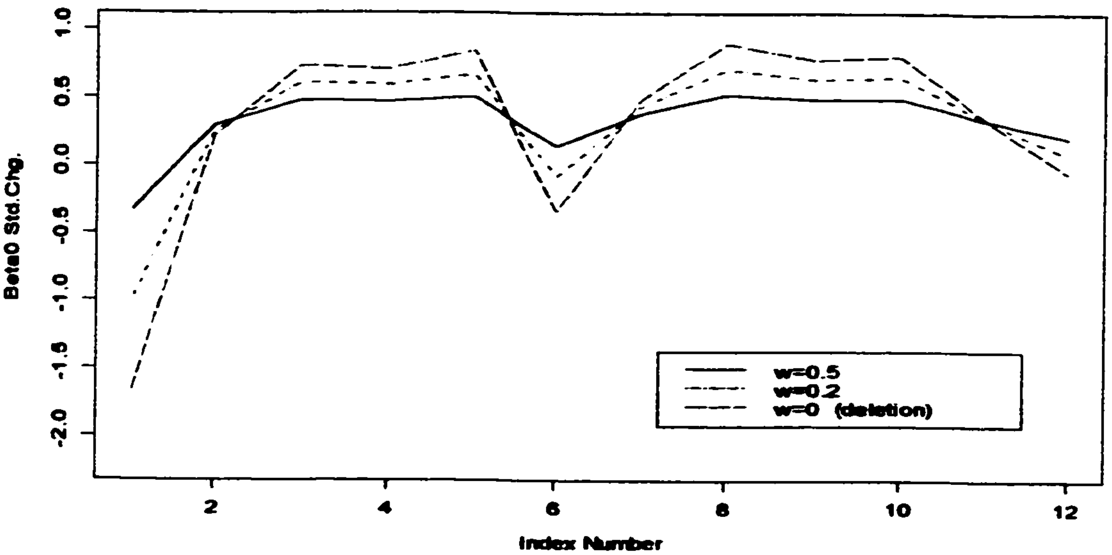

Figure 4.9: Standardized change in $\hat{\beta}_0$ for defects found on furniture produced in a certain manufacturing plant

