Linear Model Diagnostics and Measurement Error

By

Ishara Siverpersad

Submitted in fulfilment of the academic
requirements for the degree of
Master of Science in Statistics,
School of Statistics and Actuarial Sciences
University of KwaZulu-Natal

Pietermaritzburg
2008

Linear Model Diagnostics and Measurement Error

by

Ishara Siverpersad

Submitted in fulfillment of the academic
requirements for the degree of
Master of Science in Statistics,
School of Statistics and Actuarial Sciences,
University of KwaZulu-Natal, PMB, South Africa

March 30, 2008

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Science, University of KwaZulu-Natal, Pietermaritzburg, from March 2005 to March 2008, under the supervision of Professor Temesgen Zewotir.

This study represents original work by the author and has not been submitted in any form for any degree or diploma to any University. Where use has been made of the work of others it is duly acknowledged in the text.

March, 2008.

Student: _____

Ishara Sivarpersad

Supervisor: _____

Professor Temesgen Zewotir

Acknowledgement

This dissertation would not have been possible without the direction of my supervisor Professor Temesgen Zewotir. I wish to express my appreciation to him for his interest, his assistance and guidance.

Thank you to the members of staff, Ms B Bonhomme, Miss S Ndlovu, Mr Shaun Moodley, Miss Meera Nepal and Mr Nazim Abdool Gani who have been a great help over the past two years.

Mr Naven Chetty, Mr Santosh Ramkissoon, Mr Muregancuro Gaeten Kabera and Dr Kyle Warwick Tomlinson: thank you for your friendship, time, and advice.

I am indebted to my family for their constant encouragement and support. A big thank you to San, Sunil, Patti, Lings, Kevash, Sujen, Yudithee and Bhavin.

Heartfelt thanks you to my mother who has had no choice but to tolerate me for many years. Thank you mother for everything.

Abstract

The general linear model, the weighted linear model, and the generalized linear model are presented in detail. Diagnostic tools for the linear models are considered. In general, the standard analysis for linear models does not account for measurement error. Measurement error models which apply strictly to the linear model are considered. The methods were applied to two real data sets. The Durban South project was carried out to investigate the effects of certain characteristics such as weight, height, gender and asthma status on the pulmonary lung function of children in and around the Durban South region. The Durban South region is heavily industrialized and as a result residents exposed to numerous pollutants are in and around the area. Various models are used in an attempt to describe the Durban South data. Linear diagnostic tools are extensively used to check the adequacy of the models.

The second data set is the Dendrometer data in which measurement error is suspected to play a big role. This data set is analysed to identify how rainfall, temperature, humidity and solar radiation affect the growth of the stem radius of eucalyptus. The covariates in these data appear to be highly correlated. Multicollinearity diagnostics are therefore used to check for multicollinearity. The linear regression model as well as measurement error models are investigated for their goodness of fit to the data. Of particular interest here is the effect measurement error has on the Dendrometer data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most useful models in statistical analysis is the general linear model. The general linear model includes various linear models that are widely used today in research and industry. At an elementary level, one is introduced to the simple linear regression model where there is only one predictor variable. The multiple linear regression model is an extension of the simple linear regression model in which a number of predictor variables are considered. In general, the objective of multiple linear regression modelling is to investigate the relationship of the response with a number of predictor variables simultaneously because almost always more than one predictor variable influences the response. When data has observations that are said to have a greater 'weight' than other observations on the model, weighted linear modelling is applied. The weighted linear model is commonly used to remedy the problem of unequal variances of the errors in the responses (Neter et al., 1996). Another valuable extension of the general linear model is to the generalized linear model (GLM). This type of model is essential when the response distribution is non-normal.

Estimation of model parameters is an essential aspect in linear modelling. The least-squares estimation method, and the method of maximum likelihood are commonly used to estimate the general linear model parameters. For the weighted linear model parameters, one can make use of the weighted least-squares method for estimation.

Generalized linear model parameters are usually estimated using the maximum likelihood method. In the case of generalized linear models, the estimates of the unknown parameters are obtained by means of iterative estimation methodologies, which include the Newton-Raphson algorithm and the iteratively reweighted least-squares algorithm. Once the parameter estimates are obtained, then the standard errors of the parameter estimates are also estimated.

In the above mentioned linear models, and in fact with all linear models, what is generally of interest is how the various predictor variables/covariates or factors affect the response (significance or insignificance of their individual and interaction effects with other predictor variables/covariates) and if the model adequately describes or represent the data. If the model does not represent the data then results obtained are inaccurate. To check the validity, appropriateness, and accuracy of the linear model, many diagnostic tools are utilized. These diagnostics, aid in assessing the behaviour of the variables, and how well the model accounts for the data generating mechanism. Diagnostics assess the correctness of the linear model for the data. Hence, diagnosis is an essential part of linear model analysis.

General linear model diagnostic analysis includes the analysis of residuals, in which outlying observations are identified, and the normality assumption, the constant variance assumption, and the correlation assumption are checked. Visual diagnostics such as the normal probability plot of residuals, and the residual plots (versus fitted values or time) are generally used. Outliers are often identified as observations with unusually large residuals. These observations are known to affect the parameter estimates, summary statistics such as the t or F statistics, the R-square statistic, and the residual mean square. A simple method to check the impact outliers have on these statistics is to drop the outliers and refit the model. Montogomery, Peck, and Vining (2001) emphasize that there should be strong nonstatistical evidence that the outlier is a bad observation before it is permanently removed.

Observations with a large influence are identified by the use of Cook's statistic, DIFFITS statistic, leverage, and COVRATIO (covariance ratio) statistics. The R-square

statistic and the standard errors of the regression coefficients are sensitive to influential observations. As the name suggests, influential observations have an impact on the analysis, in that the estimated model is drawn in their direction. The deletion method is also commonly used to assess the effect of influential observations on the estimated model. Once again, care has to be taken when these influential observations are being permanently deleted.

In many situations an observation that is classified as outlying can also be an influential observation, and as a result, an extreme observation is defined to be one that is outlying and/or influential. It is unacceptable when extreme observations have a significant impact on a estimated model. In a severe case, the parameter estimates may depend more on the extreme observations than on the majority of the data. A statistical model aims to represent every observation in a data set, not just a subset of the data. Consequently, one must assess the influence that extreme observations have on the model being used.

The normal probability plot of the residuals and the residual plots are used to assess assumptions as was mentioned before. These plots are also used to identify potential outliers, the need for transformations, inequality of variance, and nonlinearity.

Other diagnostic measures known as multicollinearity diagnostics are used to check high correlations amoung the covariates in a model. Multicollinearity diagnostics aid in identifying which covariates are highly correlated and how severe the correlations are. In many cases, depending on the nature of the data, remedial measures can be taken to eliminate multicollinearity.

For the detection of extreme observations when modeling with the GLM model, the diagnostics used are similar to those of the general linear model diagnostics. Further diagnostic measures for GLMs include those for assessing the appropriatness link of the function and the general goodness of fit of the model. There is no difference between the general linear model and the weighted linear model diagnostics.

The main components of linear models are the responses, the predictor/factors or co-

variate variables, and the model assumptions. It is essential to study the way in which the fitted model is affected by extreme observations and violation of model assumptions. Conclusions drawn from model fits that are seriously affected by a few extreme observations are often misleading or most probably incorrect. Likewise, when model assumptions are violated, the analysis will lead to incorrect conclusions. Accordingly, diagnostics is a vital component of linear model analysis. Given a data set, the first step is to model the data using the model that is believed to best describe the data at hand. Thereafter, complete diagnostic checks are carried out. If the diagnostic results confirm that the chosen model best fits the data and that the model assumptions have not been violated, then the final step is to interpret the results obtained. In this thesis linear model diagnostics are, therefore, reviewed and applied to linear modelling real data sets.

In addition to the standard linear model diagnostics, the focus in this thesis is to introduce and illustrate the application of a new type of 'diagnostic tool' that was proposed in 1983, during which the concept of regression models with predictor variables measured with error were introduced (Stefanski, 2000).

In a general regression situation there exists the response variable y, and the predictor variable X, and associated error ε. The error ε usually denotes the amount of random error associated with the response y, and is commonly referred to as that part of y which is unexplained by X. This standard type of regression model does not account for the possibility that in certain instances the predictor X cannot be observed directly, but instead is observed with inherent error. This error is commonly referred to as the 'measurement error', and the statistical models that account for this error are known as 'measurement error models' (Fuller, 1987).

To illustrate the presence of measurement errors in a given data set, consider an example taken from Carroll, Ruppert, and Stefanski (1995). The example contains data which are used to investigate the lung function in children. The response variable is categorical with two levels which represent the presence or absence of wheeze in children. The predictor variable of interest is the amount of exposure each child has to

nitrogen dioxide. However this predictor variable is measured as the concentration of nitrogen dioxide present in the bedroom and kitchen in each child's home. The analysis is carried out with y = 1 for presence, and y = 0 for absence respectively of wheeze in children; and the predictor variable X is the amount of exposure to nitrogen dioxide. Unfortunately, the amount of nitrogen dioxide each child in the study is exposed to exists not only in the kitchen and the bedroom of that child's home. A significant amount of nitrogen dioxide is found outside the home. Children spend a long time at school and outdoors, which allows for further exposure to the pollutant outside the home. The measurement error clearly exists, the predictor is assumed to be measuring the amount of nitrogen dioxide each child is exposed to at home, because that predictor does not account for all exposures outside the home. This illustrates one of many instances in which measurement errors can arise.

Ideally, where one is faced with the possibility of measurement errors, a measurement error model has to be found to account for the existing error. In order to perform measurement error analysis there is at times a need for additional information with regard to the data. A description of these types of additional data is given by Carroll et al. (1995). Some of these types include: replicate data, in which the error-prone variable is available in replicates; and instrumental data, which is a new predictor variable that is closely related to the error-prone predictor. When the variance of the measurement error variable can be estimated or predicted then the measurement error models can easily be applied to the data. Measurement error models are based on the nature of the data and the availability of additional information.

The standard analysis of linear models does not take measurement error into account, even though some variables included in the linear model might be measured with error. Therefore, in this study the effects of measurement error in linear models is reviewed and applied to linear modelling real data sets. The aim in this thesis is not only to present linear model diagnostics, but to introduce measurement error as another type of diagnostic tool that can be applied to data in which measurement error is believed to exist.

This thesis is divided into seven chapters. In Chapter 1, the general linear model diagnostics, the nature of measurement error, and the objectives of this thesis, were introduced.

In Chapter 2, the general linear model along with estimation of the model parameters and the hypothes tests about the model parameters are discussed. The Chapter also contains the theory of, the weighted linear model, and the generalized linear model.

In Chapter 3, the nature of outliers, leverage points, influential points, and various diagnostic tools are discussed in considerable detail. These include tools for the detection of outliers and influential observations, as well as residual analysis.

Measurement error models are presented in Chapter 4. The focus is on additive measurement error. As a result, the models considered in Chapter 4 are those which pertain only to additive measurement error. For the general linear model, assessment of the effect of measurement error is carried out by means of the asymptotic approach, the perturbation approach, the simulation approach, and the bootstrap method. The SIMEX (Simulation Extrapolation) method is also introduced and an attempt is made to extend this to the case of the generalized linear model (Hardin, Schmiediche, and Carroll, 2003).

In Chapter 5, the Durban South data are analysed. A demonstration of how and why model diagnostics are essential in a practical problem is presented.

In Chapter 6, the effect measurement errors on inferences is demonstrated using real data. The data are first analysed assuming no measurement error, and thereafter analysed taking measurement error into account.

The summary of the study results is presented in Chapter 7.

Chapter 2

The General Linear Model

2.1 Introduction

The linear model is commonly fitted to the data by means of linear regression techniques, techniques that are used to model data where predictor variable value changes the response, so that the effects that some variables may exert on others can be carefully examined. Regression analysis is one of the most extensively used statistical methods. The analysis is valued for the general inferences it can produce.

The primary assumptions of the linear model are that the responses are independent are normally distributed with constant variance. In many practical situations these assumptions do not hold. When this occurs, transformations of the response variable may be appropriate. With the occurrence of nonconstant variance, weighted least squares can be used. In general, the alternative approach to data transformation, when assumptions of normality and nonconstant variance are not satisfied, is to model using the generalized linear model (GLM).

The generalized linear model is a unification of both linear and nonlinear regression models that also allows the incorporation of non-normal response distributions. The general linear model, weighted linear model, and the GLM are briefly discussed in this chapter.

2.2 The General Linear Model

2.2.1 Model

The general linear model is defined as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, i = 1, ..., n \tag{2.1}$$

or in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

where $\mathbf{y}$ is a $(n \times 1)$ vector of observable responses, $\mathbf{X}$ is the $(n \times p)$ known matrix whose i^{th} row is $\mathbf{x}_i{}' = (1, x_{1i}, x_{2i}, ..., x_{ki})$, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of parameters to be estimated so that $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ and $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of unobservable errors. The variable n is the total number of observations made, k is the number of explanatory variables, and $p = k + 1$. Model (2.2) is referred to as the simple linear regression model when $k = 1$, and as the multiple linear regression model when $k \geq 2$.

2.2.2 Estimation of the Model Parameters

There are two commonly used methods of estimation: these are the least squares and the maximum likelihood methods. The method of least squares is a deterministic formulation involving the minimization of the error sum of squares. The goal of this method is to choose values of the parameters for the model which 'best' fits the data in the sense that they minimise the error sum of squares.

The least squares problem can also be considered from a probabilistic point of view in which the best parameter estimates corresponds to the most probable estimates. Put simply, the 'best' parameter values also maximize the likelihood of the parameters, i.e., the maximum likelihood method chooses as parameter estimates those values of the parameters that are most consistent with the sample data (Neter et al., 1996) under the $N(\mathbf{0}, \sigma^2\mathbf{I})$ error structure.

The method of least squares can easily be applied to (2.2). The method of least squares is explained in matrix terms. Firstly, the error sum of squares is a measure of the amount of variability unaccounted for by the regression plane. One wishes to find the vector of least squares estimates, $\hat{\boldsymbol{\beta}}$, that minimizes the error sum of squares which from (2.2) is given by

$$\begin{aligned} Q = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\beta. \end{aligned} \tag{2.3}$$

The error sum of squares is then differentiated with respect to $\boldsymbol{\beta}$, and the result is equated to zero, with $\boldsymbol{\beta}$ being replaced by $\hat{\boldsymbol{\beta}}$. This provides the least squares normal equation

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

Assuming $(\mathbf{X}'\mathbf{X})$ is nonsingular, the least squares estimate of $\boldsymbol{\beta}$ is the value $\hat{\boldsymbol{\beta}}$, and is the solution of the normal equation given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}. \tag{2.4}$$

The fitted values are obtained as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. The following assumptions concerning the error term ε are made:

1. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ (the model is correct).
2. $Var(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma}^2\mathbf{I}$ (constant variance).
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$ (uncorrelated errors).
4. The errors ε_i are approximately normally distributed.

Assumptions 3 and 4 imply that the errors are independent random variables. The statistical properties of the least squares estimator, $\hat{\boldsymbol{\beta}}$, can be obtained from Montgomery, Peck, and Vinning (2001). The ordinary least square estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. It can further be shown (Montgomery et al., 2001) that $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. The variance of $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is a $p \times p$ symmetric

matrix whose jth diagonal element is the variance of $\hat{\beta}_j$ and whose (ij)th off-diagonal element is the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$.

The estimator for variance σ^2 can easily be developed from the residual sum of squares. From (2.3) the residual sum of squares is given as

$$Q = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}. \tag{2.5}$$

The residual sum of squares has $n - p$ degrees of freedom associated with it, since p parameters are estimated in the regression model. The residual mean square becomes

$$MS_{Res} = \frac{Q}{n-p}.$$

Montgomery et al. (2001) show that the expected value of MS_{Res} is σ^2, and hence an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = MS_{Res}.$$

Since $\hat{\sigma}^2$ depends on the residual sum of squares, any violation of the assumptions on the model errors, or any misspecification of the model form may seriously damage the usefulness of $\hat{\sigma}^2$ as an estimate of σ^2. Because σ^2 is computed from the regression model residuals, it is said (Montgomery et al., 2001) to be a model dependent estimate of σ^2.

The other common method of estimation is the 'maximum likelihood' in which the log likelihood function of $\boldsymbol{\beta}$ or the log of the joint density function of the error terms is maximized to yield an estimator equivalent to the least squares estimator of $\hat{\boldsymbol{\beta}}$. This is explained in detail.

The model is given by expression (2.2), in which the errors are normally and independently distributed, with constant variance σ^2, or $\boldsymbol{\varepsilon}$ is distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$. The normal density function for the errors as given by Montgomery et al. (2001) is

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2}.$$

The likelihood function is the joint probability density function (p.d.f) of $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, or $\prod_{i=1}^{n} f(\varepsilon_i)$, which may be regarded as a function of β and σ^2 and, when so regarded,

is denoted by

$$L(\boldsymbol{\varepsilon}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}.$$

Now since $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, the likelihood function becomes

$$L(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}.$$

What is of interest is the value of $\boldsymbol{\beta}$ and σ^2 that maximize the probability $L(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)$ of obtaining the particular observed sample $x_1, x_2, ..., x_n$. Certainly, the maximizing values of $\boldsymbol{\beta}$ and σ^2 would seemingly be good estimates of the parameters because they would provide the largest probability of observing the particular sample. Since the likelihood function $L(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \sigma^2)$ and its logarithm are maximized for the same values of $\boldsymbol{\beta}$ and σ^2, either can be used. However, it is convenient to work with the log of the likelihood

$$lnL(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.6}$$

To find the maximum value of $lnL(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)$ equation (2.6) is differentiated with respect to $\boldsymbol{\beta}$ and σ^2 and equated to zero to give the estimates in (2.4) and

$$\hat{\sigma}^2 = \frac{Q}{n} \tag{2.7}$$

respectively. The estimate for $\hat{\boldsymbol{\beta}}$ is the same as that obtained from the least squares estimation. However, the estimate of σ^2 differs from that obtained using the least squares method.

2.2.3 Hypothesis Testing

Hypothesis testing includes testing the overall significance of the model as well as testing the significance of the individual regression coefficients.

The test for the overall significance of the linear regression model is to see if there is indeed a linear relationship between the response variable and any of the predictor variables. The hypotheses are:

1. $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0.$

2. $H_1 : \beta_j \neq 0$ for at least one j.

To test these hypotheses the analysis of variance (ANOVA) is carried out. The ANOVA table is given in Table 2.1. If the null hypothesis is rejected this implies that at least one of the predictor variables contributes to the model, i.e., the regression coefficient of one or more predictors does not equal zero.

Table 2.1: Analysis of Variance Table

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	Q	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

In the ANOVA table part of the total sums of squares (SS_T) are due to the regression (SS_R) and the remainder are due to residuals (Q). To obtain the partitioning, one begins with the identity

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

By squaring both sides of the above equation, summing over all n observations, rearranging and cancelling the equation becomes

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The left hand side of the equation is the corrected sums of squares of the observations, i.e., the total sum of squares SS_T which measures the total variability in the responses. SS_T can be rewritten as

$$SS_T = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \mathbf{y}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}. \tag{2.8}$$

The first component of SS_T is $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ which measures the the amount of variability in the observations y_i accounted for by the regression plane; and the second

component, $Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, as mentioned measures the amount of variation due to the residuals (that is, not explained by the regression line).

From (2.5) and (2.8) Q can be rewritten as

$$Q = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \left[\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]$$

or

$$Q = SS_T - SS_R.$$

Therefore, the regression sum of squares is

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

If the null hypothesis is true, then SS_R/σ^2 follows a χ^2_k distribution, which has the same number of degrees of freedom as the number of regressor variables in the model (Montgomery et al., 2001). The authors also show that $Q/\sigma^2 \sim \chi^2_{n-k-1}$ and that Q and SS_R are independent. From which

$$F_0 = \frac{SS_R/k}{Q/(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

follows the F distribution with degrees of freedom k and $n-k-1$.

Montgomery et al. (2001) also showed that if at least one $\beta_j \neq 0$, then F_0 follows a noncentral F distribution with k and $n-k-1$ degrees of freedom and a noncentrality parameter of $\lambda = \frac{\boldsymbol{\beta}^{*'}\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}^*}{\sigma^2}$, where $\boldsymbol{\beta}^* = (\beta_1, \beta_2, ..., \beta_k)'$ and $\mathbf{X}_c$ is the 'centered' model matrix. This noncentrality parameter also indicates that the observed value of F_0 should be large if at least one $\beta_j \neq 0$. As a result, in order to test the H_0 : $\beta_1 = \beta_2 = ... = \beta_k$ the test statistic F_0 is first computed and H_0 is rejected if the test statistic exceeds $F_{\alpha,k,n-k-1}$, where α is the level of significance. This test procedure is summarized in the ANOVA table given in Table 2.1.

Once it has been established that the linear regression model has a significant overall fit, i.e., at least one of the regressors is important, one needs to identify the regressor/s.

The hypotheses for testing the significance of any individual regression coefficient, such as β_j, are:

1. $H_0 : \beta_j = 0$.

2. $H_1 : \beta_j \neq 0$.

If $H_0 : \beta_j = 0$ is not rejected, then this indicates that x_j can be deleted from the model. The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

where c_{jj} is the jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$, corresponding to $\hat{\beta}_j$. The null hypothesis is rejected if $|t_0| > t_{\alpha/2,n-k-1}$. It is emphasized by Montgomery et al. (2001) that this is really a partial or marginal test because the estimated regression coefficient $\hat{\beta}_j$ depends on all of the other regressor variables $x_{ij}(i \neq j)$ that are in the model. Thus, this is a test of the contribution of x_i, given the other regressors in the model.

2.2.4 Confidence Intervals

In addition to the point estimates of the regression coefficients one may also construct confidence interval estimates of these parameters. The width of a confidence interval is a measure of the overall quality of the regression plane.

Since the marginal distribution of any regression coefficient $\hat{\beta}_j$ is normal with mean β_j and variance $\sigma^2 c_{jj}$, where c_{jj} is the jth diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}}, j = 0, 1, ..., k$$

has a student t with $n-p$ degrees of freedom. Therefore a $100(1-\alpha)$ percent confidence interval for the regression coefficient β_j is given by

$$\hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 c_{jj}}.$$

Some aspects concerning the general linear model have been briefly discussed. These aspects included the least squares estimation method, hypothesis testing, and the construction of confidence intervals. The weighted linear model will be considered next.

2.3 Weighted Linear Model

In certain modelling applications the weights of the data points tend to vary in that each datum point does not provide equally precise information about the variation in the responses. In other words, the variances of the responses are not constant. When this occurs it is not reasonable to assume that every observation should be treated equally. A direct analysis method to allow for this variation in error variance is the method of weighted least squares. The weighted least square criterion gives less precisely measured points less influence and highly precise points more influence on the estimated model.

The method of weighted least squares takes into account the behaviour of the random errors in the data. It incorporates non-negative constants, called weights, associated with each datum point into the fitting criterion. The weighted model is then optimized in a similar manner as the ordinary linear model to yield parameter estimates which account for the precision of the information contained in the data by use of weights.

The weighted linear is model (2.2) with

$$\epsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{W}^{-1})$$

where $\mathbf{W^{-1}} = cov(\mathbf{y})$ is an $n \times n$ diagonal matrix whose j^{th} element is σ_j^2.