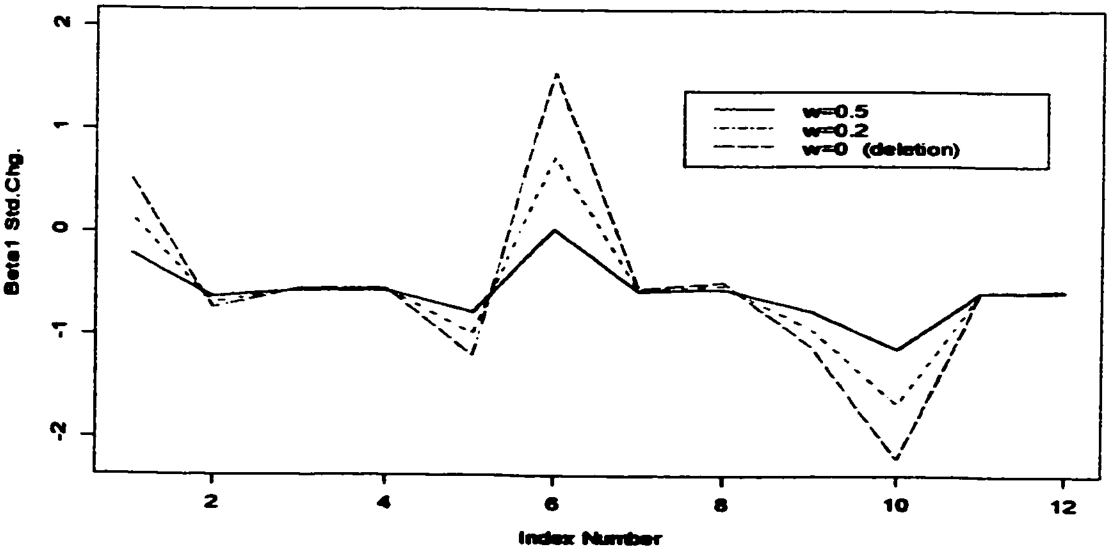

Figure 4.10: Standardized change in $\hat{\beta}_1$ for furniture damage data

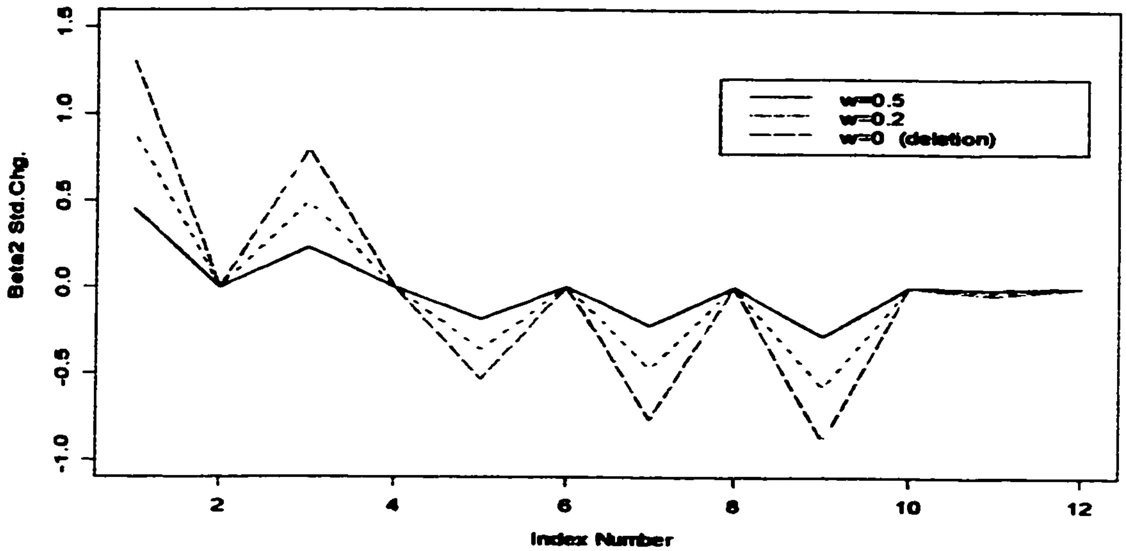

Figure 4.11: Standardized change in $\hat{\beta}_2$ for furniture damage data

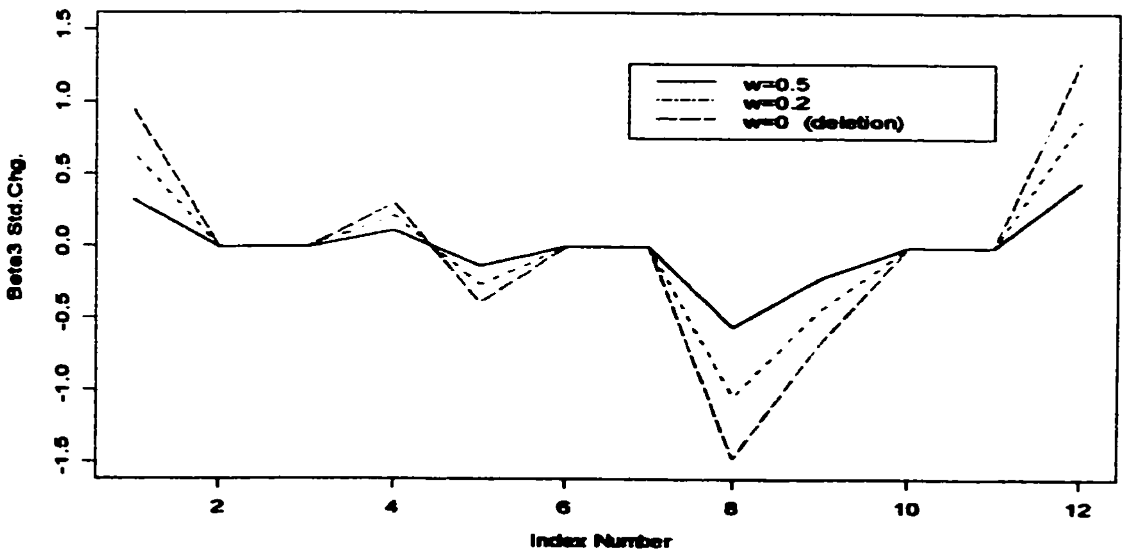

Figure 4.12: Standardized change in $\hat{\beta}_3$ for furniture damage data

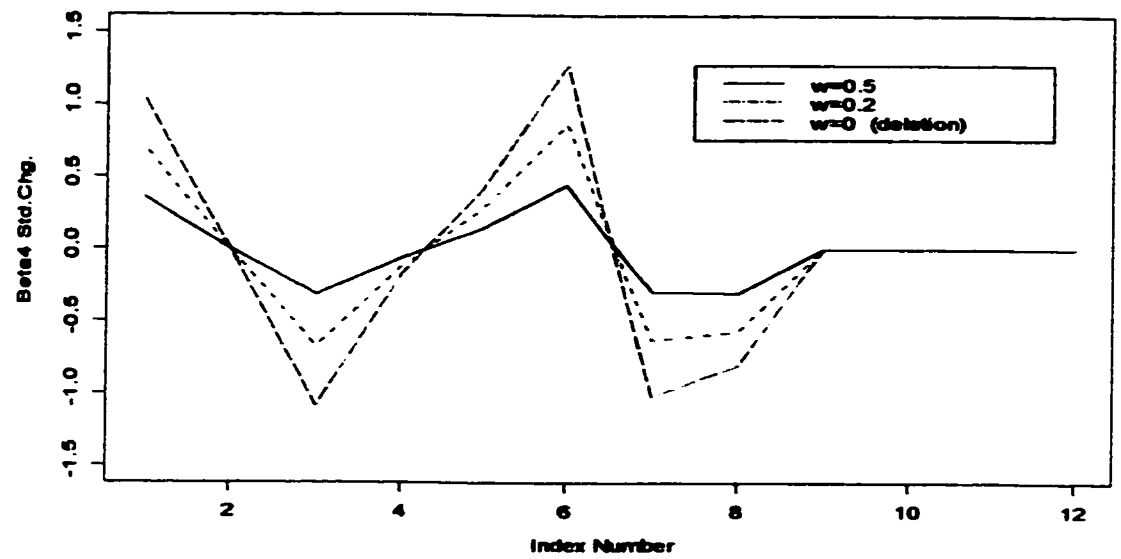

Figure 4.13: Standardized change in $\hat{\beta}_4$ for furniture damage data

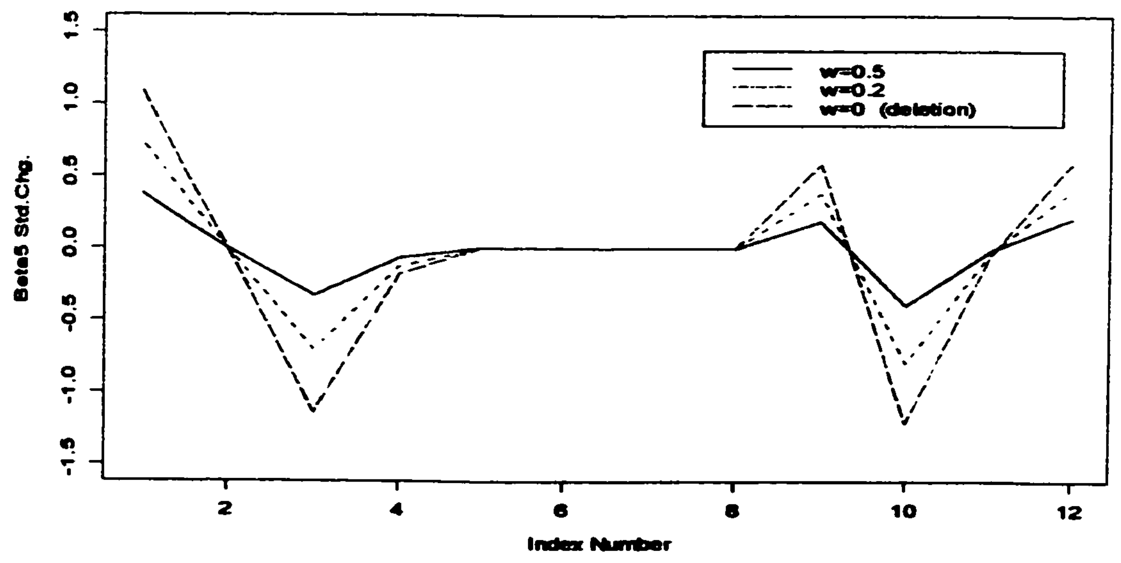

Figure 4.14: Standardized change in $\hat{\beta}_5$ for furniture damage data

4.4 Gamma Data

The next set of data are taken from McCullagh and Nelder, 1989, "Generalized Linear Models", p.300. They describe blood clotting times, in seconds, for normal plasma diluted at nine different percentage concentrations(X) with a prothrombin-free agent. The blood clotting is induced by two lots of thromboplastin. Bliss(1970) fitted a hyperbolic model by using an inverse transformation of the data to the first lot only. Here, the data assumes a gamma distribution with the inverse link applied to each lot separately, since some initial plots indicate that the two intercepts and slopes are different for the two lots. Some of the output from the program in Appendix A.3 is summarized below.

Table 4.3: *Blood clotting times in seconds for 9 percentage concentrations of plasma and for 2 lots*