2.3.1 Weighted Least Squares Method

The weighted least squares method is similar to the ordinary least squares method in that the error sum of squares are minimised by choice of $\boldsymbol{\beta}$. However, with the weighted least squares method, the error sum of squares that are minimised are given by

$$WQ = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2.$$

where $\mathbf{x}'_i = (1, x_{1i}, x_{2i}, ..., x_{ki})$. The $\boldsymbol{\beta}$ which minimises WQ, called the weighted least squares estimator, is given by

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

The method of weighted least squares has the advantage over the ordinary least squares method in that it is able to handle regression situations in which the data points are of varying quality. The biggest disadvantage of weighted least squares is that in many cases the weights are not known exactly. Neter et al. (1996) discuss two methods of obtaining estimates for the variances which are used to obtain the weights, and they also briefly discuss the case when variances are known up to a proportionality constant. The first method for estimating the weights involves obtaining the ordinary least squares residuals and thereafter estimating the variance function or the standard deviation function by regressing either the squared residuals or the absolute residuals on the appropriate predictor variable(s). Here the fitted values from the estimated variance or standard deviation function are used to obtain the weights. Neter et al. (1996) further advise that the iteratively reweighted least squares method be used when the estimated regression coefficients by the WLS method differ substantially from the estimated regression coefficients obtained by ordinary least squares. The other method of estimating weights requires replicated data. Here the error variance is estimated from the replicate observations that are made at each combination of levels of the predictor variables. It should be noted that when weights are estimated from small numbers of replicated observations, the results of the analysis can be very badly and unpredictably affected (Carroll and Ruppert, 1988). The inference procedures when weights are estimated are now only approximate (Neter et al., 1996) because the estimation of the variances introduces another source of variability. According to Neter et al. (1996), this approximation is often good when the sample size is not too small.

2.4 Generalized Linear Models

The linear model is an important statistical model that is widely used. However, its use is limited to those settings where the response distributions are normal. Also, if the means of the responses are naturally restricted to a range of values, the linear model may not be appropriate since the linear predictor, $\mathbf{X}'\boldsymbol{\beta}$, can take on any value. Furthermore, it may not be realistic to assume that the variance of the data is constant for all the responses. In its basic form the linear model is unable to account for such effects.

As a result, the linear model can be generalized to allow it to overcome these limitations. The generalizations accept forms of non-normal data, and also link the mean response to the linear predictor in possible linear fashions. These generalizations of the linear model result in a class of models known as the family of generalized linear models (GLMs).

The class of generalized linear models has gained much popularity in recent years. This is as a result of the flexibility of generalized linear models in addressing a variety of statistical problems and of the availability of software to fit the models. The SAS (Statistical Analysis Software) system provides the GENMOD (Generalized Linear Modelling) procedure for generalized linear modelling.

The GLM model which consist of a single response variable, and the predictor variable/s, is a member of the exponential family of probability distributions. Generalized linear modelling transforms the relationship between the linear predictor and the mean response, such that a nonlinear relationship can be modelled as a linear one. This admits a model specification allowing for continuous or discrete responses, and allows a description of the variance as a function of the mean response. The GLM family members are linearized by means of a link function, and are fitted using maximum likelihood techniques, with the help of iterative algorithms.

The generalized linear models were first defined by Nelder and Wedderburn (1972). Dobson (1990), Myers, Montgomery, and Vinning (2002) and Aitkin et al. (1989) have

many examples and applications of generalized linear models. McCullagh and Nelder (1989) and McCulloch (2001) give a thorough account of statistical modelling using generalized linear models.

As shown by Hilbe (1993a) and Francis, Green, and Payne (1993) in Hardin and Hilbe (2001), generalized linear models are characterized by an expanded itemized list given by

1. A random component for the response, $\mathbf{y}$, which has a distribution following the exponential family.

2. A linear predictor component
$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

3. A known monotonic, one-to-one, differentiable link function $g(.)$ relating the linear predictor to the mean response follows.
$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = E(\mathbf{y}).$$

4. The covariance matrix of $\mathbf{y}$ may depend on $\mathbf{X}$ but only through $\boldsymbol{\mu}$.

2.4.1 The Link and Variance Functions

The mean $\boldsymbol{\mu} = E(\mathbf{y})$ and the linear predictor, $\mathbf{X}\boldsymbol{\beta}$, are related by means of the link function as follows:
$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}.$$
The variance function $\mathbf{V}(.)$ relates the $Var(\mathbf{y})$ to the mean $\boldsymbol{\mu}$ as follows
$$Var(\mathbf{y}) = \mathbf{a}(\phi)\mathbf{V}(\mu)$$
where $a(\phi)$ is the dispersion factor.

The assumptions in generalized linear modelling are that the n observations are independent, the variance function, the dispersion factor, the link function, and the

explanatory variables are of the correct form, and that there is no undue influence of the individual observations on the fit. The above mentioned assumptions are given by Hardin and Hilbe (2001).

2.4.2 The Exponential Family of Distributions

The probability density function of a distribution that belongs to the exponential family is usually written as:

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

where θ is the canonical (natural) parameter and ϕ is the dispersion (scale) parameter required to produce standard errors following a distribution in the exponential family of distributions. The joint density of the sample of observations, y_i, given parameters θ and ϕ, is given by the product of the densities over the individual observations since the observations are independent. Conveniently, the joint probability density function may be expressed as a function of θ and ϕ given the observations y_i. This is called the likelihood, L, and is written as

$$L(\theta,\phi;y_1,y_2,...,y_n) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}.$$

In order to obtain the estimates of (θ_i,ϕ) that maximize the likelihood function, it is convenient to work with the log likelihood

$$\mathcal{L} = \mathcal{L}(\theta,\phi;y_1,y_2,...,y_n) = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}. \tag{2.9}$$

The log likelihood is used because the values of (θ_i,ϕ) that maximize the likelihood are the same values that maximize the log likelihood.

2.4.3 Estimation

To obtain maximum likelihood estimates of the θ_i, the following identities are used (Hardin and Hilbe, 2001)

$$E\left(\frac{\partial \mathcal{L}}{\partial \theta_i}\right) = E\frac{y_i - b'(\theta_i)}{a(\phi)} = 0 \tag{2.10}$$

and

$$E\left\{\frac{\partial^2 \mathcal{L}}{\partial \theta_i^2} + \left(\frac{\partial \mathcal{L}}{\partial \theta_i}\right)^2\right\} = 0. \tag{2.11}$$

From (2.10) it follows that

$$b'(\theta_i) = E(y_i) = \mu_i \tag{2.12}$$

where μ_i is the mean value parameter. From (2.11) the following is obtained

$$\begin{aligned} E\left\{\frac{\partial^2 \mathcal{L}}{\partial \theta_i^2} + \left(\frac{\partial \mathcal{L}}{\partial \theta_i}\right)^2\right\} = 0 &= -\frac{b''(\theta_i)}{a(\phi)} + \frac{1}{a(\phi)^2}E(y_i - \mu_i)^2 \\ &= -\frac{b''(\theta_i)}{a(\phi)} + \frac{1}{a(\phi)^2}Var(y_i) \end{aligned}$$

which implies that

$$b''(\theta_i) = \frac{1}{a(\phi)}Var(y_i). \tag{2.13}$$

Solving for $Var(y_i)$, it can be seen that

$$Var(y_i) = b''(\theta_i)a(\phi). \tag{2.14}$$

The above can be rewritten as

$$Var(y_i) = V(\mu_i)a(\phi) \tag{2.15}$$

because the variance is a function of the mean μ_i, through $b''(\theta_i)$. As such, $b''(\theta)$ is referred to as the variance function. From (2.12) to (2.15), it follows that

$$\frac{\partial \mu_i}{\partial \theta_i} = V(\mu_i).$$

Using identities (2.10), and (2.11), models that allow a parameterization of the mean $\boldsymbol{\mu}$ in terms of covariates $\mathbf{X}$ and their coefficients $\boldsymbol{\beta}$ can be developed. The covariates are

introduced through a known (invertible) function that links the mean $\boldsymbol{\mu}$ to the linear predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, such that

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta}$$

and

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

Finally, since $\eta_i = \sum_{j=1}^{p} x_{ji}\beta_j$, then

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ji}.$$

As pointed out by Hardin and Hilbe (2001), the linear predictor is not constrained to lie within any range; in fact $\eta \in \Re$. What the link function does is map the linear predictor to the range of the response variable and thereby ensures that the variance function is non-negative. As a result, an unconstrained linear model in terms of the form of the coefficients and associated covariates is maintained.

Hardin and Hilbe (2001) show, using the chain rule, that the derivatives of the log likelihood with respect to the β_j are given by

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) x_{ji}, j = 1, 2, ..., p \tag{2.16}$$

Newton's method is a linear Taylor series approximation in which the derivative of the log-likelihood is written in a Taylor series expansion. This method can be used to obtain the parameter estimates. The method is illustrated by Hardin and Hilbe (2001). The first and second derivative of the log-likelihood are given as $l'(\boldsymbol{\beta}) = \partial l/\partial \boldsymbol{\beta}$, and $L''(\boldsymbol{\beta}) = \partial^2 L/(\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}')$. Solving the estimating equations

$$\mathcal{L}'(\boldsymbol{\beta}) = 0 \tag{2.17}$$

for $\boldsymbol{\beta}$ provides the parameter estimates. By expanding, simplifying, and solving for $\boldsymbol{\beta}$, the Taylor series becomes

$$\boldsymbol{\beta} \approx \boldsymbol{\beta}^{(0)} - \frac{\mathcal{L}'(\boldsymbol{\beta}^{(0)})}{\mathcal{L}''(\boldsymbol{\beta}^{(0)})}.$$

The estimation may be iterated using

$$\boldsymbol{\beta}^{(r)} \approx \boldsymbol{\beta}^{(r-1)} - \frac{\mathcal{L}'(\boldsymbol{\beta}^{(r-1)})}{\mathcal{L}''(\boldsymbol{\beta}^{(r-1)})}, \tag{2.18}$$

for $r = 1, 2, ..., n$ and a reasonable starting value of $\boldsymbol{\beta}^{(0)}$.

The above estimation method is exact in the case of the linear regression model, i.e., when the function is truly quadratic. The estimation algorithms are the Newton-Raphson and the iteratively reweighted least squares (IRLS). Both optimization methods are said to be equivalent under certain conditions. However, differences in computer output do occur (Hardin and Hilbe, 2001). This is said to be as a result of the differences in starting values, convergence paths of the algorithms, and in some instances when the likelihood has a relatively flat area in one or more of the respective dimensions in the optimization problem. However, when the two algorithms are not equivalent they are equal in the limit, i.e., they have the same expected value.

The Newton-Raphson optimization algorithm, solves (2.17), and implements (2.18) without change. The observed matrix of second derivatives (hessian matrix) as derived by Hardin and Hilbe, (2001) is

$$\left(\frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_k}\right) = \sum_{i=1}^{n} \frac{1}{a(\phi)} \left(\frac{\partial}{\partial\beta_k}\right) \left\{ \frac{y_i - \mu_i}{V(\mu_i)} \left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji} \right\} \tag{2.19}$$

$$= -\sum_{i=1}^{n} \frac{1}{a(\phi)} \left[\frac{1}{V(\mu_i)} \left(\frac{\partial\mu}{\partial\eta}\right)_i^2 - (\mu_i - y_i)\psi \right] x_{ji} x_{ki}, \tag{2.20}$$

where $\psi = \left\{ \frac{1}{V(\mu_i)^2} \left(\frac{\partial\mu}{\partial\eta}\right)_i^2 \frac{\partial V(\mu_i)}{\partial\mu_i} - \frac{1}{V(\mu_i)} \left(\frac{\partial^2\mu}{\partial\eta^2}\right)_i \right\}$. Using the first and second derivative given in (2.16) and (2.19) respectively, the Newton-Raphson algorithm is implemented by optimizing (2.18) to yield the maximum likelihood estimates.

The iteratively reweighted least squares algorithm, which related to Fisher's method of scoring, is also used to obtain the maximum likelihood estimates for the exponential family.

Rewriting the (usual) updating formula from the Taylor series expansion presented in

(2.18) as

$$\delta\boldsymbol{\beta}^{(r-1)} = -\left\{\frac{\partial^2\mathcal{L}}{\partial(\boldsymbol{\beta}^{(r-1)})^T\partial\boldsymbol{\beta}^{(r-1)}}\right\}^{-1}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}^{(r-1)}} \tag{2.21}$$

and the calculation of the observed Hessian (second derivatives) is replaced with it expected values. This replacement is known as the method of Fisher scoring. Since $E(y_i-\mu_i)^2 = V(\mu_i)a(\phi)$, one can write

$$\begin{aligned} -E\left(\frac{\partial^2\mathcal{L}}{\partial\beta_j\partial\beta_k}\right) &= E\left(\frac{\partial\mathcal{L}}{\partial\beta_j}\frac{\partial\mathcal{L}}{\partial\beta_k}\right) \\ &= \sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\frac{1}{V(\mu_i)a(\phi)}x_{ji}x_{ki}. \end{aligned} \tag{2.22}$$

Substituting (2.22) and (2.16) into (2.21) and rearranging, it can be seen that $\delta\beta^{(r-1)}$ is the solution to

$$\left\{\sum_{i=1}^{n}\frac{1}{V(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\delta\beta^{(r-1)} = \sum_{i=1}^{n}\frac{y_i-\mu_i}{V(\mu_i)a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji}. \tag{2.23}$$

The value of the linear predictor at the $(r-1)th$ iteration of the algorithm is

$$\eta_i^{(r-1)} = \sum_{k=1}^{p} x_{ki}\beta_k^{(r-1)}$$

which can be rewritten as

$$\left\{\sum_{i=1}^{n}\frac{1}{V(\mu_i^{(r-1)})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\beta^{(r-1)} = \sum_{i=1}^{n}\frac{1}{V(\mu_i^{(r-1)})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\eta_i^{(r-1)}x_{ji}. \tag{2.24}$$

Summing (2.23) and (2.24) and thereafter substituting (2.21), one obtains the following

$$\begin{aligned} &\left\{\sum_{i=1}^{n}\frac{1}{V(\mu_i^{(r-1)})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\left(\delta\beta^{(r-1)}+\beta^{(r-1)}\right) = \\ &\sum_{i=1}^{n}\frac{1}{V(\mu_i^{(r-1)})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\left\{(y_i-\mu_i)\left(\frac{\partial\eta}{\partial\mu}\right)_i+\eta_i^{(r-1)}\right\}x_{ji}\times \\ &\left\{\sum_{i=1}^{n}\frac{1}{V(\mu_i^{(r-1)})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji}x_{ki}\right\}\beta^{r} = \\ &\sum_{i=1}^{n}\frac{1}{V(\mu_i^{(r-1)})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\left\{(y_i-\mu_i)\left(\frac{\partial\eta}{\partial\mu}\right)_i+\eta_i^{(r-1)}\right\}x_{ji} \end{aligned} \tag{2.25}$$

Let

$$\mathbf{W}^{(r-1)} = diag\left\{\frac{1}{V(\mu_i^{r-1})a(\phi)}\left(\frac{\partial\mu}{\partial\eta}\right)_i^2\right\} \tag{2.26}$$

$$\mathbf{z}^{(r-1)} = \left\{(\mathbf{y}-\boldsymbol{\mu})\left(\frac{\partial\eta_i}{\partial\mu_i^{(r-1)}}\right) + \boldsymbol{\eta}^{(r-1)}\right\}_{(n\times 1)} \tag{2.27}$$

so that (2.25) can be rewritten in matrix notation as

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

The above algorithm is the IRLS algorithm and it calculates the new estimate of the coefficient vector by means of weighted ordinary least squares.

Goodness of Fit

The measure of the goodness of fit of the model to the data is defined as twice the difference between the log-likelihoods of the model of interest and the saturated model (which consists of n parameters as it includes one parameter for each observation and hence results in the fitted values being equal to the observed values). Since this difference is a measure of the deviation from a perfectly fitting model, the measure is called the deviance. The competing goals in modelling are to find the simplest model (fewest parameters) that has the smallest deviance (reproduces the data).

The deviance D is given by

$$D(\mathbf{y},\boldsymbol{\mu}) = 2\left(l(\mathbf{y},\mathbf{y}) - l(\mathbf{y},\boldsymbol{\mu})\right), \tag{2.28}$$

where $l(\mathbf{y},\boldsymbol{\mu})$ is the log-likelihood function expressed as a function of the predicted mean value $\boldsymbol{\mu}$ and the vector $\mathbf{y}$ of response values. In fitting a model, one seeks the values of the parameters that minimize the deviance. Thus the optimal values, using the IRLS algorithm, are those at which the difference in deviance calculations between successive iterations is very small (less than a specified tolerance). The values of the parameters which minimize the deviance are the same as the values of the parameters which maximize the likelihood.

Estimation of the covariance of $\hat{\boldsymbol{\beta}}$

Once optimal estimates of $\boldsymbol{\beta}$ are found, it is necessary to estimate the variance matrix for $\hat{\boldsymbol{\beta}}$. The Newton-Raphson and the IRLS result in parameter estimates that differ only in numeric roundoff or differences in optimization criteria. Essentially, the IRLS is equivalent to the first term in (2.20), which is

$$E\left(\frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_j}\right) = -\sum_{i=1}^{n} \frac{1}{a(\phi)} \frac{1}{V(\mu_i)} \left(\frac{\partial\mu}{\partial\eta}\right)_i^2 x_{ji} x_{ki}. \tag{2.29}$$

The Hessian may be calculated as given in (2.20), or may be calculated using the more restrictive (naive) assumptions of the IRLS algorithm as given in (2.29). As shown by Hardin and Hilbe (2001), (2.20) and (2.29) are equivalent when the canonical link is used.

The implementation of the IRLS algorithm is not limited to using one version or the other of the Hessian. For example, PROC GENMOD in SAS has the SCORING option. This allows the user to specify how many of the initial iterations are performed using the Fisher scoring method, expected Hessian calculated using (2.29), before all subsequent iterations are performed using the observed Hessian calculated in (2.20), even though the overall optimization is Newton-Raphson (optimizes the log likelihood).

The usual variance estimate in statistical packages is calculated (numerically or analytically) as the inverse of (negative) second derivatives. In the case of generalized linear models, one may calculate the Hessian in two ways.

Assuming that the conditional mean is specified correctly, the expected Hessian can be calculated as

$$\hat{\mathbf{V}}_{EH} = \left\{ element(u,v) = \left(-\frac{\partial^2 \mathcal{L}}{\partial\beta_u \partial\beta_v}\right)\right\}^{-1}$$

for $u, v = 1, ..., p$, where p is the dimension of $\mathbf{X}$. Here $-\partial^2\mathcal{L}/(\partial\beta_u\partial\beta_v)$ is calculated using (2.29).

More generally, one can calculate the observed Hessian without the additional assump-

tion of correct conditional mean specification using

$$\hat{\mathbf{V}}_{OH} = \left\{ element(u,v) = \left(-\frac{\partial^2 \mathcal{L}}{\partial \beta_u \partial \beta_v} \right) \right\}^{-1}$$

for $u, v = 1, ..., p$, where p is the dimension of $\mathbf{X}$. Here, $-\partial^2 \mathcal{L}/(\partial \beta_u \partial \beta_v)$ is calculated using (2.20).

There are many other methods of estimating the variance of $\hat{\boldsymbol{\beta}}$ in GLMs and these include: outer product of the gradiant (OPG), Sandwich, modified sandwich, unbiased sandwich, modified unbiased sandwich, weighted sandwich, Newey-West, Jackknife, and bootstrap methods (Hardin and Hilbe, 2001). Here, only the Hessian matrix, $\hat{\mathbf{V}}_{EIM}$, is considered.

Chapter 3

Diagnostics

3.1 Introduction

The results obtained from the linear model analysis are based strictly on assumptions which in many instances are not suitable. Problems occur when inferences reached are strongly influenced by a small portion of the data and therefore tend not to reflect the data as a whole.

When the linear model is applied to any data, there has to be absolute certainty that the assumptions made are not severely violated in the sense that reasonable alternative structures that may exist for analysis must not produce different conclusions, and that a different sample should not lead to a totally different model with opposite conclusions. In any standard linear model analysis the assumptions have to be closely examined. For instance, the possibility that error variances are not constant, or that the errors deviate from normality in any way should be explored. Certain diagnostic checks may suggest that the general linear model is inappropriate, and that transformations should be considered. Also, the data has to be checked for possible outliers that may exist. Unfortunately, the adequacy of the model assumptions cannot be checked by examination of the standard summary statistics, such as the t or F statistics, or R^2 because these are global model properties.

One method of overcoming the above mentioned difficulties is to employ a fitting criterion that is not as vulnerable to unusual data and requires fewer valid assumptions. This criterion is termed 'robust model'. Another method is use of diagnostic methods. Diagnostic methods are used to identify, in the modelling process, unusual behavior of observations which is usually overlooked, and can also be used to remedy these situations.

It should be noted that robust methods and diagnostics are not mutually exclusive. This chapter considers diagnostic tools for linear models.

3.2 The Ordinary Residual

Diagnostic methods are primarily based on the study of the model residuals, so one should first define the residual.

Since errors, $\boldsymbol{\varepsilon}$, in (2.2) are unobservable, in order to check the validity of the assumptions made about the error, the residuals are used. The vector of ordinary residuals, $\mathbf{e}$, is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{3.1}$$

where $\hat{\mathbf{y}}$ is the vector of fitted values. Clearly the vector of residuals in (3.1) is the difference between what is actually observed, and what is predicted by the fitted model. The residual can also be viewed as the variability in the response variable not explained by the regression model. Thus, the $e_i, i = 1, 2, ..., n$, could be thought of as the observed errors if the model was correct.

The relationship between $\mathbf{e}$ and ε can be established firstly by defining the 'hat' matrix

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

The matrix $\mathbf{H}$ maps the vector of observed values into a vector of fitted values, i.e., $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. The vector of residuals in (3.1) becomes

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}. \tag{3.2}$$

The equation linking $\mathbf{e}$ to $\boldsymbol{\varepsilon}$ is obtained by substituting $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for $\mathbf{y}$ in equation 3.2 so that

$$\begin{aligned} \mathbf{e} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}. \end{aligned} \tag{3.3}$$

From (3.3) it can clearly be seen that the relationship between $\mathbf{e}$ and $\boldsymbol{\varepsilon}$ depends heavily on $\mathbf{H}$. If $\mathbf{H}$ has large elements in absolute value, the values of $\boldsymbol{\varepsilon}$ and $\mathbf{e}$ will differ, but if on the other hand $\mathbf{H}$ has sufficiently small elements in absolute vale, $\mathbf{e}$ can be substituted for $\boldsymbol{\varepsilon}$. It can also be seen from (3.3) that $\mathbf{e}$ will follow a normal distribution with $E(\mathbf{e}) = \mathbf{0}$ and $Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$. Variation in the ordinary residual is therefore determined by $\mathbf{H}$. The hat matrix is discussed in more detail.

3.2.1 The Hat Matrix and Leverage

The hat matrix plays a central role in regression analysis. By definition, the matrix is a linear transformation that orthogonally projects any n-vector onto the space spanned by the columns of $\mathbf{X}$ (Montgomery et al., 2001).

Properties of the hat matrix

1. The matrix is symmetric, i.e., $\mathbf{H}' = \mathbf{H}$.
2. It is idempotent, i.e., $\mathbf{H^2} = \mathbf{H}$.
3. Since $\mathbf{H}$ is idempotent and symmetric, it follows that $trace(\mathbf{H}) = rank(\mathbf{H}) = rank(\mathbf{X}) = p$, $\sum_j h_{ij}^2 = h_{ii}$, and that $\mathbf{H}$ is invariant under nonsingular linear reparameterizations.

The diagonal elements, h_{ii}, of the hat matrix $\mathbf{H}$, can be written as

$$h_{ii} = \mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{x_i}$$

where $\mathbf{x}_i'$ is the ith row of the $\mathbf{X}$ matrix. In geometric terms, the hat matrix's ith diagonal element is a standardized measure of the distance of the ith observations from the center of the x-space. From the third property above, it can be deduced that the average of the diagonal elements of a hat matrix is $\bar{h} = p/n$. A large ith diagonal element implies that the ith observation is remote in the x-space from the rest of the data.

Since $Var(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$ and $Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$, fitted values at remote points will have relatively large variances, and the corresponding residuals will have relatively small variances.

The elements h_{ij} of the $\mathbf{H}$ matrix may be interpreted as the amount of leverage exerted by the ith observation y_i on the ith fitted value $\hat{y}_i$. The diagonal elements are often called the leverages, since examination of

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_i$$

indicates via h_{ii} how heavily y_i contributes to $\hat{y}_i$. The fitted value $\hat{y}_i$ will be dominated by $h_{ii}\hat{y}_i$ if h_{ii} is large relative to the remaining terms, and clearly for any h_{ii} greater than 0, $\hat{y}_i$ will be dominated by $h_{ii}y_i$ if y_i is an outlier. High-leverage points are outliers in the x-space, but the converse is not necessarily true.

Traditionally (Montgomery et al., 2001), it is assumed that an observation with a leverage which exceeds $2p/n$ is considered a leverage point, however, this cutoff only applies to situations where $2p/n \leq 1$.

3.3 Residuals and Outlier Detection

For diagnostic analysis, several variations of the ordinary residuals are used to try and overcome some of their limitations. The ordinary residuals have a distribution that is scale dependent since the variance of $\mathbf{e}$ is a function of σ^2 and $\mathbf{H}$. The variations of the ordinary residual do not depend on σ^2 or $\mathbf{H}$, and can therefore be referred to as 'scaled residuals'. These scaled residuals are helpful in finding observations that are

outliers, or extreme, that is, observations that are separated in some fashion from the rest of the data.

3.3.1 Standardized Residuals

The variance of the residuals or errors can be estimated by

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Reg} = s^2.$$

As a result a logical scaling for the residuals would be the standardized residuals

$$d_i = \frac{e_i}{\sqrt{s^2}}, i = 1, 2, ..., n$$

The standardized residuals have mean zero and approximately unit variance. Consequently, a large absolute standardized residual ($|d_i| > 3$, say) potentially indicates an outlier (Montgomery et al., 2001).

3.3.2 Studentized Residuals

The division of the scale dependent statistic by its scale estimate produces a ratio with a distribution that does not depend on the nuisance scale parameter. This division is known as 'studentization'.

The 'internally studentized residual', also known just as the 'studentized residual' is defined as

$$s_i = \frac{e_i}{s(1-h_{ii})^{\frac{1}{2}}}$$

where $s^2 = \mathbf{e}'\mathbf{e}/(n-p) = \sum e_i^2/(n-p)$. Each residual is divided by its standard error. These studentized residuals are said to be internally studentized because the s^2 is an internally generated estimate of σ^2 obtained from fitting the model to all n observations.

For externally studentized residuals, an estimator of σ^2 that is independent of e_i is required. The estimate of σ^2 with observation i removed is given by

$$S^2_{(i)} = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}.$$

The externally studentized residual also known as the 'R-student' or the 'R-residual' can therefore be defined as

$$t_i = \frac{e_i}{S_{(i)}(1-h_{ii})^{\frac{1}{2}}}, i = 1, 2, ..., n.$$

An advantage of using the t_i is that, if e_i is large, it is emphasized even more by the fact that $S_{(i)}$ is independent of it. The t_i follows a student t distribution with $n-p-1$ degrees of freedom under the usual normality of errors assumption.

An equivalent procedure for finding the studentized residuals employs a mean shift outlier model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \gamma R + \varepsilon$$

where R is a dummy regressor variable to one for observation i and to zero for all other observations. This model is appropriate when observation i is believed to be significantly different from the rest of the data. To test $H_0 : \gamma = 0$, the test statistic

$$t_i = \frac{e_i/(1-h_{ii})^{\frac{1}{2}}}{\{[\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} - e_i^2/(1-h_{ii})]\,/\,[n-p-1]\}^{\frac{1}{2}}}$$

is calculated. This test statistic has a student t distribution under H_0, and is identical to the externally studentized residual.

3.3.3 Testing for Outliers

In most applications the data has to be checked for outliers that may occur. One way to do this, as mentioned by Fox (2005), is be to refit the mean-shift model n times, producing studentized residuals $t_1, t_2, ..., t_n$. It is not literally necessary to perform n auxiliary regressions. Usually interest then focuses on the largest absolute t_i value,

denoted by t_{max}. Because the biggest of the n test statistic is used, it is not appropriate simply to use $t(n-k-1)$ to find a p-value for the test based on t_{max}. One solution to this problem is to perform a Bonferroni adjustment to the p-value of the test based on t_{max}. This is done by letting $p' = Pr(t(n-k-1) > t_{max})$. Then the Bonferroni p-value for testing the statistical significance of t_{max} is p-value $= 2np'$. Note that a much larger t_{max} is required for a statistically significant result than would be the case for an ordinary individual t-test. Another approach for outlier detection is to construct a quantile-comparison plot for the studentized residuals, plotting against either the t or normal distribution quantiles. Outliers can also be identified by use of the maximum normed residual $|e_i|/\sqrt{\sum_{i=1}^{n} e_i^2}$ that is easy to apply.

3.3.4 Treatment of Outliers

Outliers are extreme observations which are isolated, i.e., far away from the rest of the data. Residuals that are considerably larger in absolute value than the others, say three or four standard deviations from the mean, indicate potential y-space outliers. Depending on their location in x-space, outliers can have a moderate to severe effect on the regression model.

Outliers have to be carefully examined to try and determine the reason for their unusual behaviour. Outliers may occur as a result of inaccurate measuring instruments, incorrect recording of data, failure of a measurement instrument or various other explainable events. In certain situations the outlier can be corrected (if possible) or deleted from the data set. Discarding bad values is desirable because least squares pulls the fitted equation toward the outlier because it minimizes the residual sum of squares. However, Montgomery et al. (2001) emphasize that there should be strong nonstatistical evidence that the outlier is a bad value before it is discarded. For instance, if an outlier is a perfectly plausible observation, deleting it to improve the fit of the equation can be dangerous because it can give the user a false sense of precision in estimation or prediction. The outlier may be important if it controls key model

properties and may also point out inadequacies in the model, such as failure to fit the data well in a certain region of x-space. If an outlier is a point of particularly desirable responses, then knowledge of the regressor values when that response was observed may be extremely valuable.

Various statistical tests that have been mentioned can be used for the detection of outliers. While the tests are useful for identifying outliers, these outliers cannot be hastily removed. Outliers may contain valuable information about the data. Generally, the regression model is fitted without the outliers to check the effect that outliers have on the regression analysis. For instance, the t or F statistic, the R^2 statistic, and the residual mean square may be very sensitive to the outliers. The regression equation is valid if it is not overly sensitive to the outliers.

3.4 Plotting Methods

Residuals can be used in a variety of graphical summaries to identify inappropriate assumptions. A number of different residual plots can be used to investigate the adequacy of the fit of a regression model and to check the underlying assumptions. In the following sections, basic residual plots are discussed.

3.4.1 Normal Probability Plot

Severe violation of the normality assumption is serious because the t or F statistics and confidence and prediction intervals depend on the normality assumption. As pointed out by Montgomery et al. (2001), if the errors come from a distribution with thicker or heavier tails than normal, the least squares fit may be sensitive to a small subset of the data. Furthermore, heavy-tailed error distributions often generate outliers that 'pull' the least squares fit too much in their direction.

A simple method of checking the normality assumption, aside from the histogram and the stem and leaf plot, is to construct the normal probability plot of the residuals.

The normal probability plot is a plot of each residual vs its expected value under the normality assumption. The resulting plot should be approximately a straight line if the normality assumption holds. Departures from a straight line indicate that the distribution is not normal. In practice, samples taken from a normal distribution will obtain an exact straight line. Montgomery et al. (2001) note that small sample sizes ($n \leq 16$) often produce normal probability plots that deviate substantially from linearity, and that in larger sample sizes ($n \geq 32$) the plots are much better behaved. It is also noted that the normal probability plots often exhibit no unusual behaviour even if the errors ε_i are not normally distributed.

If the normal probability plot shows large residuals, sometimes the corresponding observations are outliers.

3.4.2 Plot of Residuals Against Other Values

Draper and Smith (1998) perform checks for the time effects, nonconstant variance, and the possible need for transformation and curvation. Using the residuals are plotted vertically against:

1. The time order of the data, if known.
2. The corresponding fitted values $\hat{y}_i$ from the fitted model.
3. The corresponding x_i values if there is only one predictor variable; or, in general, each set of x_{ij}, where $j = 1, 2, ..., k$ represents the x's in the regression.

In all of the cases, a satisfactory plot is one that shows a (more or less) horizontal band of points. The points ought to appear random with no pattern. The residuals should also be plotted in whatever way that is sensible for the particular problem under consideration.

3.5 Measures of Influence

3.5.1 Cook's Statistic

Influential observations cannot be detected using the model-fitting process or by analysis of the residuals. They have a significant effect on the parameter estimates, and are usually detected by the Cook's statistic. Cook's statistic is a measure of the squared distance between the least-squares estimate $\hat{\boldsymbol{\beta}}$ from all the n observations and the estimate obtained after deleting the ith observation, say $\hat{\boldsymbol{\beta}}_{(i)}$. This distance can be expressed in a general form as

$$D_i(\mathbf{M}, c) = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{M}(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}})}{c}, i = 1, 2, ..., n.$$

The usual choices of $\mathbf{M}$ and c are $\mathbf{M} = \mathbf{X}'\mathbf{X}$ and $c = pMS_{Reg}$, so that the above statistic becomes

$$D_i(\mathbf{X}'\mathbf{X}, pMS_{Res}) = D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}})}{pMS_{Res}}.$$

Observations i with large values of D_i have considerable influence on the least-squares estimate $\hat{\boldsymbol{\beta}}$. As noted by Montgomery et al. (2001), the magnitude of D_i is usually assessed by comparing it to $F(\alpha, p, n-p)$.

The D_i statistic can also be rewritten in two other ways. The first way is

$$D_i = \frac{s_i^2}{p}\frac{Var(\hat{y}_i)}{Var(e_i)} = \frac{s_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}, i = 1, 2, ..., n$$

where s_i is the internally studentized residual. Here the first term in the product is a measure of discrepancy, and the second is a measure of leverage. The second way is when the statistic is given by the squared scaled distance

$$D_i = \left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\right)'\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\right)/(ps^2)$$

where $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\beta}_{(\mathrm{i})}$. The D_i is the squared distance between (the ends of) the vectors $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_{(i)}$, divided by ps^2. If omission of the ith observation makes little difference

to the fitted values, D_i will be small and therefore a large D_i indicates the deletion of observation i greatly affects the fitted values.

Cook's statistic can also be extended to evaluate the influence of observations in pairs, triplets and so on. The size-adjusted cutoff for Cook's statistic, is

$$D_i = \frac{4}{n-p-1}.$$

Absolute cutoffs for D, such as $D_i > 1$, risk missing relatively influential data according to Fox (2005).

3.5.2 DFFITS Statistic

Another statistic for measuring influence is the DFFITS statistic

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}, i = 1, 2, ..., n$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of the ith observation. The denominator is just a standardizer, since $Var(\hat{y}_i) = \sigma^2 h_{ii}$. Thus, $DFFITS_i$ is the number of standard deviations by which the fitted value $\hat{y}_i$ changes if observation i is removed.

The $DFFITS$ statistic can be calculated using the following formula

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} t_i.$$

The cutoff suggested for the $DFFITS$ statistic is $2\sqrt{p/n}$. It should be noted that all cutoffs are merely suggestions and should not be strictly followed.

3.5.3 Covariance Ratio

The influential diagnostics mentioned above are used to provide insight about the effect of observations on the estimated coefficients, $\hat{\beta}_j$, and the fitted values $\hat{y}_i$. However, the

above do not provide information about the overall precision of estimation. In order to do this, the generalized variance of $\hat{\boldsymbol{\beta}}$ is defined as

$$GV(\hat{\beta}) = \left|Var(\hat{\boldsymbol{\beta}})\right| = \left|\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right|$$

so that the role of the ith observation on the precision of estimation can be measured by

$$COVRATIO_i = \frac{\left|\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1} S_i^2\right|}{\left|(\mathbf{X}'\mathbf{X})^{-1} MS_{Res}\right|}.$$

If $COVRATIO_i$ is greater than 1, the ith observation improves the precision of estimation, while if $COVRATIO_i$ is less than 1, inclusion of the ith point degrades precision. The $COVRATIO_i$ can be calculated using the following formula

$$COVRATIO_i = \frac{\left(S^2_{(i)}\right)^p}{MS^p_{Res}} \left(\frac{1}{1-h_{ii}}\right).$$

The cutoff suggested by Belsley, Kuh, and Welsch (1980) is that if $COVRATIO_i > 1 + 3p/n$, or if $COVRATIO_i < 1 - 3p/n$, then the ith point should be considered influential. The lower bound is appropriate when $n > 3p$. These cutoffs are recommended only for large samples.

3.5.4 Treatment of Influential Observations

The criterion of deciding whether an influential observation should be discarded is similar to that of deciding whether an outlier should be removed.

The observation in question should be closely examined to try and determine why it is an influential observation. As a general rule, if there is an error in recording a measured value, or if the sample point is indeed invalid or not part of the population that was intended to be sampled, then discarding the observation is appropriate. However, if analysis reveals that the influential point is a valid observation, then it cannot be removed.

The alternative, suggested by Montgomery et al. (2001), as a compromise between deletion and retainment of influential observations, is that of robust estimation techniques. This is said to down weight observations in proportion to residual magnitude or influence, so that a highly influential observation will receive less weight than it would in a least-squares fit.

3.6 Multicollinearity Diagnostics

In order to fit the model (2.2) the solution to (2.4) must exist. A unique solution to the normal equations will only exist when the inverse of $\mathbf{X}'\mathbf{X}$ exists, i.e., $\mathbf{X}'\mathbf{X}$ is not singular. If $\mathbf{X}'\mathbf{X}$ is singular this implies that at least one column of $\mathbf{X}$ is linearly dependent on the other columns. In this case it is said (Draper and Smith, 1998) that collinearity or multicollinearity exists among the columns of $\mathbf{X}$. In general one assumes that the word 'multicollinearity' means 'near-dependence' in the $\mathbf{X}$ columns. The phrase 'near-dependence' refers to the existence of multicollinearity when $det(\mathbf{X}'\mathbf{X})$ does not equal zero. If the $det(\mathbf{X}'\mathbf{X}) = 0$ then exact multicollinearity exists.

Sources of multicollinearity sometimes occur as a result of the method of data collection employed, constraints on the model or in the population, model specification, and also overparametrized models. Multicollinearity has many effects on the least-squares estimates of the model parameters. Some of these effects include large variances and covariances of the least-squares estimator of the parameters, $\beta_j, j = 1, 2, ..., n$, estimates being too large in absolute value and the eigenvalues of $\mathbf{X}'\mathbf{X}$ being too small.

Several techniques have been proposed for detecting muticollinearity. The discussion that follows is of a few diagnostic measures that aid in identifying the degree of multicollinearity, and the regressors involved.

3.6.1 Examination of the Correlation Matrix

The basic measure of multicollinearity is the correlation matrix. If the elements of the off-diagonal element of this matrix are close to unity in absolute value, then the corresponding regressors are considered to be highly correlated.

3.6.2 Variance Inflation Factors (VIFs)

When the model (2.2) is fitted by least squares, the variances of the estimates are (Montgomery et al., 2001)

$$V(\hat{\beta}_i) = VIF_i(\sigma^2/S_{ii}), i = 1, 2, ..., k$$

where $S_{ii} = \sum_{u=1}^{n}(X_{iu} - \bar{X}_i)^2$ is the corrected sum of squares of the $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{in})'$ column. If an $\mathbf{X}_i$ column is orthogonal to all other columns of the $\mathbf{X}$ matrix, $VIF_i = 1$. Thus VIF_i is a measure of how much σ^2/S_{ii} is inflated by the linear relationship of other columns of $\mathbf{X}$ with $\mathbf{X}_i$ column.

The VIFs are the diagonal elements of the inverse of the correlation matrix. If R_i^2 is the multiple correlation coefficient obtained when the ith predictor variable $X_i, i = 1, 2, ..., k$, is regressed against all the remaining predictors X_j with $j \neq i$, then it can be shown (Draper and Smith, 1998) that

$$VIF_i = (1 - R_i^2)^{-1}.$$

3.7 GLM Diagnostics

With the GLMs, one assumes that the link and variance functions are correctly specified. Also, it is assumed that the linear structure expresses the relationship of the explanatory variables with the response variable. Various statistics and techniques useful for assessing the validity of these assumptions are considered here.

3.7.1 Deviance

In a maximum likelihood estimation, one assessment method of the goodness of fit model is to compare the fitted model to a fully specified model (a model with as many independent parameters as there are observations). The scaled deviance S is given in terms of the likelihoods of the (fitted) model L_m and the full model L_f by

$$S = -2ln(L_m/L_f).$$

One may generalize this concept to a generalized linear model such that

$$S = D/\phi,$$

where D is the deviance. The deviance for the generalized linear model is given in (2.28). Another definition for deviance is (Hardin and Hilbe, 2001)

$$D = 2\phi\mathcal{L}(y;y) - \mathcal{L}(y;u),$$

where $\mathcal{L}(y;y) = ln(L_f)$, and $\mathcal{L}(y;u) = ln(L_m)$. The deviance can be used to compare models, and the smaller the deviance the better the fitted model.

3.7.2 Leverage

In ordinary regression, the diagonal elements of the hat matrix are used as a measure of leverage. With GLMs the hat matrix

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{1/2}$$

is obtained from the last iteration of the updating algorithm being used in (2.6) and the leverage is interpreted in the same way as with the ordinary regression.

3.7.3 Residual Analysis

In regression analysis, residuals are used to check violations of assumptions such as homogeneous variances. Analysis of residuals is important in fitting the GLM as well,

as it provides guidance concerning the overall adequacy of the GLM model and assists in verifying the GLM model assumptions.

Response residuals

These residuals are simply the difference between the observed and fitted values which is given by

$$r_i^R = y_i - \hat{\mu}_i.$$

In GLM the response residuals are technically not appropriate for use, since $Var(y_i)$ is not constant (Myers et al., 2002).

Working residuals

These residuals are the difference between the working (synthetic) response and the linear predictor at convergence which is given by

$$r_i^W = (y_i - \hat{\mu}_i)\left(\frac{\partial \hat{\eta}}{\partial \hat{\mu}}\right)_i,$$

and are associated with the working response and the linear predictor in the IRLS algorithm given in (2.27).

Pearson residuals

The Pearson residual given as

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

is a re-scaled version of the working residual. The sum of the squared Pearson residuals is equal to the Pearson chi-squared statistic.

A large residual (in absolute value) indicates a failure on the part of the model to fit a particular observation (Hardin and Hilbe, 2001). A common diagnostic for detecting outliers is to plot the standardized Pearson residuals versus the observation number.

Partial residuals

Partial residuals, which are used to assess the form of a predictor and are thus calculated for each predictor (Hardin and Hilbe, 2001), are defined as

$$r_{li}^{T} = (y_i - \hat{\mu}_i)\left(\frac{\partial\hat{\eta}}{\partial\hat{\mu}}\right)_i + (x_{il}\hat{\beta}_l)$$

where $l = 1, ..., k$ and k is the number of predictors. In the above equation, $(x_{il}\hat{\beta}_l)$ refers to the ith value of the lth predictor times the lth fitted coefficient.

Deviance residuals

The deviance residual is defined as

$$r_i^D = \sqrt{\hat{d}_i^2}(sign(y_i - \hat{\mu}_i)).$$

Computational formulas of d_i^2 for individual families are given by Hardin and Hilbe, (2001). In general, the deviance residual (standardized or not) is preferred to the Pearson residual for model checking since its distributional properties are similar to those of the residuals arising from fitting linear regression models.

Adjusted deviance residuals

The deviance residual may be adjusted to make the convergence to the limiting normal distribution faster. The adjustment removes an $O(n^{-1/2})$ term, and the residual is defined as

$$r_i^{D_a} = r_i^D + \frac{1}{6}\rho_3(\theta)$$

where $\rho_3(\theta)$ is defined for individual families by Hardin and Hilbe (2001).

Likelihood residuals

A standardized residual is one in which the variance of the residual has been standardized to take into account the correlation between y and $\hat{\mu}$ (Hardin and Hilbe, 2001). The base residual has been multiplied by the factor $(1 - h)^{-1/2}$.

Likelihood residuals are a combination of standardized Pearson residuals, $r_i^{P'}$, and standardized deviance residuals, $r_i^{D'}$. The likelihood residual is defined as

$$r_i^L = sign(y_i - \hat{\mu}_i)\sqrt{h_i(r_i^{P'})^2 + (1 - h_i)(r_i^{D'})^2}.$$

Score residuals

The score residual given as

$$r_i^S = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \left(\frac{\partial \hat{\eta}}{\partial \mu}\right)_i^{-1}$$

is related to the score function (estimating equation), which is optimized (Hardin and Hilbe, 2001).

3.7.4 Cook's Distance

By using the one-step approximation to the change in the estimated coefficients when an observation is left out of the estimation procedure, one can approximate Cook's distance (Hardin and Hilbe, 2001) with

$$C_i = \left(\hat{\boldsymbol{\beta}}_{(i)}^* - \hat{\boldsymbol{\beta}}\right)^T \mathcal{I} \left(\hat{\boldsymbol{\beta}}_{(i)}^* - \hat{\boldsymbol{\beta}}\right)$$

where $\mathcal{I}$ is the Fisher information defined as the inverse of the Hessian. Furthermore,

$$\hat{\boldsymbol{\beta}}_{(i)}^* = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}x_i\hat{r}_i^R}{1 - \hat{h}_i}$$

where $\hat{\mathbf{V}} = diag(\mathbf{V}(\hat{\mu}))$ is an $(n \times n)$ matrix, $\mathbf{X}$ is the $(n \times p)$ matrix of covariates, $\hat{h}_i$ is the ith diagonal of the hat matrix, $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector, and $\hat{r}_i^R$ is the estimated response residual. The estimate $\hat{\boldsymbol{\beta}}_{(i)}^*$ is the one-step approximation to the jackknife estimated coefficient vector (Hardin and Hilbe, 2001).

3.7.5 Diagnostic Plots

To check for systematic departures from the GLM model, the standard diagnostic plots are considered. The deviance residuals can be used for constructing diagnostic plots.

The deviance residuals are very nearly the same as those generated by the best possible normalizing transformation. Myers et al. (2002) plot the deviance residuals against the fitted values and against the regressors. To check normality they use the Pearson residuals. These plots are analogous to the common residual and normal plots used in multiple regression. They have exactly the same interpretation.

It is suggested by McCullagh and Nelder (1989) and Myers (2002) that the absolute values of the deviance residuals against the fitted values are plotted, when transformed to the constant scale. A poorly chosen variance function should produce a trend in this plot.

To check the appropriatness of the individual covariates, Hardin and Hilbe (2001) suggest plotting residuals versus one of the predictor variables. These plots should not depict a null trend, because the null pattern is the same as that for the residuals versus the fitted values.

Added variable plots check whether an omitted covariate, u, should be included in the linear predictor. First, one would obtain the unstandardized residuals for u as the response, using the same linear predictor and quadratic weights as for y. The unstandardized residuals for y are then plotted against the residuals for u. There should be no trend if u is correctly omitted.

3.7.6 Model Statistics

Akaike information criterion (AIC)

The Akaike information criterion is generally used to compare models, in order to choose the best model that fits the data. A comparison may be made with nonnested models or models fitted to different samples. The criterion is such that the lower the value, the better fitting is the model. The formula is given by

$$AIC = \frac{-2\mathcal{L}(M_k) + 2p}{n}$$

where $\mathcal{L}(M_k)$ is the likelihood for model k.

Bayesian information criterion (BIC)

The Bayesian information criterion is similar to the AIC in that the measure may be used to compare nonnested models or models fitted to different samples. The criterion is such that a saturated model has criterion zero. Again, the smaller the value is, the better fitting is the model. If the BIC is positive, then the saturated model is preferred (Hardin and Hilbe, 2001). The formula is given by

$$BIC = D(M_k) - (df)ln(n)$$

where $D(M_k)$ is the deviance of model k, and df are the degrees of freedom.