	% Concentration								
Clotting Time	5	10	15	20	30	40	60	80	100
lot 1	118	58	42	35	27	25	21	19	18
lot 2	69	35	26	21	18	16	13	12	12

Model (lot 1)	inverse		Model (lot 2)	inverse	
Parameters	β_0	β_1	Parameters	β_0	β_1
Estimates	-0.01655439	0.01534312	Estimates	-0.02390848	0.02359922

Data (lot 1)	118	58	42	35	27	25	21	19	18
Fitted Values	122.86	53.26	40.01	34	28.07	24.97	21.61	19.73	18.48
Data (lot 2)	69	35	26	21	18	16	13	12	12
Fitted Values	71.06	32.86	25	21.37	17.74	15.84	13.75	12.58	11.8

If the level of significance is 0.05, then the 95th percentile of the $\chi_7^2 = 14.1$. The value obtained through (2.37) is much less than that: $D^* = 0.017$ for lot 1 and $D^* = 0.013$ for lot 2. Thus, the gamma distributed blood clotting times provides a good model fit for both lots.

In the graphs that follow, some diagnostic tools are used to assess which observations exert some influence on the fitted model for lot 1. The first two index plots agree that observation 2, which is the 10% concentration of the prothrombin-free agent, is not well fitted by the inverse model of the blood clotting times. However, the two standardized change in coefficient plots for the intercept($\hat{\beta}_0$) and the percentage of agent concentration($\hat{\beta}_1$) agree that the 5% concentration level is greatly influential on the model fit, depending on the level of perturbation($w = 0.5, 0.2$) or on a case deletion($w = 0$) altogether.