3.7.7 Assessing the link function

A plot of the vector, $\mathbf{z}$, given in (2.27) against the linear predictor, $g(\boldsymbol{\mu})$, provides an informal check of the appropriatness of the link function. If one rewrites (2.27) as

$$z_i = g(\hat{\mu}_i) + g'(\hat{\mu}_i)r_i^R \tag{3.4}$$

then by plotting $\mathbf{z}$ against the $g(\hat{\boldsymbol{\mu}})$, one can assess the link function. Whe the graph is a straight line, then the link function $g(.)$ is appropriate.

Chapter 4

Measurement Error Model

4.1 Introduction

A measurement error model refers to a linear model with the independent variables whose values are measured with errors. This error is usually ignored in a standard regression analysis. Usually, it is assumed that the dependent variable, y, is the only variable with error. This assumption is never likely to be satisfied, especially when the data are complex measurements.

Measurement errors occur due to a number of reasons which include sampling errors, analysis errors, errors in scale, rounding errors, etc. The error can be explicit or implicit, it can be fully and clearly expressed, or it can be implied and indirect. Carroll et al., (1995) as well as Fuller (1987) give a number of studies which illustrate occurences of measurement error.

Overlooking the possibility of measurement error in variables when estimating regression parameters results in asymptotically biased, i.e., inconsistent estimators. This is the motivation for investigating measurement error models. In general, data that are exposed to a great deal of measurement errors will produce very different results when this measurement error is accounted for when compared to those results obtained when measurement error is disregarded. Any statistical analysis aims to produce the most

accurate results so that misinterpretation of the model at hand is minimized.

It is surprising that the analysis for measurement error models is not included in the standard statistical software packages available. However, Hardin and Carroll (2003) introduce software for analysing certian measurement error models with the STATA package. Lederer and Kuchenhoff (2006) have also introduced similar software with the **R** package.

4.1.1 The Additive Measurement Error Model

Measurement error models for the response $\mathbf{y}$, is a linear model with two types of predictors: the predictor, $\mathbf{Z}$, representing measurements which, for all practical purposes, are considered to be error free; and the predictor $\mathbf{X}$ representing unobservable measurements due to errors. Instead, $\mathbf{W} = f(x, u)$, is observed where u represents measurement errors. In the model, values of the predictor, $\mathbf{Z}$, are treated as fixed constants.

Measurement error models are said to be based upon two major defining characteristics. The first characteristic includes the error structure and the data structure. The error structure refers to the specification of the model for the measurement errors. Such models include the Classical Measurement Error Models, Error Calibration Models, and Regression Calibration Models (Carroll et al., 1995). The classical error model, in its simplest form, is appropriate when an attempt is made to determine $\mathbf{X}$ directly, but one is unable to do so because of various errors in measurement. In such a circumstance, it sometimes makes sense to hypothesize an unbiased additive error model, which is given by

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{4.1}$$

where $\mathbf{U}$ is independent of $\mathbf{X}$, has mean zero, and variance $\sigma^2 = 1$.

The data structure refers to the type of data available to help asses the measurement error. These may include replicate data, validation data, and instrumental data.

The second defining characteristic of measurement error models is determined by properties of the unobserved true values, $\mathbf{X}_i$, i = 1,2,...,n. Stefanski et al. (1995) make the distinction between classical functional models, in which the $\mathbf{X}$'s are regarded as a sequence of unknown fixed constants, and classical structural models, in which the $\mathbf{X}$'s are regarded as random variables.

The effect of measurement error on the regression coefficient in the simple linear model

$$\mathbf{y} = \beta_0 + \beta_x \mathbf{x} + \varepsilon$$

and the model 4.1, under the assumption that the $\mathbf{x}$ are random variables with var($\mathbf{x}$) $= \sigma_x^2 > 0$, is investigated fully by Fuller (1987). Assuming

$$(x, \varepsilon, U)' \sim N\left[(\mu_x, 0, 0)', diag\left(\sigma_x^2, \sigma_\varepsilon^2, \sigma_\mu^2\right)\right] \tag{4.2}$$

where $\sim$ N is an abbreviation for 'distributed normally and independently,' and diag$\left(\sigma_x^2, \sigma_\varepsilon^2, \sigma_\mu^2\right)$ is a diagonal matrix with the given elements on the diagonal.

It follows from the structural model (4.2) that the vector, $(\mathbf{y}, \mathbf{x})'$, where $\mathbf{y}$ is defined by the simple linear model and $\mathbf{x}$ is defined by model (4.1), has a bivariate normal distribution with mean vector

$$E\left[(\mathbf{y}, W)\right] = (\mu_Y, \mu_W) = (\beta_x + \beta_1 \mu_x, \mu_x)$$

and covariance matrix

$$\begin{bmatrix} \sigma_Y^2 & \sigma_{WY} \\ \sigma_{WY} & \sigma_W^2 \end{bmatrix} = \begin{bmatrix} \beta_x^2 \sigma_x^2 + \sigma_\varepsilon^2 & \beta_x \sigma_x^2 \\ \beta_x \sigma_x^2 & \sigma_x^2 + \sigma_U^2 \end{bmatrix}.$$

Let $\hat{\beta}_W$ be the regression coefficient computed from the observed variables. Then

$$\hat{\beta}_w = (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{Y}. \tag{4.3}$$

By the properties of the bivariate normal distribution

$$\begin{aligned} E[\hat{\beta}_w] &= (\sigma_W^2)^{-1} \sigma_{WY} \\ &= \frac{\sigma_x^2}{(\sigma_x^2 + \sigma_U^2)} \beta_x \\ &= \lambda \beta_x \end{aligned}$$

where $\lambda = \frac{\sigma_x^2}{(\sigma_x^2+\sigma_U^2)} < 1$. For the bivariate model with independent measurement error in $\mathbf{x}$, the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{w}$ produces an estimator that is attenuated to 0. The attenuating factor, λ, is called the reliability ratio (Fuller, 1987). The residual variance of the regression of $\mathbf{y}$ on $\mathbf{w}$ is

$$var(y/w) = \sigma_\varepsilon^2 + \frac{\beta_x^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2}.$$

As a result, the data with measurement error are more noisy, with an increased error about the line.

4.2 Methods of Estimating Inaccuracy Due to Measurement Error

Chatterjee and Hadi (1988) briefly discuss three different approaches of estimating the amount of inaccuracy introduced by measurement errors in the general linear model. These approaches include :

- the asymptotic approach,
- the perturbation approach, and
- the simulation approach.

The bootstrapping approach, which can also be used to try and quantify measurement error, is briefly discussed.

4.2.1 The Asymptotic approach

The asymptotic approach is fundamentally based on a large sample theory used predominantly in econometrics. With this approach the effects of the errors on the regression coefficients are examined as the size of the sample becomes infinitely large, i.e.,

as the sample size n approaches infinity. Consider the general linear model (2.2). The ordinary least squares (OLS) asymptotic results for (2.2) are derived as follows:

The OLS estimators can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4.4}$$

substituting (2.2) into (4.4) obtains

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

The probability limit of OLS is given as

$$plim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} plim\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right).$$

Since $\mathbf{X}$ and ε are uncorrelated, it is assumed that

$$plim\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right) = 0.$$

Hence

$$plim\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}.$$

The OLS estimator of β is therefore unbiased and consistent.

Now suppose $\mathbf{X}$ is unavailable, and the variable $\mathbf{W}$ is measured with additive error, $\mathbf{U}$ as in (4.1). Substituting (4.1) into (2.2) the model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The estimator for $\boldsymbol{\beta}$ which takes into account the measurement error is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_w &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} \\ &= \boldsymbol{\beta} + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\beta}).\end{aligned} \tag{4.5}$$

Assuming the following (Chatterjee and Hadi, 1988)

$$\begin{aligned} plim\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right) &= Var(\mathbf{W}) \\ plim\left(\frac{\mathbf{U}'\mathbf{U}}{n}\right) &= Var(\mathbf{U}) \\ plim\left(\frac{\mathbf{U}'\boldsymbol{\varepsilon}}{n}\right) &= 0 \end{aligned}$$

and

$$plim\left(\frac{\mathbf{U}}{n}\right) = 0$$

where $Var(\mathbf{W})$ and $Var(\mathbf{U})$ denote the variance covariance matrices of $\mathbf{W}$ and $\mathbf{U}$ respectively, it follows from (4.5) that the probability limit of $\hat{\boldsymbol{\beta}}_w$ is

$$plim\hat{\boldsymbol{\beta}}_w = \boldsymbol{\beta} - [\,Var(\mathbf{W})]^{-1}\,Var(\mathbf{U})\boldsymbol{\beta}.$$

Therefore

$$plim\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta} = -[\,Var(\mathbf{W})]^{-1}\,Var(\mathbf{U})\boldsymbol{\beta}. \tag{4.6}$$

It can clearly be seen that inconsistent estimates of $\boldsymbol{\beta}$ are obtained when measurement error is present in the independent variables, $\mathbf{X}$.

Furthermore, the regression coefficients for all the variables are affected, even those that are measured without error. The measurement error can be estimated by considering the norm of $\|plim\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}\|$, from which the measure of the effects of errors in measurement then becomes

$$trace\left([Var(\mathbf{W})]^{-1}Var(\mathbf{U})\right).$$

Under the assumption that $Var(\mathbf{U})$, is diagonal the effect of errors of measurement can then be assessed as

$$\sum_{i=1}^{n} r_{ii}\sigma_i^2 \tag{4.7}$$

which is known as the perturbation index, where $R = (W'W)^{-1}$, and σ_i^2 is the variance of U_i of the independent variable i with error.

As previously mentioned, the asymptotic results are for large samples with large sizes. The results do not apply to small samples.

4.2.2 The Perturbation Approach

The perturbation approach estimates measurement error by examining the effects of the changes introduced in the independent variables. This is discussed in detail by Chatterjee and Hadi (1988). A brief outline of the approach is given below.

Suppose that the model (2.2) has additive measurement error $\mathbf{U}$, as shown in (4.1). The true least-squares estimator, $\hat{\boldsymbol{\beta}}$, in (2.2) cannot be obtained and instead $\hat{\boldsymbol{\beta}}_w$ is calculated. A result connecting $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_w$ is derived as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_w + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{U}\hat{\boldsymbol{\beta}}_w - (\mathbf{W}'\mathbf{W})^{-1}\mathbf{U}'\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}\right) + o\left(\|\mathbf{U}\|\right)^2. \tag{4.8}$$

In order to study the effect of changing the independent variables equation (4.8) can be written in a different form by removing the order term and making use of

$$\mathbf{W} = \mathbf{X} + \sum_{j=1}^{k} \delta_j u_j'$$

where δ_j is the jth column of $\mathbf{U}$, and u_j is the jth unit vector

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_w + (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\sum_{j=1}^{k}\delta_j u_j'\hat{\boldsymbol{\beta}}_w - (\mathbf{W}'\mathbf{W})^{-1}\sum_{j=1}^{k}\delta_j u_j'\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}\right). \tag{4.9}$$

When the jth column of $\mathbf{W}$ is perturbed, $\mathbf{U}$ reduces to $u_j\delta_j'$. By multiplying (4.9) by v_i' the following is obtained

$$\hat{\beta}_i - \hat{\beta}_{wi} = \left\{\hat{\beta}_{wj}v_i'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' - v_i'(\mathbf{W}'\mathbf{W})^{-1}v_j\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}\right)'\right\}\delta_j \tag{4.10}$$

Letting

$$f_{ij}^2 = \left\|\hat{\beta}_{wj}v_i'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' - v_i'(\mathbf{W}'\mathbf{W})^{-1}v_j\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\beta}}\right)'\right\|,$$

and taking the norms of equation (4.10)

$$\left\|\hat{\beta}_{wi} - \hat{\beta}_i\right\| \leq f_{ij}^2\left\|\delta_j\right\|. \tag{4.11}$$

it can then be shown that

$$f_{ij}^2 = \hat{\beta}_{wj}^2 d_{ii} + d_{ij}^2 SSE. \tag{4.12}$$

This is a simple bound for the relative error of $\hat{\beta}_{wi}$. The bound obtained from (4.12) is presented as

$$\frac{\left|\hat{\beta}_{wi} - \hat{\beta}_i\right|}{\left|\hat{\beta}_{wi}\right|} \leq \frac{f_{ij}\sqrt{\|\delta_j\|}}{\left|\hat{\beta}_{wi}\right|}.$$

The variable, σ_j^2, denotes the variance of the errors of measurement in the jth variable, taking it as the norm of δ_j, the relative error in $\hat{\beta}_{wi}$ due errors in the jth variable becomes

$$\frac{\left|\hat{\beta}_{wi} - \hat{\beta}_i\right|}{\left|\hat{\beta}_{wi}\right|} \leq \frac{f_{ij}\sigma_j}{\left|\hat{\beta}_{wi}\right|}$$
$$= \frac{\sigma_j}{\left|\hat{\beta}_{wi}\right|}\sqrt{c_{ii}\hat{\beta}_{wj}^{\ 2} + c_{ij}^2 SSE}.$$

The perturbation approach does not put a restriction on the sample size of the data as does the asymptotic approach. However, the joint effects of measurement errors occurring simultaneously in all the variables is not considered with the perturbation approach.

4.2.3 The Simulation Approach

The simulation approach overcomes the limitations of the previous two methods. It is the most recommended solution and sometimes is the only available one.

The observed independent variables measured with error are represented in matrix form $\mathbf{W}$. Suppose $\mathbf{W}$ consists of k variables, and for each $\mathbf{W}_j$, $j = 1, ..., k$, there are n observations. The idea of the simulation approach is to randomly generate the error and add this error to the observed variable in order to create a new data set. Thereafter, the ordinary least squares regression equation is fitted to the new data set, and the regression coefficients thus produced are studied. A large number of new data sets are created and analysed, so that the distribution of all regression coefficients can be examined.

The observed independent variables in matrix form are

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1k} \\ w_{21} & w_{22} & \cdots & w_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nk} \end{pmatrix}.$$

The simulation approach is as follows :

1. For each W_j the random error, U_j, is generated from a probability distribution that is believed to describe the probability law of the measurement error, U_j.

2. The new data set is generated

$$\mathbf{W}_{(new)} = \begin{pmatrix} w_{11}+u_{11} & w_{12}+u_{11} & \cdots & w_{1k}+u_{11} \\ w_{21}+u_{21} & w_{22}+u_{22} & \cdots & w_{2k}+u_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1}+u_{n1} & w_{n2}+u_{n2} & \cdots & w_{nk}+u_{nk} \end{pmatrix}.$$

3. The $\mathbf{W}_{(new)}$ is regressed with $\mathbf{Y}$, and the regression coefficients

$$\beta_{1_{(new)}}, \beta_{2_{(new)}}, ..., \beta_{k_{(new)}}$$

   are obtained.

Steps 1 to 3 are repeated many times so that the distribution of the regression coefficients can be studied.

With the simulation approach the effect of the measurement error occurring simultaneously in all the independent variables can be analysed and the data set maintains its own sample size. This overcomes the two limitations of the asymptotic and perturbation approaches. The simulation approach does, however, have its disadvantages namely, that there exists the possibility that no closed-form analytic solution may exist, and that the sensitivity of the solutions for the different factors cannot be examined without further simulation.

4.2.4 The Bootstrap Approach

Two resampling procedures that can be used in the regression context are commonly known as 'bootstrapping residuals' and 'bootstrapping pairs' (Draper and Smith, 1998).

With the first procedure, the linear model is fitted and the n residuals are obtained. A sample of size n is taken from the residuals with replacement. These sample values are then attached to the n predicted responses, $\hat{y}_i$, to give a resampled set of y's. Thus, for the models $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, the new response variables become

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

where $\mathbf{e}^*$ is a resampled set from the vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Least-squares regression is now performed on the model

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

to obtain a new set of regression coefficients. The procedure can be repeated any number of times, and the usual sample mean and sample standard deviation of each of the elements of those vector estimates can be found. These results are referred to as the 'bootstrap averages' and the 'bootstrap standard errors,' respectively.

The second bootstrapping procedure is carried out on pairs $(y_i, \mathbf{x}_i')$, where y_i is the ith observation, and $\mathbf{x}_i'$ is the ith row of the $\mathbf{X}$ matrix. The resampling involves selecting a set of n of the $(y_i, \mathbf{x}_i')$, each selected with probability $1/n$, and sampling with replacement, to obtain $\mathbf{y}^{**}$ and $\mathbf{X}^{**}$. The regression model

$$\mathbf{y}^{**} = \mathbf{X}^{**}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is now fitted by least squares. As with the first bootstrap procedure, the sampling can be done a number of times, and the estimates for the regression coefficients are then studied.

4.2.5 Illustration of the Approaches

Unfortunately, the three mentioned approaches are not implemented on standard statistical software packages. As a result, SAS programmes have been written in PROC IML to carry out the analysis. In order to verify the accuracy of these programmes, the data set used is taken from Chatterjee and Hadi (1988), and also the results obtained have been checked with those from Chatterjee and Hadi (1988).

The data set used is the Health Club data which is taken from the health records of 30 employees who were regular members of a company's health club. The independent variables which are assumed to be measured with error are : weight in pounds (X_1), resting pulse rate per minute (X_2), arm and leg strength (number of pounds an employee was able to lift) (X_3), time (in seconds) in a quarter mile trial (X_4), and the response variable is the time (in seconds) in a one mile run (y).

Asymptotic results

With the asymptotic approach, the programme written in PROC IML gave the perturbation index given in (4.7) is calculated as 0.3199349. Recall that the variance covariance matrix has been assumed to be a diagonal matrix, and as a result the effects of the measurement error can be assessed as 0.3199349. By increasing the size of the sample indefinitely, the asymptotic approach gives a perturbation index which indicates that measurement error is not severe with the Health Club data. The asymptotic results (Chatterjee and Hadi, 1988) do not provide information on individual components but rather they provide norms which bound the whole vector. A large norm might have concealed the fact that all the components are small except one. The perturbation approach aims to overcome these difficulties.

Perturbation results

The regression analysis of the Health Club data is given in Table 4.1, and the results from the PROC IML programme for carrying out the perturbation approach are presented in Table 4.2.

The analysis of the entries in Table 4.2 indicates that, if the estimates of the standard deviation are small, then the estimate of the intercept is reliable up to one significant digit. The regression coefficients of all the independent X variables are reliable up to two significant digits.

Table 4.1: Regression results for Health Club data

Regression Statistics

Multiple R	0.9236407
R Square	0.8531121
Adjusted R Square	0.82961
Standard Error	28.671492
Observations	30

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Regression	4	119360.5052	29840.13	36.29945	4.50993E-10
Residual	25	20551.36148	822.0545		
Total	29	139911.8667			

Parameter Estimates

Variable	**Coefficients**	**Standard Error**	**Standard t Stat**	**P value**
Intercept	-3.618564	56.10266855	-0.0645	0.949086
X1	1.2676336	0.286871192	4.418825	0.000168
X2	-0.525206	0.862785992	-0.60873	0.548194
X3	-0.505022	0.245920001	-2.0536	0.050614
X4	3.9030122	0.747711459	5.219944	2.11E-05

Table 4.2: Bounds for the relative errors in the regression coefficients

		Relative error in reg coef				
error in	**std dev**	**const**	**X1**	**X2**	**X3**	**X4**
X1	0.289	21.8	0.4	2.3	0.8	0.3
X2	0.289	43.8	0.4	7.2	0.9	0.3
X3	0.289	13.3	0.2	1.1	0.7	0.1
X4	0.289	62.9	0.9	6.8	1.9	1

Simulation results

For the simulation approach 1000 data sets were generated in PROC IML, and the histograms of the estimated regression coefficients were drawn using PROC UNIVARIATE.

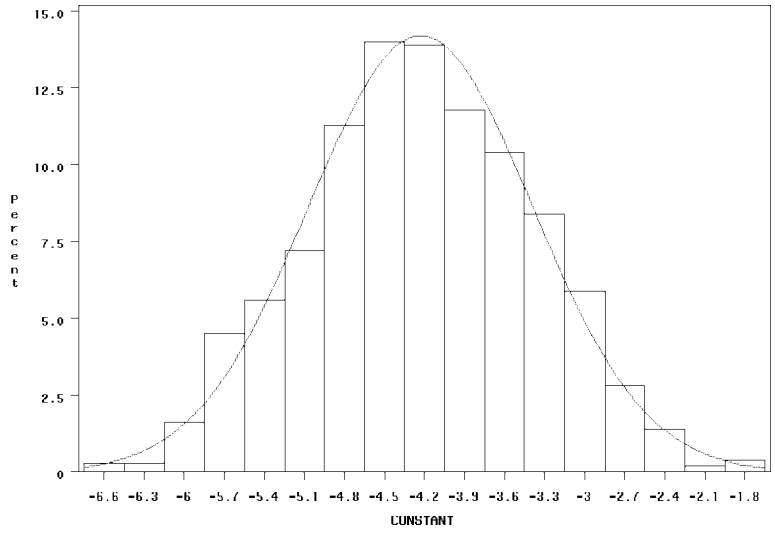

Figure 4.1: Histogram of regression coefficient for the constant.

The ideal shapes of the simulated regression coefficient values should roughly be normal. There should not be many outlying or extreme values, skewness, peaks etc. Also, the histograms should not pass through zero because this would indicate insignificance of the coefficients. By carefully examining the histograms given in Figures 4.1 to 4.5, one can see no unusual behaviour with the coefficients as the randomly generated error from a uniform distribution, with mean 0 and standard deviation 0.2887, is added to the original data 1000 times to produce 1000 values for each coefficient. As a result, the simulation approach does not indicate that measurement error plays a significant role with the Health Club data.

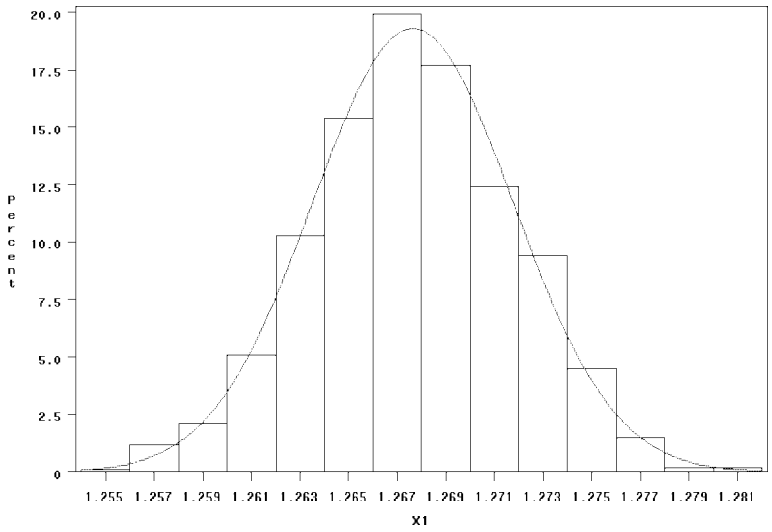

Figure 4.2: Histogram of regression coefficient for X_1.

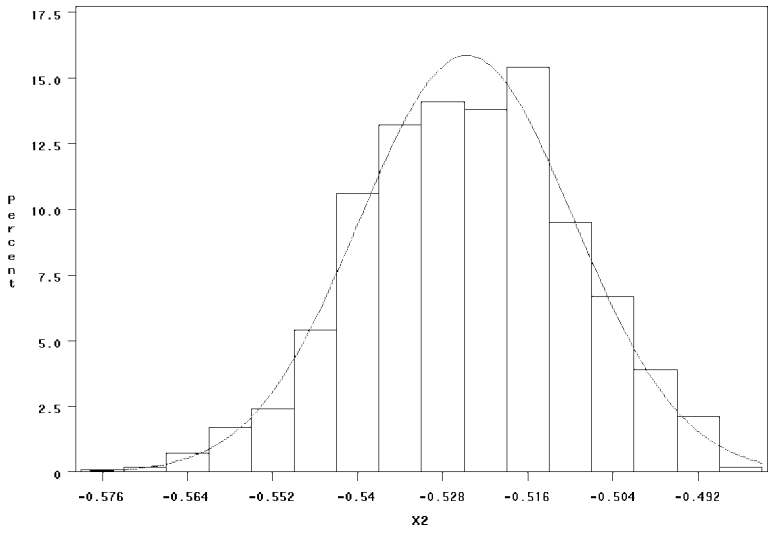

Figure 4.3: Histogram of regression coefficient for X_2.