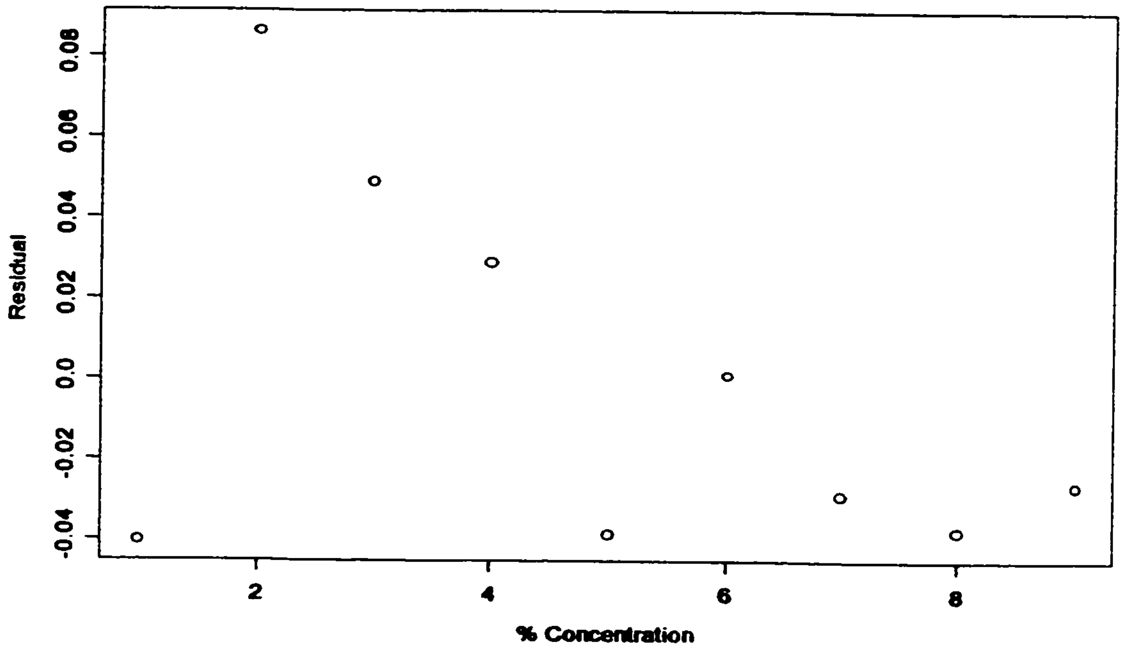

Figure 4.15: Deviance residuals for lot1 of bloodclot time

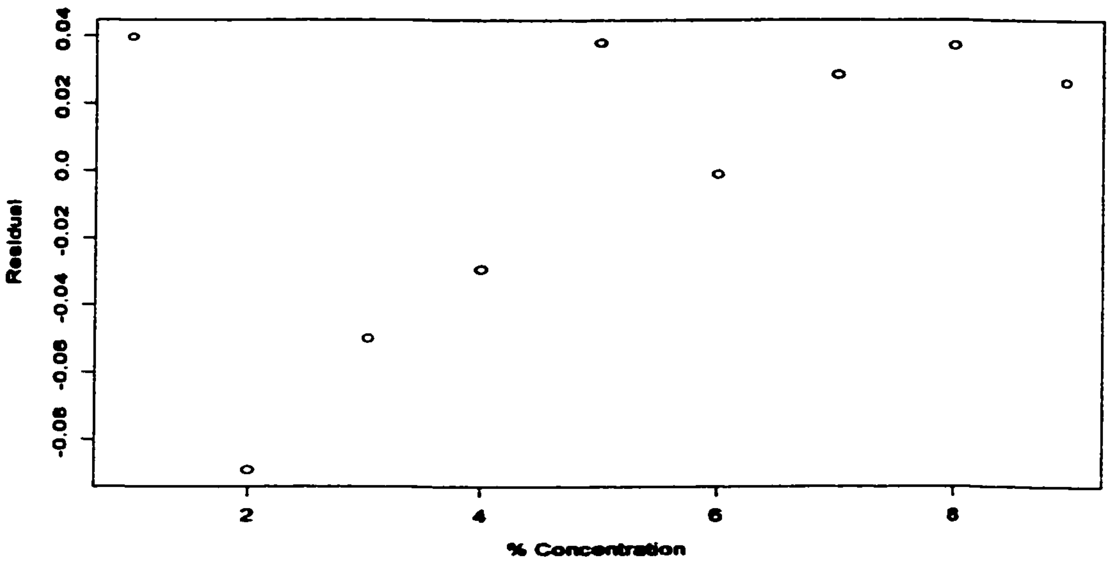

Figure 4.16: χ residuals for lot1 of bloodclot time

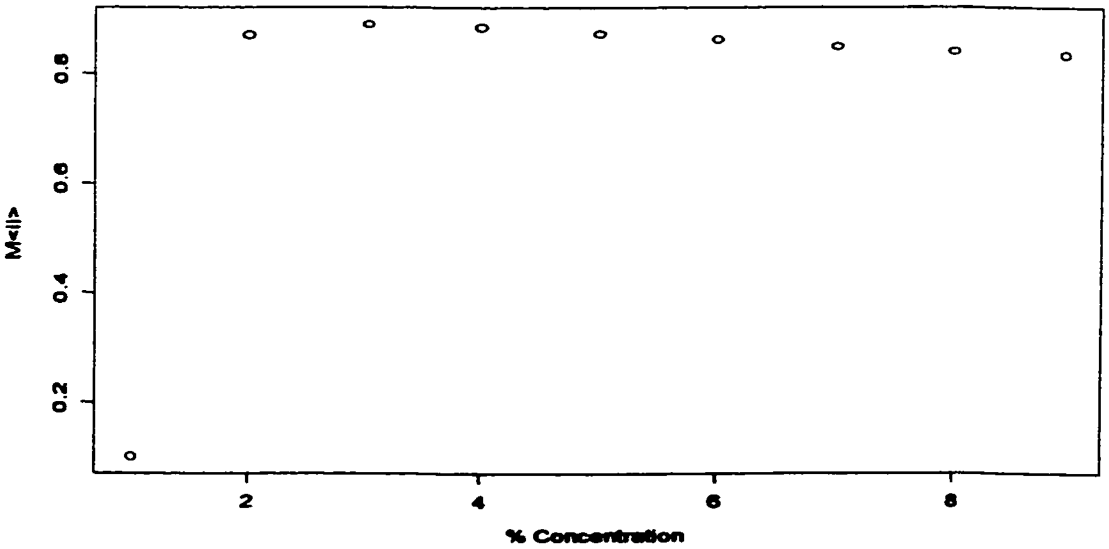

Figure 4.17: Projection matrix diagonal elements for lot1 of bloodclot time

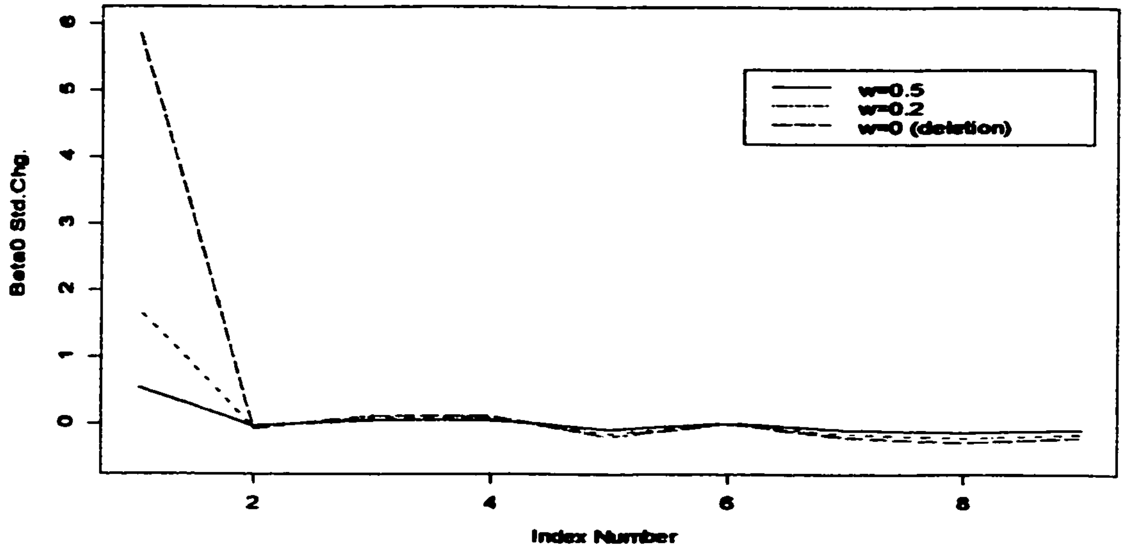

Figure 4.18: Standardized change in $\hat{\beta}_0$ for lot1 of bloodclot time

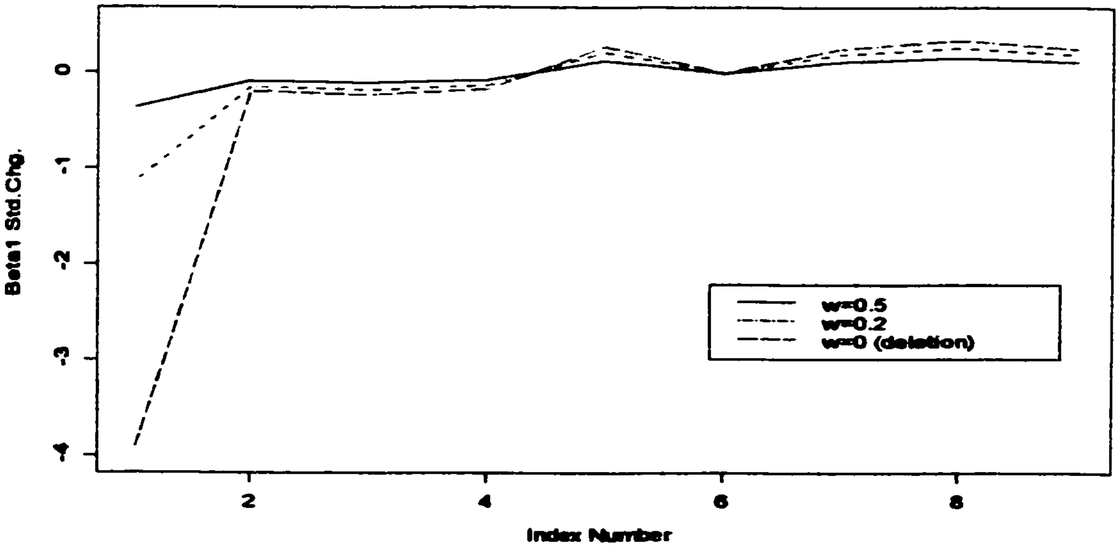

Figure 4.19: Standardized change in $\hat{\beta}_1$ for lot1 of bloodclot time

4.5 Conclusion

The diagnostic measures developed through the one-step function provide an effective counting device to modify the loglikelihood function which is not too time consuming. In fact, the one-step function presents an adequate way of detecting and quantifying the effect of outlying observations and extreme points for GLMs. It is noteworthy to mention that for logistic regression, the *Hauck-Donner* phenomenon can occur (see [27], p.225). When the $\hat{\beta}_j$ are large, the t statistic goes to zero. This implies that highly significant $\hat{\beta}_j$ may have non-significant t ratios. For example, when dealing with fitted values that are very close to either one or zero, then a dual conflict of the Hauck-Donner phenomenon and convergence problems may arise. This can be seen when dealing with a very large dataset of say, 1000 observations, and about fifty binary explanatory variables, whereby one of the covariates is always one to confirm the presence of a disease, for example. Then the resulting fitted probabilities with respect to that covariate must necessarily be one, and hence its associated regression coefficient, $\hat{\beta}_j = \infty$. This in turn implies that the maximum likelihood estimates do not exist.

Since the generalized linear models are members of the exponential family distributions, the computations and diagnostic measures described here can be extended to a greater scope to lead to applications in time series models and survival models. Some research work on diagnostic measures for survival models has been investigated by D. Pregibon.