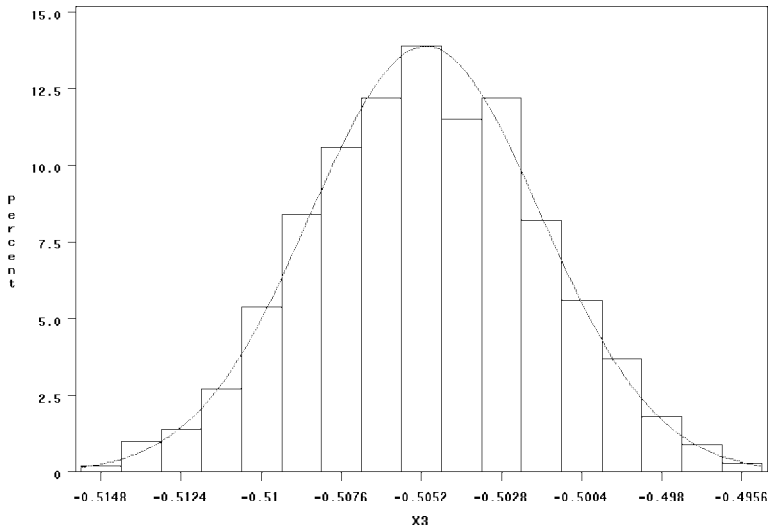

Figure 4.4: Histogram of regression coefficient for X_3.

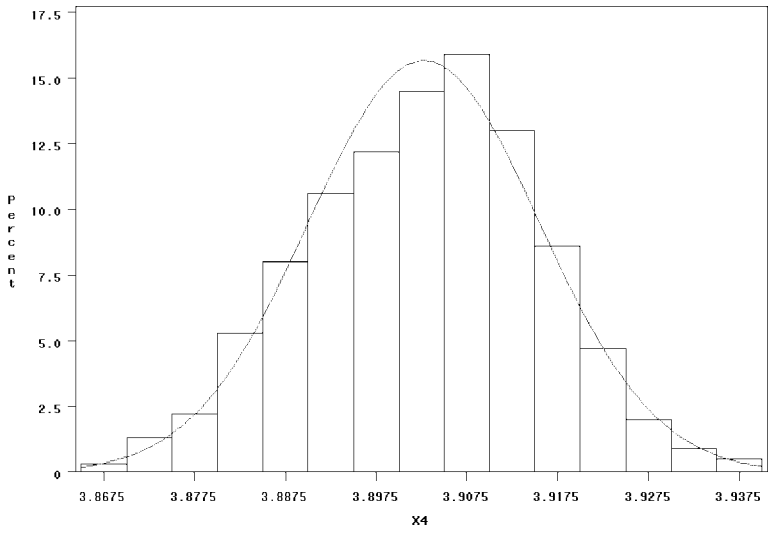

Figure 4.5: Histogram of regression coefficient for X_4.

Bootstrapping results

The bootstrap approach is applied to the Health Club data, in which 1000 data set is sampled with replacement in PROC IML, as explained in Section 3.2.4. The following histograms were drawn using PROC UNIVARIATE :

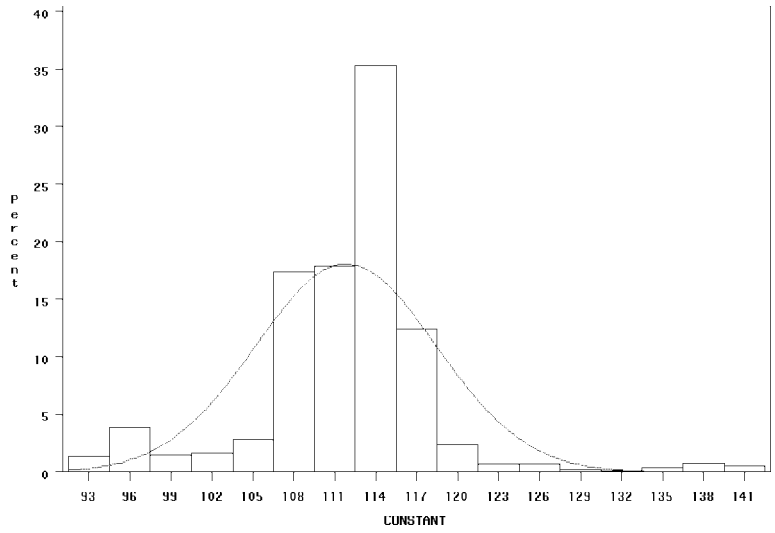

Figure 4.6: Histogram of regression coefficient for the constant using bootstrap approach.

As with the simulation results, the histograms here do not indicate any major deviations in the consistency of the coefficients. The histograms all appear normally distributed. As a result, it is concluded that measurement error is not significant with regard to the bootstrap method.

4.3 Simulation Extrapolation

Simulation extrapolation (SIMEX) (Carroll et al., 1995) is a simulation-based method of estimating and reducing bias due to measurement error. The SIMEX estimate is obtained firstly by simulating a large number of data sets where each data set has a

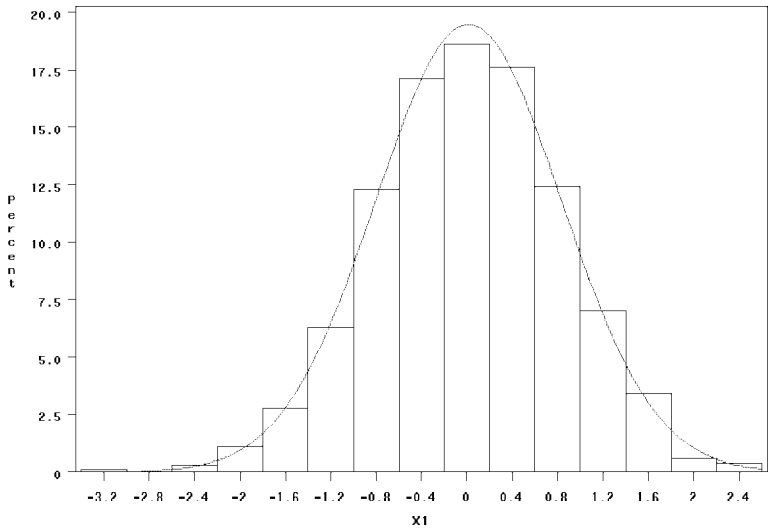

Figure 4.7: Histogram of regression coefficient for X_1 using bootstrap approach.

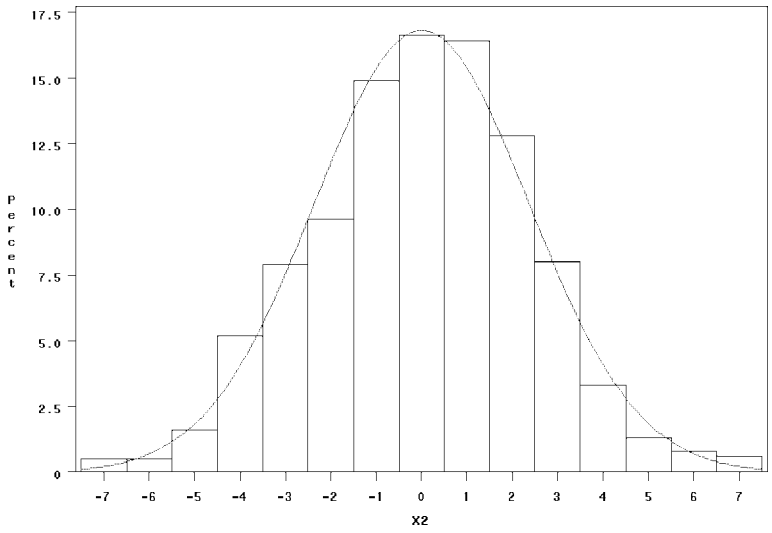

Figure 4.8: Histogram of regression coefficient for X_2 using bootstrap approach.

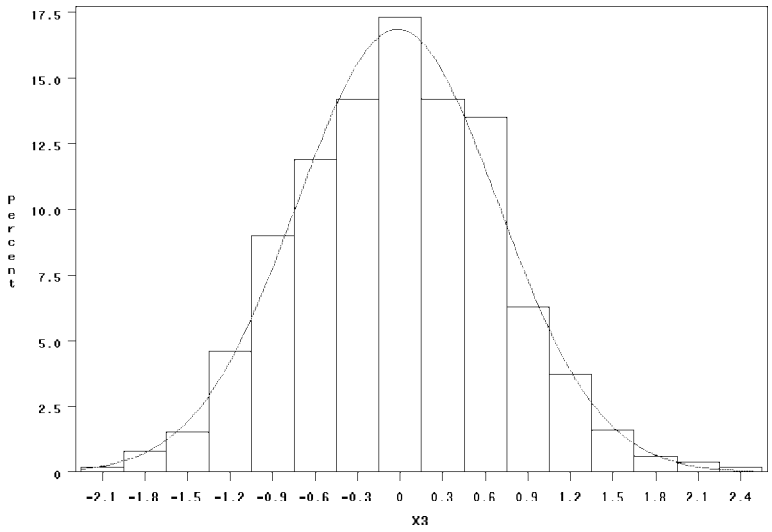

Figure 4.9: Histogram of regression coefficient for X_3 using bootstrap approach.

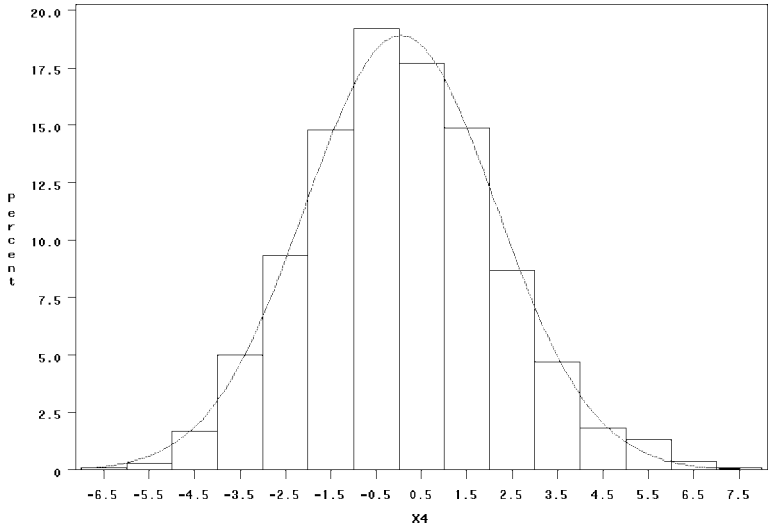

Figure 4.10: Histogram of regression coefficient for X_4 using bootstrap approach.

larger measurement error variance than the previous one. Thereafter, a graph showing the trend of measurement error-induced bias against the variance of the added measurement error is used to extrapolate to the case of no measurement error.

The basic idea of the SIMEX algorithm is clearly defined in the context of the simple linear regression model given as

$$\mathbf{y} = \beta_0 + \beta_x \mathbf{X} + \boldsymbol{\varepsilon} \tag{4.13}$$

with additive measurement error as shown in (4.1). Recall the least-squares estimate of β_x, which can be obtained from the observed variables W, is

$$\hat{\beta}_w = \frac{\hat{\beta}_x \sigma_x^2}{(\sigma_x^2 + \sigma_u^2)}.$$

The above estimate is calculated from the original data set. Suppose that in addition to that data set there exists $M - 1$ other data sets, each having increasingly larger measurement error variance

$$(1 + \lambda_m)\, \sigma_u^2$$

where $0 = \lambda_1 < \lambda_2 <, ..., < \lambda_M$. For the mth data set the estimate of β_m would be

$$\hat{\beta}_{w_m} = \frac{\beta_x \sigma_x^2}{[\sigma_x^2 + (1 + \lambda_m)\, \sigma_u^2]}.$$

Consider the function

$$G(\lambda) = \frac{\beta_x \sigma_x^2}{[\sigma_x^2 + (1 + \lambda)\, \sigma_u^2]}$$

where $\lambda \geq 0$. Extrapolating the above function to $\lambda = -1$, will yield the estimate

$$G(-1) = \beta_x.$$

With SIMEX the above approach is imitated as follows:

1. *Simulation step*: Each data set created has a larger measurement error variance $(\lambda_m \sigma_u^2)$ than the previous. The mth data set will have a total measurement error variance of

$$\sigma_u^2 + \lambda_m \sigma_u^2 = (1 + \lambda_m)\sigma_u^2.$$

2. *Re-estimation step*: Estimates of β_x are obtained from each of the resulting data sets.

3. Steps (1) and (2) are repeated a large number of times, and the average value of the estimates of β_x for each m is calculated.

4. These averages are plotted against the λ values, and extrapolant techniques are used to obtain the SIMEX estimate ($\lambda = -1$).

4.3.1 The SIMEX Algorithm for Simple Linear Regression

The basic SIMEX algorithm applies to the simple linear model in which only one single, scalar, predictor X is subject to additive error as in equation (4.1). The algorithm involves the following steps:

1. With the simulation step for any $\lambda \geq 0$, define

$$W_{b,i}(\lambda) = W_i + \sqrt{\lambda} U_{b,i}$$

where $i = 1, ..., n$, $b = 1, ..., B$, and $\{\mathbf{U}_{b,i}\}_{i=1}^{n}$ are the computer generated pseudo-errors which would follow the distribution believed to be that of the measurement error.

2. Once the values of the predictor have been generated, $\hat{\beta}_{x_b}(\lambda)$ is estimated and the average of these estimates are calculated as

$$\hat{\beta}_x(\lambda) = B^{-1} \sum_{b=1}^{B} \hat{\beta}_{x_b}(\lambda).$$

The estimate, $\hat{\beta}_x(\lambda)$, is the sample mean of $\left\{\hat{\beta}_{x_b}(\lambda)\right\}_1^B$, and is therefore the average of the estimates obtained from a large number of experiments with the same amount of measurement error.

3. The simulation component of SIMEX involves plotting the points $\left\{\hat{\beta}_{\mathbf{x}}(\lambda_m), \lambda_m\right\}_1^M$ on a graph.

4. The extrapolation step entails modelling each of the components of $\left\{\hat{\beta}_x(\lambda)\right\}$ as functions of λ for $\lambda \geq 0$, and extrapolating the fitted models back to $\lambda = -1$. The extrapolated values yield the SIMEX estimate, $\hat{\beta}_{x_{simex}}$ of β_x.

4.3.2 SIMEX with Multiple Linear Regression

In order to illustrate the SIMEX approach with the Health Club data, the special case of SIMEX with multiple linear regression is considered.

The multiple linear regression model is defined as

$$y_i = \beta_1 + \beta_z' Z_i + \beta_x X_i + \varepsilon_i. \tag{4.14}$$

The variable, Z_i, denotes the independent variables measured without error, and X_i denotes the independent variable measured with error. From equation (4.14) it is clearly seen that this special case applies strictly to the model in which multiple independent variables are error free, whereas only one independent variable is measured with error.

Let $\beta = (\beta_1, \beta_z', \beta_x)'$, from the simulation step the SIMEX estimate becomes

$$\hat{\beta}(\lambda) = \left\{ \sum_{i=1}^{n} \begin{pmatrix} 1 & Z_i^t & W_i \\ Z_i & Z_i Z_i^t & Z_i W_i \\ W_i & W_i Z_i^t & W \end{pmatrix} \right\}^{-1} \times \left\{ \sum_{i=1}^{n} \begin{pmatrix} y_i \\ Z_i y_i \\ W_i y_i \end{pmatrix} \right\}.$$

Solving the system of equations it is found that

$$\hat{\beta}_v(\lambda) = (V^t V)^{-1} V^t y - \frac{(V^t V)^{-1} V^t W (W^t y - W^t V (V^t V)^{-1} V^t y)}{W^t W - W^t V (V^t V)^{-1} V^t W + \lambda \sigma^2}$$

$$\text{and } \hat{\beta}_x(\lambda) = \frac{W^t y - W^t V^t V)^{-1} V^t y}{W^t W - W^t V^t V)^{-1} V^t W + \lambda \sigma^2}$$

where $\beta_v = (\beta_1, \beta_z^t)^t, V^t = (V_1, V_2, ..., V_n)$, with $V_i = (1, Z_i^t)^t$.

4.3.3 Illustration of the SIMEX Method

With the Health Club data it is now assumed that the variable $\mathbf{X}_4$ is measured with error so that SIMEX results in the case of multiple linear regression can easily be

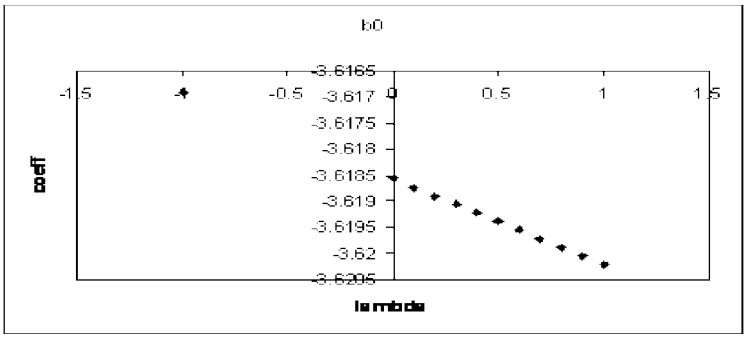

Figure 4.11: Coefficient extrapolation for β_0.

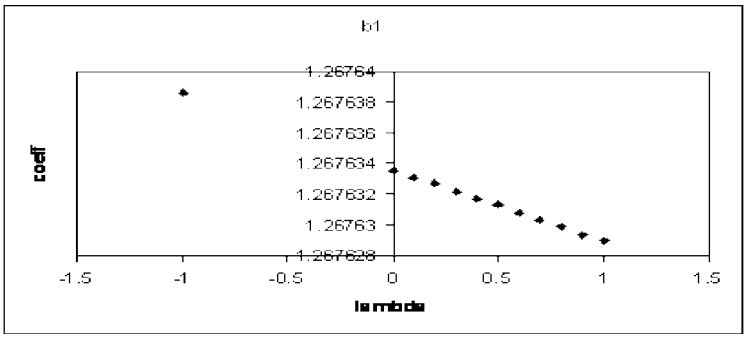

Figure 4.12: Coefficient extrapolation for β_1.

demonstrated. A simple programme written in PROC IML is used to carry out the analysis. The graphs are obtained for each of the regression coefficients. These graphs are presented in Figures 4.11 to 4.15. All graphs clearly show the SIMEX estimate obtained for each coefficients.

4.3.4 Fitting GLMs using SIMEX with Additive Measurement Error

Cook and Stefanski (1994) present a class of nonlinear models that is exact for many linear regression models with least squares estimation which is the generalized least squares estimation of an exponential mean model. They assume that the observed

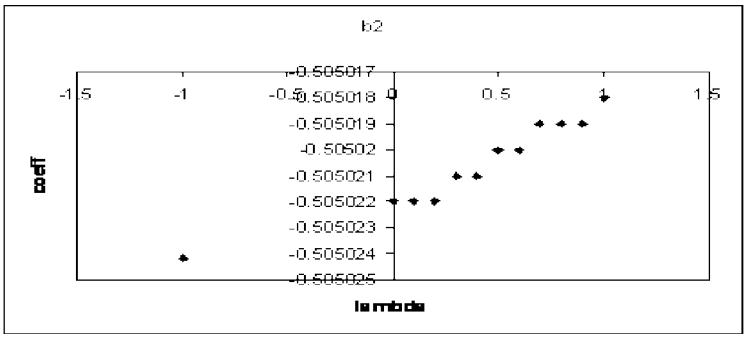

Figure 4.13: Coefficient extrapolation for β_2.

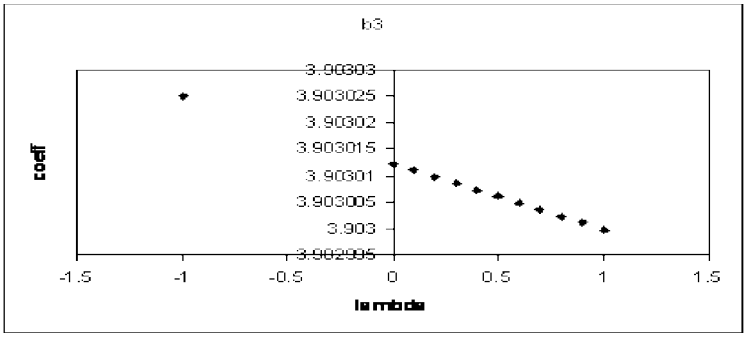

Figure 4.14: Coefficient extrapolation for β_3.

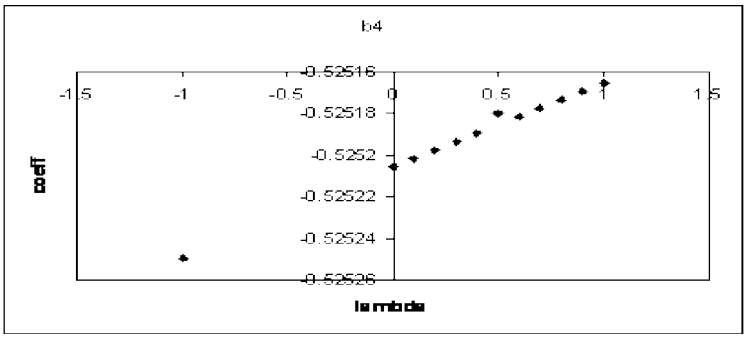

Figure 4.15: Coefficient extrapolation for β_4.

data Y_i, Z_i, W_i for $i = 1, ..., n$ are such that

$$W_i = X_i + \sigma U_i$$

where σ^2 is the known measurement error variance.

Cook and Stefanski (1994) assume that X is a scalar, and that $E(Y/X) = \exp(\beta_1 + \beta_X X)$, with conditional variance $V(Y/X) = \kappa^2 \exp(\tau\beta_1 + \tau\beta_X X)$ for some constants κ and τ. If (W, X) has a bivariate normal distribution and generalized least squares is the method of estimation, then $\hat{\beta}_1(\lambda)$ and $\hat{\beta}_U(U)$ consistently estimate

$$\beta_1(\lambda) = \beta_1 + (1 + \lambda)\frac{\mu_X\sigma^2\beta_X + \beta_X^2\sigma_X^2\sigma^2/2}{\sigma_X^2 + (1 + \lambda)\sigma^2}$$

and

$$\beta_X(\lambda) = \frac{\beta_X\sigma_X^2}{\sigma_X^2 + (1 + \lambda)\sigma^2}$$

where $\mu_X = E(X), \sigma_X^2 = V(X)$, and $\sigma^2 = V(W/X)$. Cook and Stefanski (1994) point out that the nonlinear extrapolant is clearly asymptotically exact for estimating β_X, and that the same holds for β_1. It is further said that the parameters in the nonlinear extrapolant function bear little resemblance to their counterparts in the linear model.

Carroll et al. (1995) fit the generalized linear model with additive normally distributed measurement error by means of two functional methods referred to as the 'conditional-score' and 'corrected-score' methods. In the case of the multiple linear regression model the corrected-score is equivalent to the usual method of moments regression estimator.

Chapter 5

Durban South Data Analysis

5.1 Introduction

The Durban South data which are described in detail below are first analysed assuming thatit fits the linear regression model. It cannot be said with certainty that this model is appropriate for the data. The assumptions associated with the model need to be checked. The aptness of the model for the data is examined before inferences based on the model are made.

In this chapter a full description of the Durban South data is given, followed by analysis of the data which includes the various diagnostic checks associated with the linear regression model. The diagnostics consist of basic graphical techniques, outliers, and influential detections. Some remedial measures are considered.

5.2 The Data Set and Objective of the Study

The data set used in the analysis was obtained from the Durban South health survey. The area is located in the Durban South Industrial Basin and as a result is exposed to air pollution emitted by numerous industries and refineries. The study was conducted on children from various schools in the area.

What is of primary interest with the Durban South data set is whether or not a relationship exists between the respiratory condition of the children involved and their individual characteristics which include weight, height, gender and their asthmatic status.

5.2.1 Basic Biological Background Information

Disorders of the respiratory system and respiratory disease are very common in children. Children suffer from various respiratory disorders and these are known as acute respiratory infections (ARI) which are responsible for 20 percent of the deaths of children under the age of five years in South Africa, according to Coovadia and Wittenberg (2000). ARI is also the leading cause of death of children living in the third world. Asthma and pneumonia constitute a few of the respiratory disorders, and acute pneumonia is responsible for 90 percent of the child fatalities. Respiratory disease is considered extremely dangerous since more than half the total lung tissue (or function) may be lost before an individual complains of symptoms such as dyspnea, coughing and chest pain, or displays signs such as tachypnea, rales, and wheezing. These symptoms and signs may be overlooked or may be subtle in young children.

The anatomy of the respiratory system is explained by Behrman and Kliegman (1990).

Pulmonary function testing (PFT) is used to assess the functional status of the lungs. The test is used to determine how much volume of air and how fast the air in the lungs can be moved in and out of the lungs. It also relates to how stiff the lungs and chest wall are, the diffusion characteristics of the membrane through which the gas moves (determined by special tests), and how the lungs respond to physical therapy procedures to the chest.

Static lung volumes reflect the elastic properties of the lungs and chest wall. Vital capacity (VC) is the maximum volume of air that can be expired slowly after a full inspiratory effort. Simple to perform, the test is one of the most valuable measurements of pulmonary function. Because VC decreases as a restrictive lung disorder worsens, it

can be used along with the diffusing capacity to follow the course of such a disorder and its response to therapy. The VC also reflects the strength of the respiratory muscles and is often used to monitor the course of neuromuscular disorders.

Forced vital capacity (FVC), similar to VC, is the volume of air expired with maximal force. FVC is usually measured along with expiratory flow rates in a simple spirometer.

Dynamic lung volumes reflect the calibre and integrity of the airways. A spirometer records the lung volume against time during an FVC manoeuvre. Forced expiratory volume in one second (FEV_1) is the volume of air forcefully expired during the first second after a full breath, and normally accounts for $> 75\%$ of the FVC. The value can also be obtained by using the AirWatch device which is a hand-held digital monitor that measures and displays the rate of the airflow.

Prolongation of expiratory flow rates is increased by bronchospasm (in asthma), impacted secretions (in bronchitis), and loss of lung elastic recoil (in emphysema). In fixed obstruction of the upper airway, flow is limited by the calibre of the narrowed segment rather than by dynamic compression, resulting in equal reduction of inspiratory and expiratory flow rates.

In restrictive lung disorders, increased tissue elastic recoil tends to maintain the calibre of the larger airways so that at comparable lung volumes, flow rates are often higher than normal. (Tests of small airways function, however, may be abnormal.)

There are a few variables such as gender, weight, and height which have an impact on the lung function of one individual compared to another.

Pelosi, Caironi, and Gattinoni (2002) define obesity as a metabolic disease in which adipose tissue represents a proportion of body mass tissue greater than normal. Many actiologic factors may be implicated in determining obesity, such as genetic, environmental, socio-economic, and individual factors, such as age and gender. In normal conditions, adipose tissue represents 15% to 18% of body weight in males, and about 25% in females.

A number of indices have been developed for the quantification of obesity. These

include height/weight tables, calculation of the ratio between the actual and the 'ideal' weight of an individual and the calculation of the body mass index (BMI). The 'ideal' weight expressed in kilograms is computed by subtracting 100 (in men) and 105 (in women) from the individuals height expressed in centimetres. A person weighing more than 120% of 'ideal' weight may be considered obese, and greater than 200% of 'ideal' weight as pathologically obese.

The most widely used measurement for defining obesity is the BMI, calculated as weight/height^2 (kg/m^2). This index is often used to monitor childhood obesity. The index is, however, known to mislabel a healthy child as overweight or obese (Ellis, Abrams, and Wong, 1999).

Ulger et al. (2006) study the effect of childhood obesity on respiratory function tests and airway hyperresponsiveness in a cross-sectional study. In their study, FEV_1/FVC ratios of the study and control groups were similar. This meant both respiratory resistance and airway resistance rise significantly with the level of obesity. These findings suggest that in addition to the elastic load, obese subjects have to overcome increases in respiratory resistance resulting from the reduction in lung volumes which is related to being overweight.

Usually, the lung volumes and capacities of males are larger than the lung volumes and capacities of females. Even when males and females are matched for height and weight, males are found to have larger lungs than females. Therefore, there exists a gender-dependent lung size difference.

Asthma is characterized by spastic contraction of the smooth muscles in the bronchioles, which causes extremely difficult breathing (Guyton and Hall 2000). It is a common childhood condition. The asthma status of an individual can be determined by various lung function testing techniques available, amongst these are the FEV_1 test.

Berhane et al. (2000) study the effects of gender and asthma on the pulmonary function in children. They conclude that large deficits with regard to flow rates are seen in children who have had asthma for a longer time. Large deficits in flow rates in both

large and small airways were also observed in males and females for whom asthma was reported to have been diagnosed before the age of 3 years.

Studies in children and adolescents have shown that the level of lung function is related to height (Nasirumbi, 2006). In general, it was found that the taller a person the larger his or her lung volume was, and the greater the FEV_1 reading.

The Durban South data which are part of a health survey were obtained from communities in the Durban South and the northern residential areas. These areas were included in the study because of their exposure to various pollutants. One source of the pollution was sulphur dioxide being released from the numerous industries in and around the Durban South region. As a result of this area being heavily industrialized there is much pollution released. What is of interest in this thesis is how the respiratory health of the children that have been included in this study who live in and around these polluted areas is being influenced by their individual weight, height, gender and asthma status.

Data were collected from five primary schools. A measure of the lung function of each child was obtained using FEV_1 values. The original data set consisted of five measures of FEV_1 values obtained for each child, four times a day, over fourteen school days and was simplified to fourteen FEV_1 values per child by using a within-day variability formula (Nasirumbi, 2006).

The only data made available for the present study includes the weight, height, sex, asthma status, the average of the 14 FEV_1 values, and the corresponding standard deviation of the average FEV_1 values, for 225 children.

5.3 Analyses and Results

The analysis is begun with an attempt to fit the linear model. When a regression model such as the linear regression model, with or without weights, is considered for any application, one cannot be certain in advance that the model is going to represent

the data well. The adequacy of the model should be checked by means of the various diagnostics tools.

The weighted linear model is discussed in Chapter 1. The vector of responses is the average of the 14 FEV_1 values obtained from 225 children. This variable is denoted as *FEV1*. The predictor variables are *weight, height, sex, statusL* and *statusM*. Weight, height and sex denote the weight, height and sex of each child respectively. If a child has a normal asthmatic status then this is denoted by *statusL*, if a child has a mild asthmatic level then this is denoted by *statusM*. If the child has neither normal asthmatic nor mild asthmatic status then he/she is regarded as a persistent asthmatic. The data is weighted by the inverse variance of the average 14 FEV_1 values. The fit results are presented in Table 5.1.

Table 5.1: Results for Weighted Linear Regression Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10.20603	2.04121	2.93	0.0140
Error	219	152.64158	0.69699		
Corrected Total	224	162.84761			

Root MSE	0.83486	R-Square	0.0627
Dependent Mean	0.49964	Adj R-Sq	0.0413
Coeff Var	167.09421		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t value	Pr > |t|
Intercept	1	0.44848	0.15302	2.93	0.0037
Weight	1	0.00086670	0.00106	0.82	0.4134
Height	1	0.00976	0.12904	0.08	0.9398
Sex	1	0.03352	0.01331	2.52	0.0125
StatusL	1	-0.03449	0.01703	-2.02	0.0441
StatusM	1	0.00597	0.01531	0.39	0.6972

The results indicate that the overall fit of the model is significant, with F-prob = 0.0140. The parameter estimates *weight, height*, and *statusM* are insignificant at alpha level 0.05. Parameter estimates *sex* and *statusL* are significant. The R-Square statistic is 0.0627. Therefore, only 6.27% of the amount of variation about the mean is explained

by the weighted linear regression model.

Before making the inferences about parameters of interest, it is important to examine the plots of residuals and other diagnostics. The plot of studentized residuals vs the predicted values are presented in Figure 5.1.

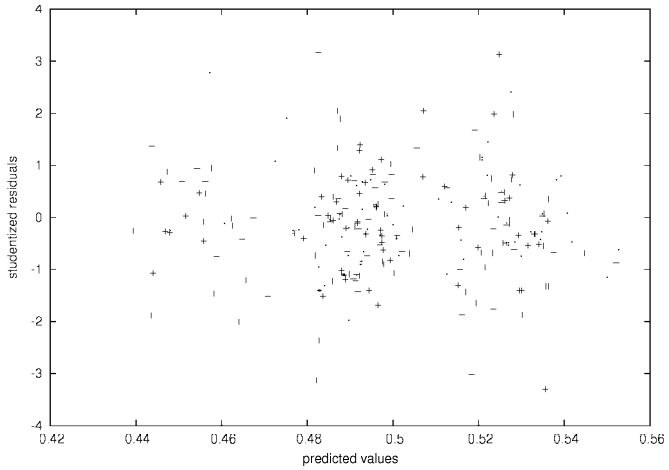

Figure 5.1: The graph of studentized residuals vs predicted values.

The residual plot appears satisfactory since it has a random scatter of points. It does not seem to depict nonlinearity of the mean response ($FEV1$) and nonconstant error variance.

The diagnostic measures discussed in the chapter 3 are applied to the weighted linear regression analysis. High leverage, Cook's distance, the DFFITS statistic, COVRATIO, as well as the outlying studentized residuals are used. Of the 225 students in the study, observations 8, 129, 12, 138, 65, 95, 44, and 174 are considered outlying and / or influential.

For the Durban South data an observation whose leverage exceeds 0,044444444 is considered a high leverage point. The leverage values of the observations are graphically represented in Figure 5.2. Observations which lie above the line indicate high leverage. Approximately 32 observations are classified as high leverage.

The measures of influence include Cook's distance. For these data, if the calculated

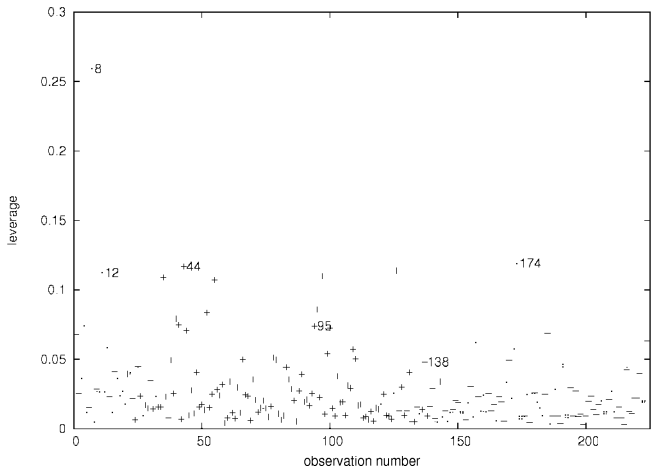

Figure 5.2: Leverage values for the weighted linear regression model.

Cook's distance exceeds 0.01826484, the corresponding observation is influential. In total, 22 of the 225 observations have a high Cook's distance; this total is roughly 10% of observations. Figure 5.3 presents a summary of the Cook's distances obtained for the data set. Observations which lie above the line have a high Cook's distance, and the labelled observations are those that have been classified as extreme observations.

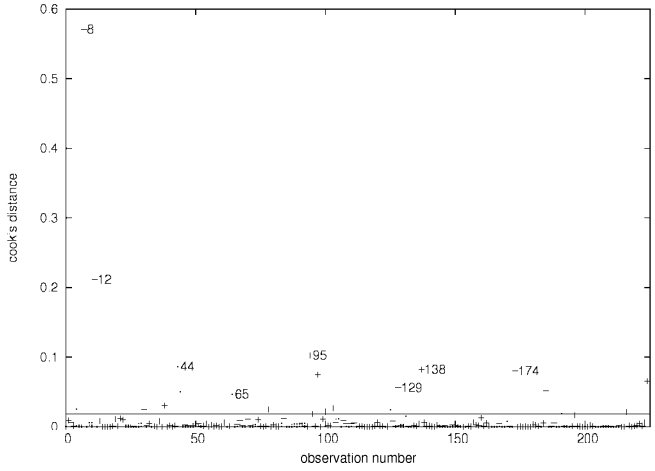

Figure 5.3: Cook's distance for the weighted linear regression model.

Another measure of influence is the DIFFIT's statistic. With the Durban South data

any observation whose DIFFIT's statistic is bigger than 0.298142395 in absolute value can be considered to be influential. The graphical summary of the DIFFIT's statistics are given in Figure 5.4. In total, 24 observations have DIFFIT's statistics which are an indication of high influence. From Figure 5.4, it can be seen that the observations which fall within the two lines are not influential according to the DIFFIT's statistic.

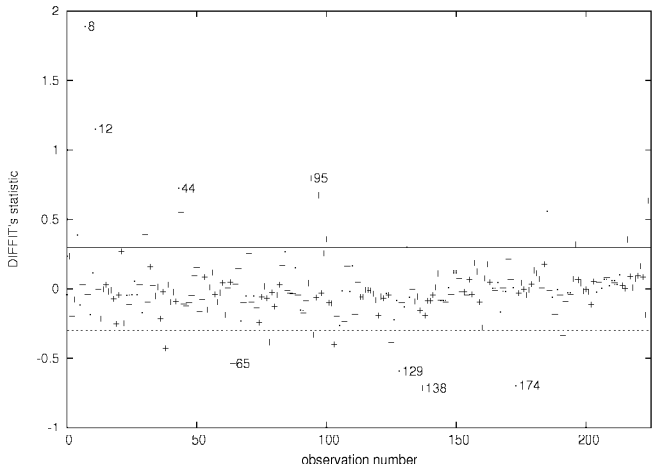

Figure 5.4: DIFFIT's statistics for the weighted linear regression model.

Recall the COVRATIO statistic is used as a measure of model performance which falls under measures of influence. An observation with a COVRATIO value that is less than one is indicative that the corresponding data point does not contribute to model performance. The summary of the COVRATIO statistics for the 230 observations is given in Figure 5.5. From Figure 5.5, it can be seen that 31 observations have COVRATIO value that is less than 1.

In general, the intersection of all problematic sets consists of observations 8, 129, 12, 138, 65, 95, 44, and 174. These are the most outlying and/or influential cases. To assess the effect of these extreme observations, each of them is deleted one by one to see how much influence the observations have on the overall performance of the model.

When observations 65, 129, and 95 are sequentially deleted, the R-square statistic, as well as the overall model significance increase slightly. When observations 8, 12, 138,

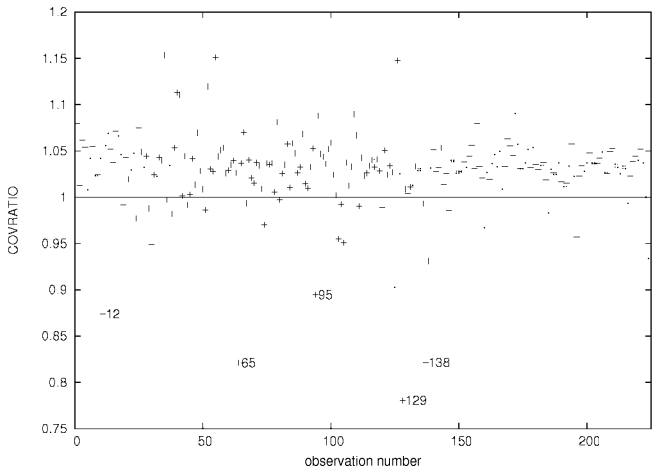

Figure 5.5: COVRATIO statistics for the weighted linear regression model.

44, and 174 are deleted, in general, the F-prob and R-square statistics decrease slightly. Significance of parameter estimates does not vary drastically.

Out of the 8 observations the most change occurs when observations 8, 12, 95, and 129 are deleted. These observations were deleted in pairs. The deletion of the paired observations did not have a huge effect on the overall model performance or significance of parameter estimates. In general, when observations 12 and 129, 12 and 95, as well as 129 and 95 were deleted, the R-square statistic reached its highest value of around 0.0986. The R-square statistic is far too small, indicating that variation is not being explained by the predictor variables. The predictor variables *sex* and *asthmaL* remain significant for most deletions. Other than these, the deletion of individual and paired observations did not result in a better fit.

As mentioned in Chapter 3, a probability plot is a graph of the studentized residuals against the quantiles of a user-specified distribution. Usually, the goal of constructing a probability plot is to visually (subjectively) evaluate the null hypothesis that the data are well fitted by the specified distribution. Frequently, the null hypothesis is that the data set has a normal distribution. If the graph of plotted points in a probability plot appears linear to the eye, and with little scatter or deviation about the line, it can be

concluded that the data appear to be from the specified distribution. If the plotted points do not approximate a straight line, the type of departures from linearity provide information about how the actual data distribution deviates from the hypothesized distribution. Figure 5.6 shows a probability plot for the studentized residuals. The null hypothesis used to obtain this plot was the errors in the response $FEV1$ follow the normal distribution. However, the plotted points are not well fitted by a straight line at the tail ends.

The normal probability plot not only determines if the data are approximately normally distributed, but also offers insight into how the data are distributed if they are not normally distributed. This insight is helpful in deciding what transformation to use to induce normality (Zewotir and Galpin, 2004).

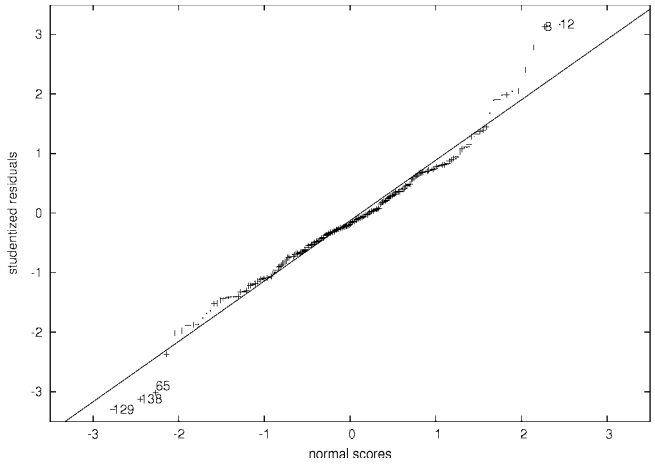

Figure 5.6: Normal probability plot for the weighted linear regression model.

Table 5.2: Tests for normality for the linear regression model

Test	Statistic	p-value
Shapiro-Wilk	W 0.98294	(Pr<W) 0.0082
Kolomogorov-Smirnov	D 0.052311	(Pr>D) 0.1362
Cramer-von Mises	W-sq 0.120637	(Pr>W-sq) 0.0616
Anderson-Darling	A-sq 0.857891	(Pr>A-sq) 0.0275

The results of the formal tests for normality of the errors are given in Table 5.2. The

Kolomogorov-Smirnov and the Cramer-von Mises tests indicate that the normality assumption is valid, whereas the remaining tests do not. Violation of normality could be due to the presence of outliers (Zewotir and Galpin, 2004).

Since observations 8, 12, 65, 129, and 138 are seen to stand out in Figure 5.6, these observations are deleted from the data to see if and how they influence normality. Individual deletion of extreme observations 8 and 12 result in the probability plot not appearing much different from that shown in Figure 5.6. The only difference is that observations have been deleted from Figure 5.6. Conclusions of the formal tests for normality are the same as those from Table 5.2.

Deletion of the extreme observation 12 seems to cause a distinct change in the conclusions of the formal tests for normality. The results for these tests are presented in Table 5.3, and the corresponding normal probability plot is shown in Figure 5.7.

Table 5.3: Tests for normality when obs 12 is deleted

Test	Statistic	p-value
Shapiro-Wilk	W 0.98294	(Pr<W) 0.0777
Kolomogorov-Smirnov	D 0.052311	(Pr>D) >0.1500
Cramer-von Mises	W-sq 0.120637	(Pr>W-sq) 0.2137
Anderson-Darling	A-sq 0.857891	(Pr>A-sq) 0.1306

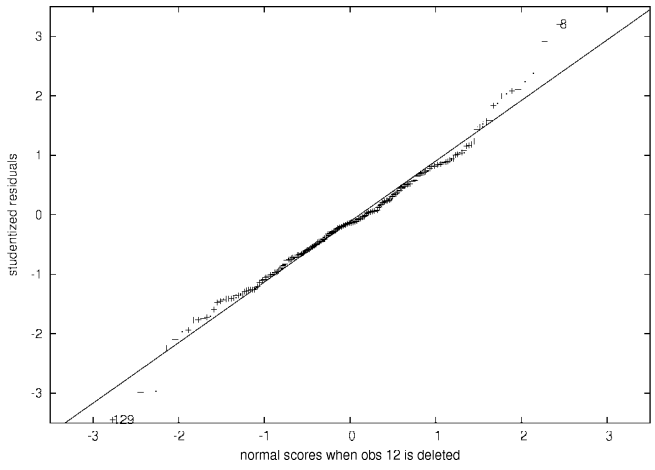

Figure 5.7: Normal probability plot when obs 12 is deleted.

All normality tests clearly indicate that, without the extreme observation number 12, the normality assumption is not violated. However, the probability plot given in Figure 5.7 does not seem to appear much different to that obtained for the full data as shown in Figure 5.6. It is closer to the straight line in the center but it still has a few outliers. Observations 65 and 138 are sightly less outlying, and observations 8 and 12 appear more outlying than they do in Figure 5.6.

When the extreme observation 129 is deleted, only the Shapiro-Wilk test for normality indicates that the data is not normal. The remaining tests indicate normality but the normality is not as strong as that when observation 12 is deleted. The resulting probability plot obtained here strongly resembles that shown in Figure 5.6, and this occurs also when observation 138 is deleted. However, the formal tests for normality when observation 138 is deleted are similar to those shown in Table 5.2.

Since the deletion of extreme observation 12 induces normality according to the normality tests, and when this observation is deleted in a pair with any of the other extreme observations, the resulting normality tests also favour the normality assumption. Paired observations 12 and 8, 12 and 65, 12 and 129, as well as 12 and 138 are deleted from the data. With the probability plots the residuals become slightly closer together and two of the most extreme observations are removed from the data set and as a result, the tests for normality indicate normality. The most normal data according to the normality test are obtained when extreme observations 12 and 129 are deleted. The results for the normality tests are presented in Table 5.4 and the corresponding normal probability plot is given in Figure 5.8.

Table 5.4: Tests for Normality when obs 12 and 129 are deleted

Test	Statistic	p-value
Shapiro-Wilk	W 0.991218	(Pr<W) 0.1991
Kolomogorov-Smirnov	D 0.046175	(Pr>D) >0.1500
Cramer-von Mises	W-sq 0.059434	(Pr>W-sq)>0.2500
Anderson-Darling	A-sq 0.447008	(Pr>A-sq)>0.2500

From the normality plot presented in Figure 5.8 it can be seen that, aside from the

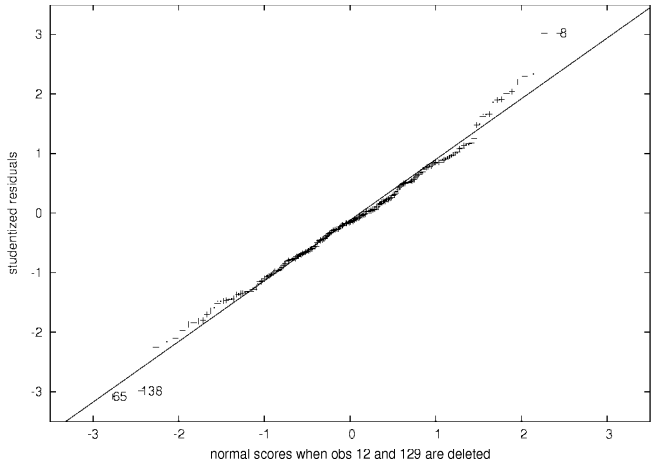

Figure 5.8: Normal probability plot when obs 12 and 129 are deleted.

extreme observations 8, 65, and 138, most of the data fall very close to the straight line. The data that appear to deviate the most from the straight line in Figure 5.8 are those found toward the right-hand tail.