Appendix A

Programs for Parameter Estimation for Different Families

A.1 MLE program for binomial family

```
# Binomial data program: sufficient statistic is the proportion of y to m
# for the ith observation.
#
muhati <- function(y,m) {
   muhat <- rep(NA, length(y))
   for (i in 1:length(y)) {
      if (y[i]/m[i]==0 || y[i]/m[i]==1) {
         muhat[i] <- (y[i]+0.5)/(m[i]+1)
      } else {
         muhat[i] <- y[i]/m[i]
      }
   }
   muhat
}

# Gather all information(z - adjusted dependent variate, X - covariates,
# W - weightvalue) in a dataframe - GenDataFrame.
#
GenDataFrame <- function(zValue, X, weightValue, nRows) {
   # Generate matdata.
   matdata <- data.frame(zValue, X[1,])
   if (nRows >= 2) {
     for (j in 2:nRows) {
        matdata <- data.frame(matdata, X[j,])
     }
   }
   matdata <- data.frame(matdata,weightValue)
}

# Purpose of this function is to create and execute the command:
# betaValue <- lm(zValue ~ x[1,]+x[2,]+x[3,]+etc..., matdataValue)$coefficients
# through concatenation of each covariate Xi.
#
Genlm <- function(zValue, X, matdataValue, nRows, weightValue) {
   # cat-file & parse file to generate Beta0
   cat("betaValue <- lm(zValue ~ X[1,]",file="tmp.1")
   if (nRows>=2) {
     for (i in 2:nRows) {
       cat("+X[", file="tmp.1", append=T)
       cat(i, file="tmp.1", append=T)
       cat(",]", file="tmp.1", append=T)
     }
   }

   cat(", weights = weightValue)$coefficients", file="tmp.1", "\n", append=T)
   #cat(", matdataValue, weightValue)$coefficients", file="tmp.1", "\n", append=

   # Now execute the created command
   eval(parse(file="tmp.1"), local=T)
}


# Purpose of this function is to create and execute the command:
# etahat <- betaValue[1] + betaValue[2]*X[1,] + betaValue[3]*X[2,] +
  betaValue[4]*X[3,] + etc.
#
Genetahat <- function(betaValue, X, nRows) {
   # Need a for loop to generate the required command.
   cat("etahat <- betaValue[1]",file="tmp.1")
   for (k in 1:nRows) {
      cat("+ betaValue[", file="tmp.1", append=T)
      cat(k+1, file="tmp.1", append=T)
      cat("]*X[", file="tmp.1", append=T)
      cat(k,file="tmp.1", append=T)
      cat(",]", file="tmp.1", append=T)
   }

   # Now execute the created command.
   eval(parse(file="tmp.1"))
}

# This part is made to measure for binomial data - need to extract
# pertinent statistics.
#
iterbin <- function(y, X, m, itmax=50) {
   # Find out how many Xi's, by the length of a column.
   nRows <- length(X[,1])

   muhat0  <- muhati(y,m)
   etahat0 <- logit(muhat0)
   weight0 <- rep(1, length(y))
   z0 <- etahat0 + ((y/m)-muhat0)/(muhat0*(1-muhat0))
   matdata <- GenDataFrame(z0, X, weight0, nRows)

   beta0 <- Genlm(z0, X, matdata, nRows, weight0)

   n <- 0
   for (i in 1:itmax)     {
      n <- n+1

      etahat <- Genetahat(beta0, X, nRows)
      muhat  <- exp(etahat)/(1+exp(etahat))
      weight <- m*muhat*(1-muhat)
      z <- etahat + m*((y/m)-muhat)/weight
      matdata <- GenDataFrame(z, X, weight, nRows)
      beta <- Genlm(z, X, matdata, nRows, weight)
      if (sum(abs(beta-beta0)) <= 10^(-10))  {
        return(list("Pass"=T, coefficients=beta, fittedValues=m*muhat,
        adjustedValue=z, Variance=weight, iterations=n))
      }
      beta0 <- beta
   }

   list("Pass"=F, coefficients=beta, iterations=n)
}
```

A.2 MLE program for Poisson family

```
# Poisson data program: sufficient statistic is the mean of y
# for the ith observation.
#y is vector of the sum of counts
muhatpoi <- function(y,m) {
  y/m
}


# Gather all information(z - adjusted dependent variate, X - covariates,
#  W - weightvalue) in a dataframe - GenDataFrame.
#
GenDataFrame <- function(zValue, X, weightValue, nRows) {
   # Generate matdata.
   matdata <- data.frame(zValue, X[1,])
   if (nRows >= 2) {
     for (j in 2:nRows) {
        matdata <- data.frame(matdata, X[j,])
     }
   }
   matdata <- data.frame(matdata,weightValue)
```

```
}


# Purpose of this function is to create and execute the command:
# betaValue <- lm(zValue ~ x[1,]+x[2,]+x[3,]+etc..., matdataValue)$coefficients
#
Genlm <- function(zValue, X, matdataValue, nRows, weightValue) {
   # cat-file & parse file to generate Beta0
   cat("betaValue <- lm(zValue ~ X[1,]",file="tmp.1")
   if (nRows>=2) {
     for (i in 2:nRows) {
       cat("+X[", file="tmp.1", append=T)
       cat(i, file="tmp.1", append=T)
       cat(",]", file="tmp.1", append=T)
     }
   }

   cat(", weights=weightValue)$coefficients", file="tmp.1", "\n", append=T)

   # Now execute the created command
   eval(parse(file="tmp.1"), local=T)
}


# Purpose of this function is to create and execute the command:
# etahat <- betaValue[1] + betaValue[2]*X[1,] + betaValue[3]*X[2,] +
# betaValue[4]*X[3,] + etc
#
Genetahat <- function(betaValue, X, nRows) {
   # Need a for loop to generate the required command.
   cat("etahat <- betaValue[1]",file="tmp.1")
   for (k in 1:nRows) {
      cat("+ betaValue[", file="tmp.1", append=T)
      cat(k+1, file="tmp.1", append=T)
      cat("]*X[", file="tmp.1", append=T)
      cat(k,file="tmp.1", append=T)
      cat(",]", file="tmp.1", append=T)
   }
   # Now execute the created command.
   eval(parse(file="tmp.1"))
}


# This part is made to measure for poisson data - need to extract
# pertinent statistics.
#
iterpoi <- function(y, X, m, itmax=100) {
   # Find out how many Xi's, by the length of a column.
   nRows <- length(X[,1])
   etahat0 <- log(muhatpoi(y,m))
   weight0 <- rep(1, length(y))
   z0 <- etahat0
   # Generate matdata.
   matdata <- GenDataFrame(z0, X, weight0, nRows)

   # cat-file & parse file to generate Beta0
   # beta0   <- lm(z0 ~ x[1,]+x[2,]+x[3,]+etc..., matdata)$coefficients
   beta0 <- Genlm(z0, X, matdata, nRows, weight0)

   h <- 0
   for (i in 1:itmax)     {
      h <- h+1
      # etahat <- beta0[1] + beta0[2]*X[1,] + beta0[3]*X[2,]+beta0[4]*X[3,]
      etahat <- Genetahat(beta0, X, nRows)

      muhat  <- exp(etahat)
      weight <- muhat
      z <- etahat + (y-muhat)/weight
      # Generate matdata
      matdata <- GenDataFrame(z, X, weight, nRows)
      beta <- Genlm(z, X, matdata, nRows, weight)

      if (abs(max(beta-beta0))<= 10^(-10))  {
        return(list("Pass"=T, coefficients=beta, fittedValues=muhat,
        adjustedValue=z, Variance=weight, iterations=h))
      }
      beta0<-beta
   }
   list("Pass"=F, coefficients=beta, iterations=h)
}
```