When the remaining extreme observations 8, 65, 129, and 138 are deleted in pairs the results obtained do not significantly deviate from the above mentioned results for deletion of individual observations with regard to the normality plots and formal tests of normality.

To further analyse the data, the stem and leaf display as well as the boxplot of the studentized residuals is considered. The stem and leaf display illustrated in Figure 5.9 shows that the studentized residuals are roughly symmetric. However, the weighted linear model does not fit the data well. Deletion of the extreme observations does not improve the goodness of fit of the model. To try to overcome this lack of fit and accommodate for the extreme observations, various transformations are considered in the following sections as is the generalized linear model with gamma distribution and beta distribution.

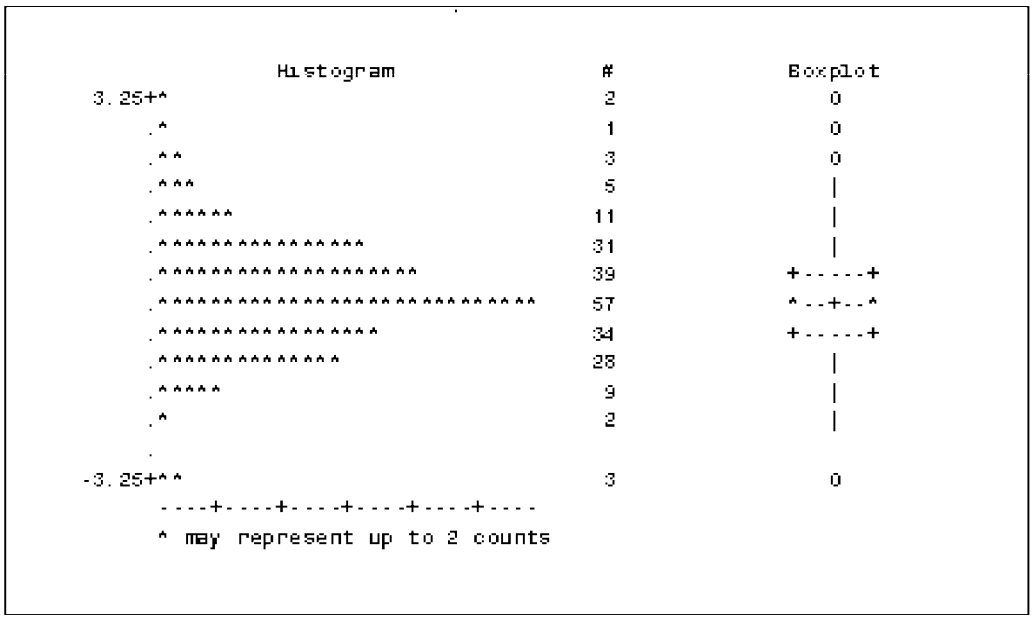

Figure 5.9: The stem and leaf and boxplot display for the studentized residuals.

5.3.1 Log Transformation

The values of the response variable, *FEV1*, lie roughly between 0.25 and 0.7; the linear analysis reveals that outliers are present; the variable, *FEV1* is bounded below by zero and the largest *FEV1* value is almost more than three times larger than the smallest *FEV1* value. Also the *FEV1* data are ratios by definition. These are all indications that a log transformation should be applied to the data (Osborne, 2002). Also, the log transformation works well for data where the residuals get bigger for bigger values of the response variable. Such trends in the residuals occur often, because the error or change in the value of an outcome variable is often a percent of the value rather than an absolute value. For the same percent error, a bigger value of the variable means a bigger absolute error, so residuals are bigger too. Taking logs pulls in the residuals for the bigger values. From Figure 5.10 it can be seen that the residuals are increasing with increasing *FEV1* values.

In order to apply the log transformation the *FEV1* values are reflected. The data are reflected by multiplying each log transformed *FEV1* value by -1 (Zewotir and Galpin,

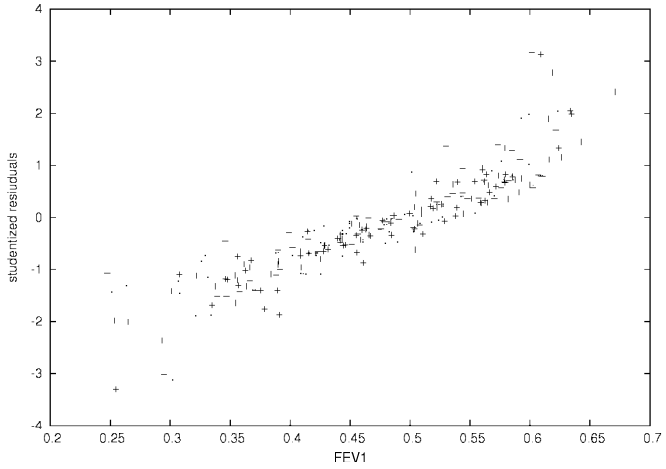

Figure 5.10: Plot of response variable versus the residuals.

2004).

If the results presented in Table 5.5 are compared to those presented in Table 5.1 the overall fit of the data with F-prob = 0.0265, although still significant at alpha level 0.05, has decreased from that obtained for the linear model which was F-prob = 0.0140. The R-square statistic is far too low at 0.0559 and has decreased when compared to the 0.0627 obtained for the linear regression analysis. The significance of the parameter estimates have decreased when compared to the weighted linear regression estimates. The p-values for these estimates indicate that only the predictor variable *sex* remains significant and the predictor *statusL* is no longer significant with the transformation.

To check the normality of the data, the residual plot is shown in Figure 5.11. If this plot is compared to that shown in Figure 5.1 for the linear regression analysis, clearly the two plots differ in that the range for the predicted values has now changed due to the log transformation being used here. With the linear regression model the predicted values lie between 0.42 and 0.56, whereas with the reflected log transformation the predicted values now lie between 0.6 and 0.85. The points reflected in log linear residual plot does appear more randomly scattered than those in the linear regression residual plot. The residual plots in Figures 5.1 and 5.11 do not have outlying observations that stand

Table 5.5: Results for reflected Log Linear Transformation Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	44.08189	8.81638	2.60	0.0265
Error	224	744.02630	3.39738		
Corrected Total	229	788.10819			

Root MSE	1.84320	R-Square	0.0559
Dependent Mean	0.71639	Adj R-Sq	0.0344
Coeff Var	257.28816		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t value	Pr > |t|
Intercept	1	0.82931	0.33783	2.45	0.0149
Weight	1	-0.00194	0.00233	-0.83	0.4060
Height	1	-0.02125	0.28490	-0.07	0.9406
Sex	1	-0.06909	0.02938	-2.35	0.0196
StatusL	1	0.06980	0.03760	1.86	0.0648
StatusM	1	-0.01384	0.03380	-0.41	0.6825

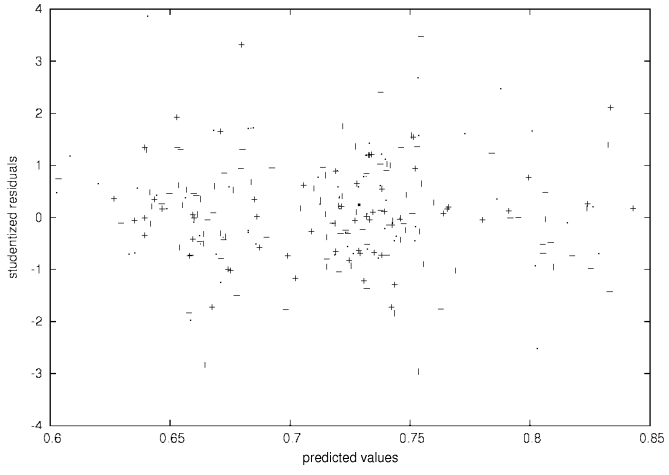

Figure 5.11: Residual Plot for reflected log linear regression model.

out. However, with the reflected log linear analysis, extreme observations also do exist; and a summary of observations with extreme studentized residuals, Cook's distance, leverage, and DIFFIT's statistic are presented in Table 5.6, as well as COVRATIO values of those extreme observations that lie below 1.

Table 5.6: Summary of diagnostics for log linear model

obs	studentized residual	Cook's distance	leverage	DIFFIT's statistic	COVRATIO
129	3.866981	0.0768795		0.702012	0.693153961
138	3.477941	0.10167897	0.048014055	0.801743	0.767790713
65	3.3148	0.05609992		0.593937	0.777852005
12	-2.96086	0.18485598	0.112308464	-1.07243	0.906167438
8	-2.83798	0.46985585	0.25927252	-1.70687	
126	2.68084	0.03131006		0.439713	0.863344109
95	-2.51893	0.08432253	0.073848953	-0.72017	0.930320687
104	2.47284	0.0398654		0.494734	0.901123013
174	2.110663	0.10018685	0.118892162	0.781538	
225	-1.97371	0.04386969	0.063292531	-0.51649	0.985266189
31	-1.84079	0.02024971		-0.35049	0.969598217

In Table 5.6 is presented a summary of the diagnostics for the 11 most extreme observations in the data. The observations that are extreme here are, in general, the same as those observations that were extreme for the linear regression analysis. Only diagnostic measures for the particular observations that are not within the required range have been included in Table 5.6. For instance, the COVRATIO for observation 8 is not given because is it above 1 and therefore contributes to the goodness of fit for the model. Observation 12 has a COVRATIO below 1 and as a result has been included in Table 5.6.

Deletion of the extreme observations 31, 65, 95, 126, 129, and 138 appears to improve the goodness of fit of the model and the significance of the parameter estimates. When observations 65, 95, 126, and 138 are deleted the parameter estimate for *statusL* becomes significant at alpha level 0.05%. Deletion of observations 8, 12, 104, 174, and 225 decreases overall model goodness fit and the significances of parameter estimates. All observations listed in Table 5.6 can be regarded as being outlying and/or influential.

The residual plots obtained when the various extreme observations are deleted, are

presented in Figures 5.12, 5.13, 5.14, 5.15, and 5.16. These residual plots are included because deletion of the individual extreme observations caused changes in the plots as compared to in Figure 5.11.

By examining the residual plots for the deleted observations it can be seen that, when extreme observations 31, 95, and 225 are deleted, the resulting residual plot (although slightly more scattered) resembles the most that obtained in Figure 5.11. For deletion of the extreme observations 126, 138, and 174 the scatter in the residual plots is somewhat less and the residuals lie closer to each other when compared to those in Figure 5.11. Deletion of the extreme observation 65 results in the least desirable residual plot in which residuals are close together in three distinct groups. When observations 12, 104, and 129 are deleted the residual plot is much more random than before with the most random results obtained when observation 129 is deleted. With the deletion of the extreme observation 8, the resulting residual plot is concentrated towards the centre of the plot and has a few residuals outlying. There does appear to be a possible 'v' shape with this residual plot.

The three most extreme observations are considered to be those with the most outlying studentized residuals. These are observations 65, 129, and 138. These observations are paired and deleted from the data and a summary of the results obtained is given in Table 5.7.

Table 5.7: Results for deletion of extreme observations

observations deleted	F-prob	R-Sqr	p-value for sex	p-value for $statusL$
65 and 129	0.0017	0.0847	0.0017	0.0326
65 and 138	0.0143	0.0630	0.0153	0.0129
129 and 138	0.0091	0.0677	0.0151	0.0240

From Table 5.7 it can be seen that paired deletion of the three most extreme observations results in an improvement in the goodness of fit and of the significance of the parameter estimates sex and$statusL$. However, the R-square is still small at 0.087 when observations 65 and 129 are deleted. The residual plots for the deletion of the paired

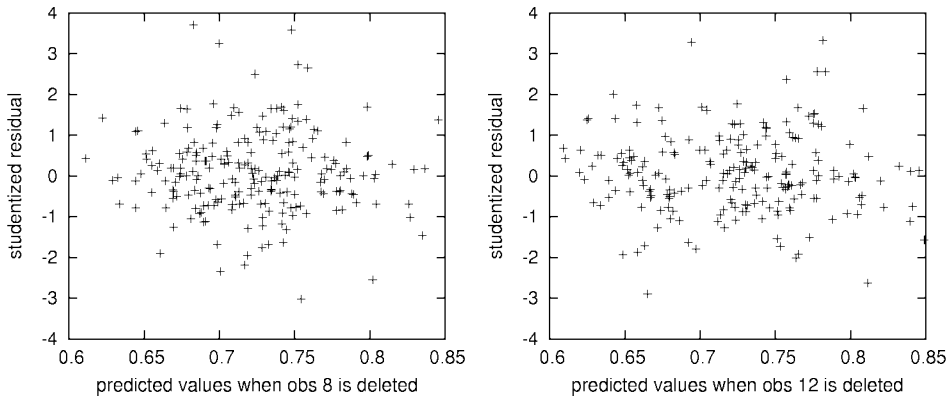

Figure 5.12: Residual Plot for reflected log linear regression model when observations 8 and 12 are deleted individually.

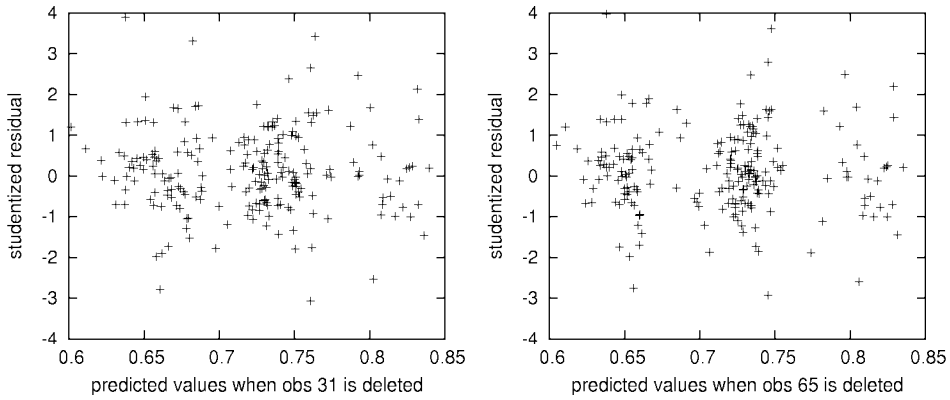

Figure 5.13: Residual Plot for reflected log linear regression model when observations 31 and 65 are deleted individually.

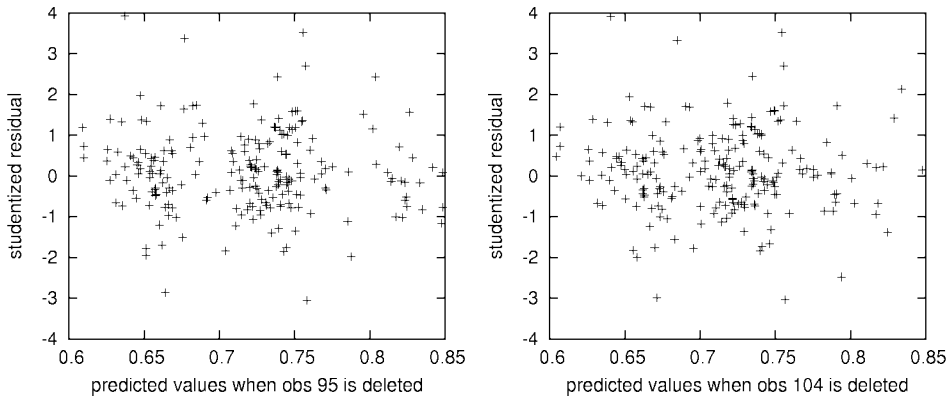

Figure 5.14: Residual Plot for reflected log linear regression model when observations 95 and 104 are deleted individually.

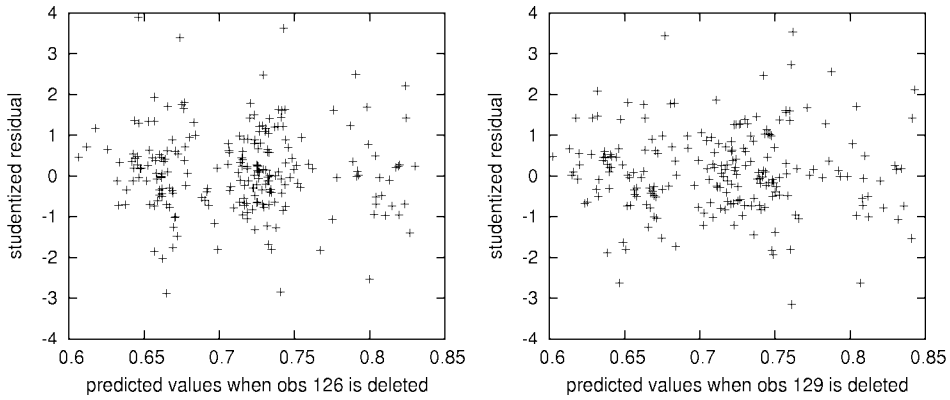

Figure 5.15: Residual Plot for reflected log linear regression model when observations 126 and 129 are deleted individually.

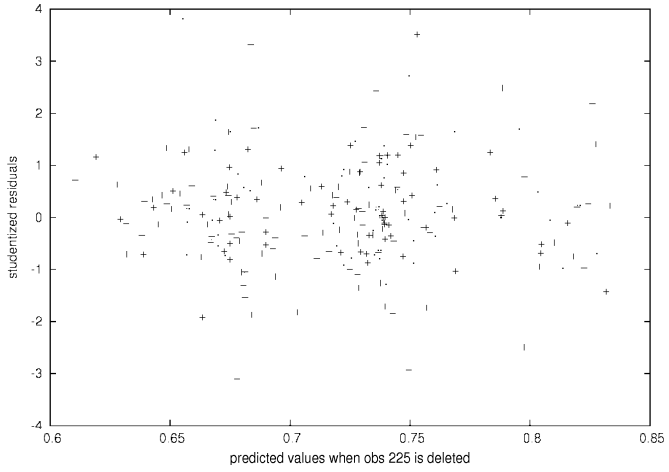

Figure 5.16: Residual Plot for reflected log linear regression model when observation 225 is deleted.

observations display more of a pattern when compared to the residual plot for the full data. When observations 129 and 138, as well as 65 and 129 are deleted the residual plot obtained strongly resembles that shown in Figure 5.11. The residual plot obtained when observations 65 and 138 are deleted is shown in Figure 5.17. The figure clearly shows a distinct pattern of three groups of residuals.

The normal probability plot for the log linear model is shown in Figure 5.18.

The four formal tests for normality for the probability plot in Figure 5.18 all agree that the normality assumption is violated. This could be as a result of outliers. Notice that in Figure 5.18 the outlier that stands out the most is observation 129. This extreme observation does not fall within the range of the probability plot since it has a studentized residual equal to 3.866981. Its approximate position is given in Figure 5.18.

When the extreme observations that have been listed in Table 5.6 are deleted from the data, observations 12, 31, 65, 95, 126, and 129 are considered to be normal according to the Kolomogorov-Smirnov test for normality. The Cramer-von Mises test also indicates normality when deletion of observations 65 and 129 are made. When the paired observations 65 and 138, and 129 and 138 are deleted all four tests for normality agree

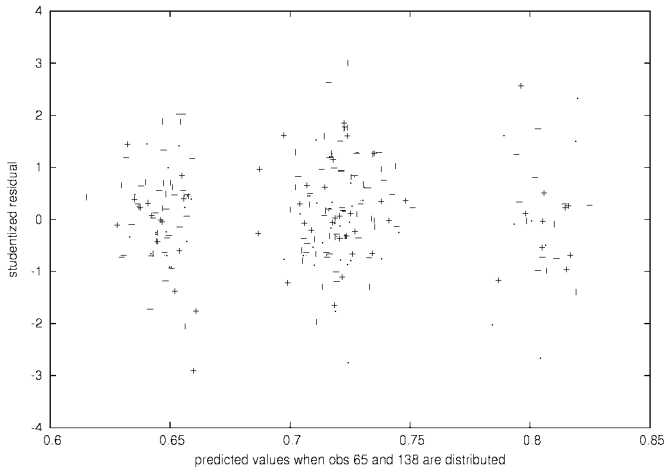

Figure 5.17: Residual Plot for reflected log linear regression model when observations 65 and 138 are deleted.

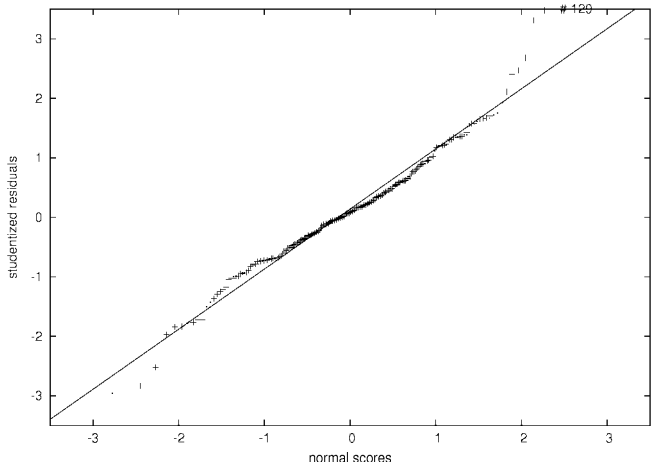

Figure 5.18: Normal probability plot for log linear model.

that the residuals are normal. The proabability plots for these deletions are shown in Figure 5.19. These plots appear normal, with the residuals falling close to the straight line, except along the bottom tail end. Only the Kolomogorov-Smirnov test indicates normality when observations 65 and 138 are deleted.

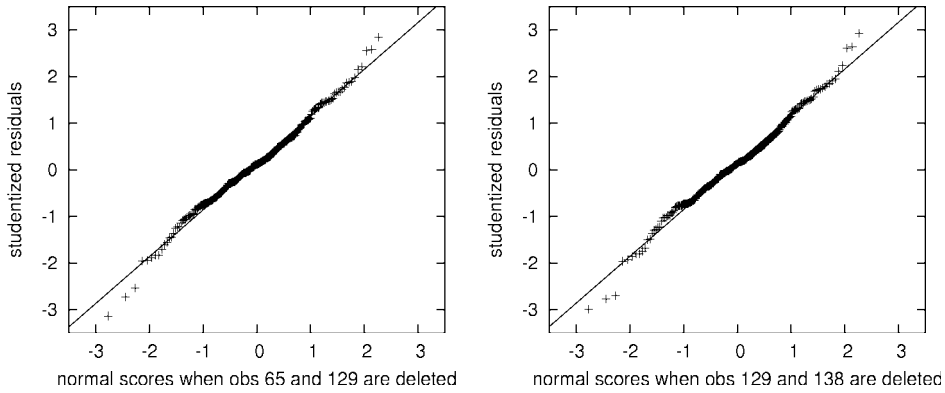

Figure 5.19: Normal probability plots when paired observations are deleted.

There is an improvement with the extreme observations. When the response of $FEV1$ are log transformed they are slighlty less outlying observations than in the linear regression model. Normality of the studentized residuals appears to be attained when certain extreme paired observations are deleted. In general, the reflected log linear analysis does not result in significant difference with regard to the goodness of fit of the model. The R-square statistics remain small.

5.3.2 Inverse Transformation

Taking the reflected inverse of the response *FEV1* essentially makes the very small numbers large and very large numbers small. This transformation has the effect of

reversing the order of the data. This transformation is used here for the same reasons the log linear transformation was used in the previous section.

The results for the reflected inverse transformation data are given in Table 5.8.

Table 5.8: Results for reflected inverse transformation model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	204.36045	40.87209	2.20	0.0549
Error	219	4060.75614	18.54227		
Corrected Total	224	4265.11659			

Root MSE	4.30607	R-Square	0.0479
Dependent Mean	-2.10025	Adj R-Sq	0.0262
Coeff Var	-205.02645		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t value	Pr > |t|
Intercept	1	-2.35780	0.78923	-2.99	0.0031
Weight	1	0.00450	0.00545	0.82	0.4104
Height	1	0.04696	0.66559	0.07	0.9438
Sex	1	0.14794	0.06865	2.16	0.0322
StatusL	1	-0.14430	0.08785	-1.64	0.1019
StatusM	1	0.03490	0.07896	0.44	0.6589

The results given in Table 5.8 are different to those for the log linear model in that the F-prob says that the model fit is insignificant at alpha level 0.05%. The R-square is still far too small at 0.0479. Here, only the parameter estimates for *sex* are significant. The overall analysis of variance and parameter estimates appear to be slightly worse than what was obtained for the linear regression model and the log linear model.