A.3 MLE program for Gamma family

```
# Gamma data program: sufficient statistic is the mean of y
# for the ith observation.
#
muhatgam <- function(y,m) {
   muhat <- rep(NA, length(y))
   for (i in 1:length(y)) {
         muhat[i] <- y[i]/m[i]
      }
   muhat
}


# Gather all information(z: adjusted dependent variate, X: covariates,
# W: weightvalue) in a dataframe - GenDataFrame.
#
GenDataFrame <- function(zValue, X, weightValue, nRows) {
   # Generate matdata.
   matdata <- data.frame(zValue, X[1,])
   if (nRows >= 2) {
     for (j in 2:nRows) {
        matdata <- data.frame(matdata, X[j,])
     }
   }
   matdata <- data.frame(matdata,weightValue)
}


# Purpose of this function is to create and execute the command:
```

```
# betaValue <- lm(zValue ~ x[1,]+x[2,]+x[3,]+etc..., matdataValue)$coefficients
#
Genlm <- function(zValue, X, matdataValue, nRows, weightValue) {
   # cat-file & parse file to generate Beta0
   cat("betaValue <- lm(zValue ~ X[1,]",file="tmp.1")
   if (nRows>=2) {
     for (i in 2:nRows) {
       cat("+X[", file="tmp.1", append=T)
       cat(i, file="tmp.1", append=T)
       cat(",]", file="tmp.1", append=T)
     }
   }

   cat(", weights=weightValue)$coefficients", file="tmp.1", "\n", append=T)

   # Now execute the created command
   eval(parse(file="tmp.1"), local=T)
}


# Purpose of this function is to create and execute the command:
# etahat <- betaValue[1] + betaValue[2]*X[1,] + betaValue[3]*X[2,] +
# betaValue[4]*X[3,] + etc
#
Genetahat <- function(betaValue, X, nRows) {
   # Need a for loop to generate the required command.
   cat("etahat <- betaValue[1]",file="tmp.1")
   for (k in 1:nRows) {
      cat("+ betaValue[", file="tmp.1", append=T)
      cat(k+1, file="tmp.1", append=T)
      cat("]*X[", file="tmp.1", append=T)
      cat(k,file="tmp.1", append=T)
      cat(",]", file="tmp.1", append=T)
   }

   # Now execute the created command.
   eval(parse(file="tmp.1"))
}
```

```
# This part is made to measure for gamma data - need to extract pertinent
# statistics.
#
itergam <- function(y, X, m, itmax=50) {
   # Find out how many Xi's, by the length of a column.
   nRows <- length(X[,1])

   muhat0 <-muhatgam(y,m)
   etahat0 <- inverse(muhat0)
   weight0 <- rep(1, length(y))
   z0 <- etahat0 + (y-muhat0)/((muhat0)^2)

   # Generate matdata.
   matdata <- GenDataFrame(z0, X, weight0, nRows)

   # cat-file & parse file to generate Beta0
   # beta0   <- lm(z0 ~ x[1,]+x[2,]+x[3,]+etc..., matdata)$coefficients
   beta0 <- Genlm(z0, X, matdata, nRows, weight0)

   h <- 0
   for (i in 1:itmax)     {
      h <- h+1
      # etahat <- beta0[1] + beta0[2]*X[1,] + beta0[3]*X[2,]+beta0[4]*X[3,]
      etahat <- Genetahat(beta0, X, nRows)

      muhat  <- inverse(etahat)
      weight <- muhat^2
      z <- etahat + (y-muhat)/weight
      # Generate matdata
      matdata <- GenDataFrame(z, X, weight, nRows)
      beta <- Genlm(z, X, matdata, nRows, weight)

      if (sum(abs(beta-beta0)) <= 10^(-10))  {
        return(list("Pass"=T, coefficients=(-1)*beta, fittedValues=muhat,
        adjustedValue=(-1)*z, Variance=weight, iterations=h))
      }
      beta0<-beta
   }
   list("Pass"=F, coefficients=beta, iterations=h)
}
```

A.4 One-step function

```
onestep <- function(X, V, W, z, i) {

  inverse <- solve(X %*% sqrt(V) %*% W %*% sqrt(V) %*% t(X))

  result <- inverse %*% X %*% sqrt(V) %*% W %*% sqrt(V) %*% z

}

w4onestep <- function(X,V,z,w,dimen){

     cat("", file="tmp.1")

     for (i in 1:dimen)     {
        W<- diag(rep(1,dimen))
        W[i,i] <- w
        temp <- onestep(X,V,W,z,i)
cat(temp, file="tmp.1", "\n", append=T)
      }

Value <- read.table("tmp.1")
}
```

Appendix B

B.1 Output for the Herbicide data

```
> iterbin(y.herb, X.herb, m.herb)
$Pass:
[1] T

$coefficients:
 (Intercept)      X[1, ]
   -2.620648 0.0001629353

$fittedValues:
 [1] 15.05633 14.98851 12.75041 12.41130 16.63094 24.07666 15.89181 14.85287
 [9] 14.06209 15.94563 16.54840 14.78505

$adjustedValue:
 [1] -2.980909 -2.476682 -2.178973 -2.742632 -2.425176 -2.089799 -2.212360
 [9] -2.465570 -2.958618 -2.593435 -2.656198 -2.605052

$Variance:
 [1] 14.03519 13.97197 11.88566 11.56955 15.22694 21.41755 14.67763
 [8] 13.84553 13.06339 14.76849 15.42607 13.78231

$iterations:
[1] 4
```

B.2 Output for One-Step function using the Herbicide data

```
> onestep(X1.herb, V,W,z)
[1] -2.6206479686  0.0001629353
> w4onestep(X1.herb, V,z, 0.2,12)
          V1           V2
 1 -2.585102 0.0001480545
 2 -2.634782 0.0001688525
 3 -2.657003 0.0001781548
 4 -2.610896 0.0001588526
 5 -2.618878 0.0001651574
 6 -2.621226 0.0001654647
 7 -2.639557 0.0001608354
 8 -2.635722 0.0001692459
 9 -2.590345 0.0001550040
10 -2.615674 0.0001624919
11 -2.616755 0.0001613054
12 -2.622156 0.0001635668
> w4onestep(X1.herb, V,z, 0,12)
          V1           V2
 1 -2.575092 0.0001438639
 2 -2.638761 0.0001705179
 3 -2.667047 0.0001823594
 4 -2.608209 0.0001577279
 5 -2.618366 0.0001657994
 6 -2.622215 0.0001697957
 7 -2.644727 0.0001602613
 8 -2.639960 0.0001710200
 9 -2.582075 0.0001528395
10 -2.614313 0.0001623706
11 -2.615644 0.0001608406
12 -2.622580 0.0001637442
```

Bibliography

[1] Aitken, M., Anderson, D. and Francis, B. (1989). *Statistical Modelling in GLIM.* Oxford University Press, New York.

[2] Anscombe, F.J. (1948). The Transformation of Poisson, Binomial, Negative Binomial data. *Biometrika,* **35** 246-254.

[3] Chaubey, Y.P. and Mudholkar, G.S. (1983). *On the Symmetrizing Transformations of random variables.* Paper unpublished.

[4] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Wiley, New York.

[5] Cox, D.R., Hinkley, D.V., Reid, N. and Snell, E.J. (1991). *Statistical Theory and Modelling: In Honour of Sir David Cox, FRS.* Chapman and Hall, London ; New York.

[6] Davison, A.C. and Gigli, A. (1989). Deviance Residuals and Normal Score Plots. *Biometrika,* **79** 211-221.

[7] Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis.* Second ed.. Wiley, New York.

[8] Firth, D. (1988). Multiplicative Errors: Log-Normal or Gamma? *J.R.Statist.Soc.B*, **50** 266-268.

[9] Green, P.J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation and some Robust and Resistant Alternatives. *J.R.Statist. Soc.*, **46** 149-192.

[10] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models, A roughness penalty approach.* Chapman and Hall, London.

[11] Hoaglin, D.C. and Welsch, R.E. (1978). The Hat Matrix in Regression and ANOVA. *Amer. Statistician*, **32** 17-22.

[12] Lindsey, J. K. (1997). *Applying Generalized Linear Models.* Springer-Verlag, New York.

[13] Mathai, A.M. and Provost, S. (1992). *Quadratic forms in random variables: theory and application.* Dekker, New York.

[14] MathSoft (1997). *S-PLUS Programmer's Guide*, Data Analysis Products Division, Seattle, WA.

[15] McCullagh, P. (1985). On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential Family Models. *International Statistcal Review*, **53** 61-67.

[16] McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

[17] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Second ed..Chapman and Hall, London.

[18] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *J. R. Statist. Soc. A*, **135** 370-383.

[19] Pregibon, D. (1980). Goodness of Link Tests for Generalized Linear Models. *Appl. Statist.*, **29** 15-24.

[20] Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics*, **9** 705-724.

[21] Pierce, D.A. and Schafer, D.W. (1986). Residuals in Generalized Linear Models. *JASA*, **396** 977-986.

[22] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York.

[23] Searle, S.R. (1971). *Linear Models*. Wiley, New York.

[24] Seber, G.A.F. (1977). *Linear Regression Analysis*. Wiley, New York.

[25] Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression*. Wiley, New York.

[26] Spector, P. (1994). *An Introduction to S and S-Plus*, Duxbury Press.

[27] Venables, W.N. and Ripley, B.D. (1999). *Modern Applied Statistics with S-Plus*, Third ed.. Springer-Verlag, New York.

[28] Williams, D.A. (1987). Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. *Appl. Statist.*, **36** 181-191.

[29] Zellner, A. (1976). Bayesian and Non-Bayesian Analysis of the Regression Model with Multivariate Student-t Error Terms. *JASA*, **354** 400-405.