The residual plot for the reflected inverse transformation model is given in Figure 5.20.

From the residual plot shown in Figure 5.20, the residuals appear somewhat random and the plot resembles that obtained and shown in Figure 5.11. Notice observation 129 has an extreme studentized residual of -4.53585, which does not fall within the range of the residual plot. Its approximate position is shown in Figure 5.20.

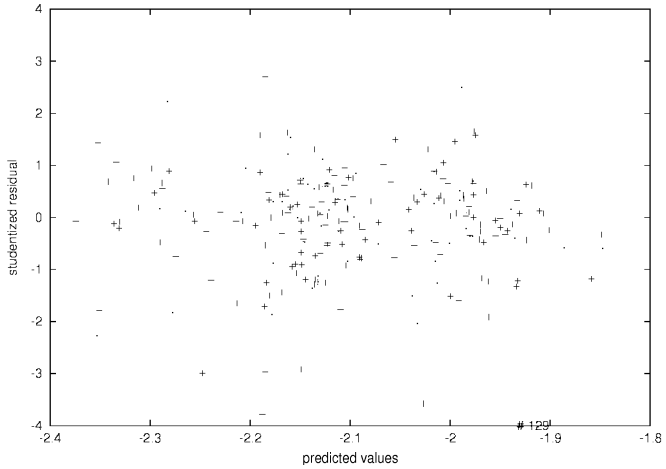

Figure 5.20: Residual plot for the Inverse Transformation Model.

The extreme observations with the reflected inverse transformations are observations 8, 12, 44, 65, 95, 129, 138, and 174. These are more or less the same observations that were found to be extreme with the reflected log linear analysis and these observations are identical to the most extreme observations with the linear regression analysis. If the extreme observations are deleted one by one with the inverse transformation, it is found that, in general, deletion of observations 12, 65, 95, and 129 improves the goodness of fit of the model. The R-square statistic increases slightly and the parameter estimates significance also increases. When observation 95 is deleted the parameter estimate for *statusL* becomes significant. The highest R-square is obtained when observation 129 is deleted and here the R-square is 0.0692. When observations 8, 44, 138, and 174 are deleted the overall fit of the model and the significance of parameter estimates decrease.

Analysis of diagnostics reveals that the Cook's statistic, leverage, DIFFIT's statistic, and COVRATIO are similar to those obtained for the linear regression analysis and the reflected log linear transformation analysis. Deletion of paired observations produces similar results to those obtained for the previous two models. The three most extreme observations are 8, 12, 95, and 129. When observations are deleted in pairs, all results indicate an improvement in goodness of fit. Parameter estimates for *sex* and *statusL*

are both significant when observations 95 and 129 are deleted. The R-square is highest when observations 12 and 129 are deleted. This R-square value is 0.0872. The residual plots produced for deletion of individual and for paired extreme observations are similar to those obtained for the reflected log linear analysis.

Concerning the normal probability plot obtained for the inverse transformation, one can see from Figure 5.21 that the data do not appear normally distributed. The plot deviates drastically from the straight line. As expected, the formal tests for normality indicate violation of the normality assumption.

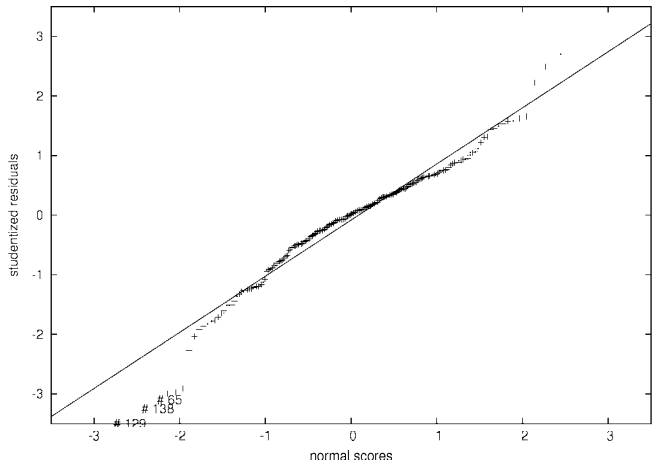

Figure 5.21: Normal probability plot for the inverse transformation model.

Extreme observations 65, 129, and 138 stand out in the probability plot for the inverse transformation model. When these observations are deleted from the data, the resulting normal probability plots do not seem to improve much. In general, deletion of the extreme observations one by one or in pairs does not have a significant effect on the normality of the reflected inverse transformation model.

If one looks at the distribution of the studentized residuals for the inverse transformation model one would expect it to be normally distributed since the data was reflected before being transformed. Judging by the histogram shown in Figure 5.22, the studentized residuals appear to be slightly positively skewed.

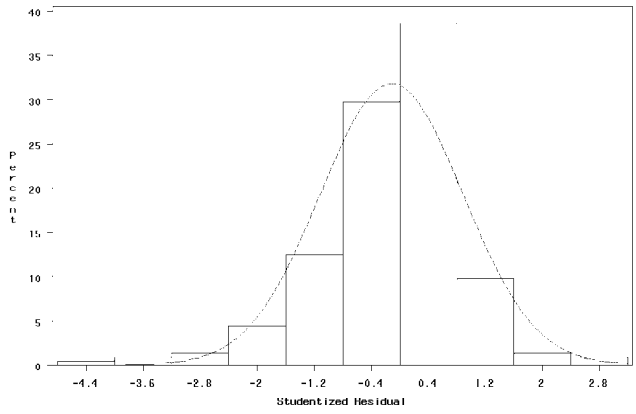

Figure 5.22: Histogram for studentized residuals.

The reflected inverse transformation does not improve model fit. The results obtained for this transformation appear to be worse than those obtained for the reflected log linear results.

Note that the reflected log transformation is a combination of the log and inverse transformation: that is, the log of the inverse of the response variable *FEV1* is regressed against the predictor variables.

5.3.3 Generalized Linear Model

The Gamma distribution

With the data in the original linear model and in the inverse model the normality assumption was not strongly justified. A generalized linear model with gamma distribution would be the possible solution for failure of the normality assumption. The gamma distribution with log and inverse link is considered. The generalized model with gamma distribution and log link is considered first and the results are presented in Tables 5.9 and 5.10.

The results for this model show that the gamma log link model does not fit the data

Table 5.9: Criteria for assessing goodness of fit

Criterion	DF	Value	Value/DF
Deviance	219	695.5848	3.1762
Scaled Deviance	219	227.6260	1.0394
Pearson Chi-Square	219	615.9797	2.8127
Scaled Pearson	219	201.5757	0.9204
Log Likelihood		177.0721	

Table 5.10: Results for GLM model

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.8171	0.3236	-1.4514	-0.1828	6.37	0.0116
Weight	1	0.0018	0.0022	-0.0025	0.0062	0.67	0.4135
Height	1	0.0324	0.2722	-0.5012	0.5659	0.01	0.9054
Sex	1	0.0659	0.0279	0.0112	0.1206	5.58	0.0182
StatusL	1	-0.0699	0.0357	-0.1399	0.0000	3.84	0.0501
StatusM	1	0.0125	0.0321	-0.0504	0.0755	0.15	0.6960
Scale	1	0.3272	0.0305	0.2726	0.3928		

well. The goodness-of-fit statistics shown in Table 5.9 have a Deviance and Pearson Chi-Square deviance greater than one. This is a clear indication that the model does not fit the data well. Only the parameter estimate for *sex* is significant at alpha level 0.05%, and *statusL* is almost significant for the same alpha level.

The residual plot which is interpreted in a similar way to that for linear regression analysis is presented in Figure 5.23.

The residual plot for the gamma log link model is similar to that obtained for the previous models. There appears to be a reasonable amount of scatter as shown in Figure 5.23. The extreme observations are the same as those obtained from the linear regression analysis except that here observation 31 has taken the place of extreme observation 186 from the linear analysis. Again, the most outlying and influential observation is 129.

The results for the gamma model with inverse link are very similar to those obtained for the gamma model with log link. The residual plot strongly resembles that shown in Figure 5.23 and once again the most extreme observation is 129. The goodness-of-fit statistics do not indicate that the model fits the data well. The parameter estimate

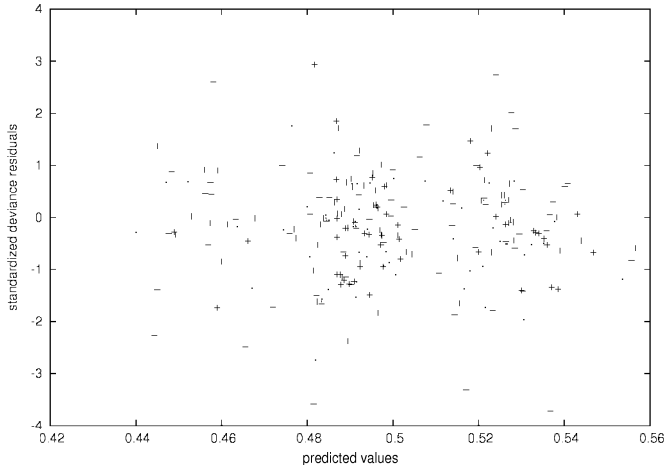

Figure 5.23: Residual plot for the gamma log link model.

sex is significant and *statusL* is significant with a p-value of 0.0500. In general, there is no significant improvement.

When the identity link is used with the gamma distribution, the results obtained are similar to those obtained for the gamma model with log link. Extreme observations conform to those previously obtained. The model fit does not differ significantly, and does not appear to improve.

If the above mentioned three gamma models are compared, it is found that the results obtained for the criteria for assessing goodness of fit are similar in all three cases. The results shown in Table 5.9 are identical to those of the inverse link, and identity link gamma models up to 3 significant digits.

When the test for the link function is performed, the plot of the vector, $\mathbf{z}$, against the linear predictor vector for all three link functions does not depict a straight line. Recall that the vector, $\mathbf{z}$, is the first gradient and is made up of the sum of the linear predictor, and the product of the derivative of the linear predictor with the raw residual, as given in (3.4). Since none of the plots depict a straight line it is evident that the second component on the right-hand side of (3.4) has a significant effect on the vector $\mathbf{z}$. As a result, one concludes that the log, identity, and inverse link with gamma distribution

do not fit the data well, and therefore the gamma distribution is inappropriate.

The Beta distribution

The probability density function for the beta distribution, $f(x)$, (Hogg and Craig, 1995) is greater than 0 for $0 < x < 1$. As a result, one considers this distribution for $FEV1$ since the response variable $FEV1$ lies within the same interval.

The analysis of the beta distribution with the logit link is carried out in PROC GLIMMIX. The fit statistics obtained using PROC GLIMMIX reveal that the effects are all insignificant. The results are presented in Tables 5.11 and 5.12.

Table 5.11: Fit statistics

-2 Log Likelihood	-405.15
AIC	-391.15
AICC	-390.63
BIC	-367.23
CAIC	-360.23
HQIC	-381.49
Pearson Chi-Square	228.84
Pearson Chi-Square DF	1.04

Table 5.12: Type III tests of fixed effects

Effect	Num DF	Den DF	F Value	Pr > F
weight	1	219	0.46	0.4967
height	1	219	1.28	0.2599
Sex	1	219	1.91	0.1689
statusI	1	219	1.62	0.2048
statusM	1	219	3.04	0.0827

The results for the generalized linear model with gamma distribution indicated that the asthma status of the children were more or less significant. The generalized linear model with beta distribution clearly indicates that all explanatory variables are insignificant at alpha level 0.05. When the logit link is tested, the plot obtained can be seen in Figure 5.24. This plot appears linear.

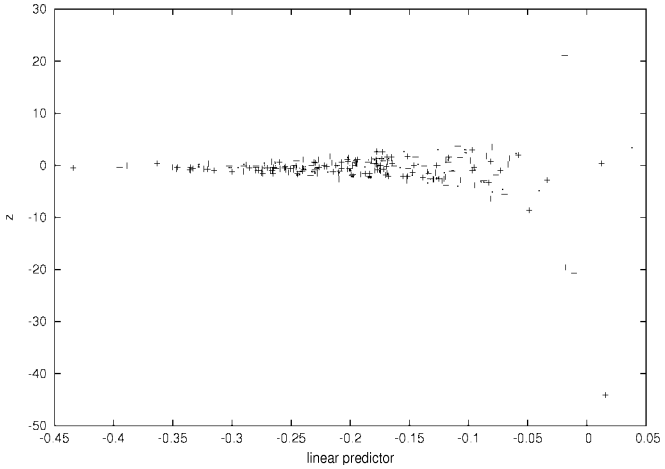

Figure 5.24: Plot of z vs linear predictor for the logit link with beta distribution .

When other links are used with the beta distribution the results do not differ drastically from those given in Tables 5.11 and 5.12. However, it is clearly seen in Figures 5.25, 5.26, and 5.27, that the other link functions are inappropriate.

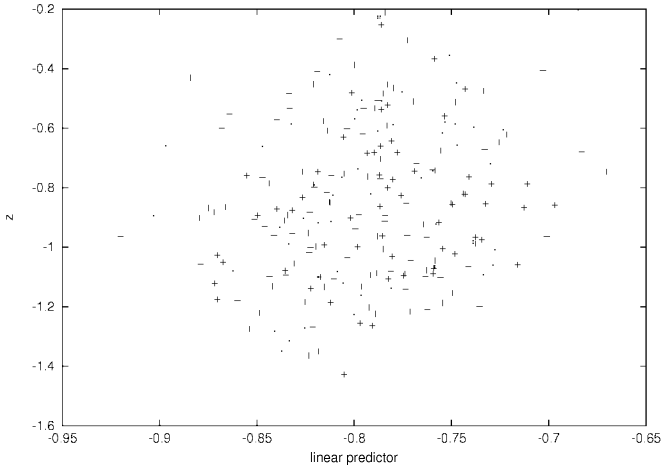

Figure 5.25: Plot of z vs linear predictor for the log link with beta distribution .

The beta distribution with logit link is has explanatory variables that are insignificant therefore the model does not fit the data.

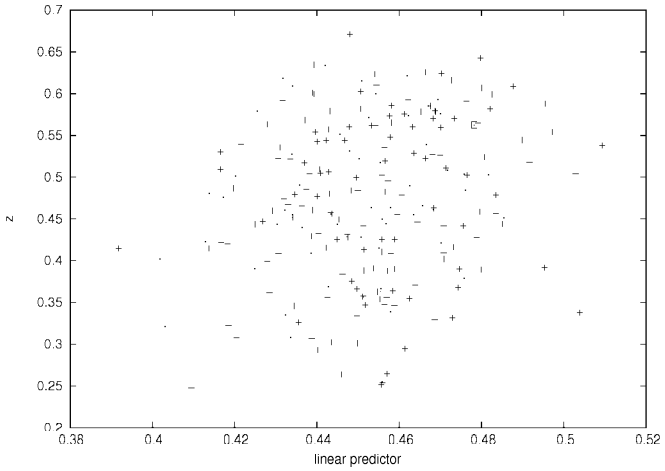

Figure 5.26: Plot of z vs linear predictor for the identity link with beta distribution .

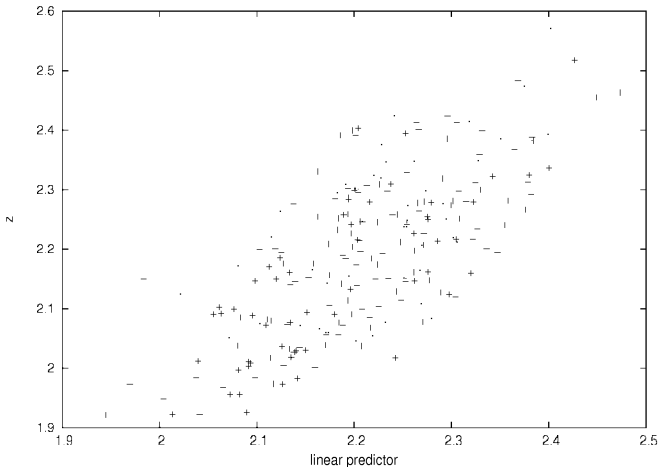

Figure 5.27: Plot of z vs linear predictor for the inverse link with beta distribution .

5.4 Summary

The models in this chapter failed to fit the Durban South data well. It was found that, in general, the most extreme observations were observations numbers 8, 12, 65, 129, and 138. The summary of these observations is given in Table 5.13.

Table 5.13: Extreme observations

obs	*height*	*weight*	*statusI*	*statusM*	*sex*	*FEV1*
8	1.29	28	0	1	1	0.6090401
12	1.28	25	0	0	0	0.60132
65	1.32	27	0	0	1	0.2946012
129	1.4	36	1	0	0	0.3563011
138	1.32	24	0	0	0	0.3019459

The average height of a student is 1.38m. Observations 12 and 8 seem to lie way below the mean height. The minimum and maximum heights are 1.09m and 1.58m respectively, and none of the extremes comes close to these values. The mode for height is 1.38m therefore there is nothing unusual about observations 65, 129, and 138 with regard to height.

The mean weight is 34.62kg and the smallest and largest weight values are 17kg and 83kg respectively, with the most frequent weight value being 30kg. By examining at the height and weight, it can be seen that student 138 is under weight and student 129 is slightly over weight. The remaining observations are slightly under weight.

Out of the 225 children 53 have an asthma status considered as normal, 61 children have an asthma status that is considered mild asthmatic and the remainder which is 111 children have a persistent asthmatic status. In Table 5.13, it is shown that observations 12, 65, and 138 are persistent asthmatic, observation 8 is mild asthmatic and observation 129 is normal asthmatic.

In total, 130 students are female and 95 are male. The mean value for the average FEV1 readings denoted as *FEV1* is 0.472677108. Observations 8 and 12 lie above the mean value and the remainder lie below, with observation 65 having the smallest FEV1 value.

The diagnostic check for the Dendrometer data did not reveal any extreme observations in terms of outliers and influential observations. Other transformations such as the square root and arcsin produce results very similar to those presented in the previous sections. Unfortunately, not one of the models fits the data or accounts for the outliers and influential observations.

Chapter 6

Dendrometer data analysis

6.1 Introduction

In this chapter, the focus is on the analysis of measurement error with the Dendrometer data. A brief description of the Dendrometer data set is given followed by the basic linear regression analysis with a few multicollinearity diagnostic checks. One of the predictor variables has the measurement error and as a result the various measurement error approaches are applied to try and determine whether the existing measurement error has a serious effect on the linear regression results.

6.2 The Dendrometer Data

The Dendrometer data were obtained from the 'dendrometer trial' which was a joint effort between the CSIR (Council for Scientific and Industrial Research)and SAPPI (South African Pulp and Paper Inc) to conduct a range of experiments to understand the processes of fibre development in plantation eucalypts.

A dendrometer is a band which is commonly used to make short-term repeated measurements of tree-stem growth. These bands are fairly easy to make and install on most trees. Dendrometer bands consist of thin straps of metal placed around a tree,

with one end passed through a collar and then connected back to itself with a spring. As the tree grows (or as it shrinks and swells diurnally), the spring allows the band to move with the changing circumference of the tree. As the stem expands, the band slips through the collar and the spring is stretched (Clark, Wynne, and Schmoldt, 2000). These measurements of either expansion or shrinkage are recorded.

The variable *dailystemincrement* is defined as being the period of time for which the stem continues to grow above a previous maximum (i.e., a net growth increase). The period of time each day for which the stem continues to grow prior to beginning to shrink is defined as being the duration of stem increment for day (times hours). The rate of increment is calculated as the change in stem size from the beginning of increment divided by the duration of increment. The variable *sum_rainfall* represents the rainfall in the data. Temperature measurements were obtained from two sensors, the MCSystems automatic weather station, and also from a HOBO logger which is an electronic instrument that records measurements such as temperature and relative humidity over time (www.onsetcomp.com/hobo). The temperature measure is denoted as *avg_temperature*. The variable *avg_solarradiation* is a measure of the total incoming solar radiation summed over the minimum specified logging interval. The wind speed is measured by an MCSystems 177 wind speed sensor. The variable *avg_windspeed* is the average of wind speed over the minimum logging interval. The relative humidity data are represented by the predictor variable *avg_rh*. This measure was obtained from two sensors, the MCSystems automatic weather station, and also from a HOBO logger.

The study aims to see how the daily stem increment of eucalypts is influenced by rainfall, temperature, solar radiation, wind speed, and humidity.

Analysing the Dendrometer data is of particular interest because they contain measurement error. The error occurs with the predictor variable used to measure the relative humidity present. This variable does not represent the true amount of relative humidity because it has been manipulated so that it does not exceed a value of 100. That is, if the relative humidity was recorded as 101 this value was changed to 100 before analysis. The true relative humidity values for the Dendrometer data set are

unavailable. As a result, measurement error analysis is carried out to try and account for the measurement error now present in the data.

The analysis is presented such that initially the measurement error is ignored. Thereafter, the presence of measurement error is acknowledged. Various techniques discussed in Chapter 4 are used to see if the measurement error has a serious impact on the results obtained when measurement error was not accounted for. This illustrates the effect of the measurement error.

6.3 Linear Regression Analysis when Measurement Error is Ignored

Firstly, the linear regression model is considered in order to try and fit the data. The response variable is *dailystemincrement* and the predictors are *avg_temperature*, *sum_rainfall*, *avg_solarradiation*, *avg_windspeed*, and *avg_rh*. The results of analysis are presented in Table 6.1.

The multiple regression function is given as

$$\begin{aligned} dailystemincrement = {} & 113.09526 + 0.99789 avg\_temperature + 0.14668 sum\_rainfall \\ & + 0.00198 avg\_rh - 101.91948 avg\_solarradiation \\ & - 9.26300 avg\_windspeed. \end{aligned} \tag{6.1}$$

From the regression function (6.1) the intercept term is equal to 113.09526. This implies that if all the regressor variables equal zero then 113.09526 estimates the mean response $E(\mathbf{y})$. The regression coefficient for the temperature which equals 0.99789 is the estimate of the change in the mean response, per unit increase in *avg_temperature* when *sum_rainfall*, *avg_rh*, *avg_solarradiation*, and *avg_windspeed* are held constant. The remaining regression coefficients are interpreted in a similar manner.

From Table 6.1 it can be seen that the overall model fit is significant, with an F-prob = <0.0001. Therefore not all regression coefficient estimates are equal to zero.

Table 6.1: Results for Linear Regression Model with the Dendrometer data

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	263009	52602	23.11	<.0001
Error	360	819540	2276.49939		
Corrected Total	365	1082548			

Root MSE	47.71268	R-Square	0.2430
Dependent Mean	53.54661	Adj R-Sq	0.2324
Coeff Var	89.10494		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t value	Pr > |t|
Intercept	1	113.09526	35.75752	3.16	0.0017
avg_temperature	1	0.99789	0.86348	1.16	0.2486
sum_rainfall	1	0.14668	0.40250	0.36	0.7158
avg_rh	1	0.00198	0.37107	0.01	0.9958
avg_solarradiation	1	-101.91948	13.73321	-7.42	<.0001
avg_windspeed	1	-9.26300	4.83390	-1.92	0.0561

The only significant parameter is *avg_solarradiation*, and the remaining estimates are insignificant at alpha level 0.05. The R-square statistic is 0.2430, therefore 24.30% of the variation in *dailystemincrement* is explained by the predictor variables.

6.3.1 Multicollinearity Diagnostics

Solar radiation is essentially the energy from the sun and it includes ultraviolet radiation, visible radiation, and infrared radiation. Humidity is the concentration of water vapour in the air. Temperature is commonly known as the physical property of a system which underlies the common notions of 'hot' and 'cold': the material with the higher temperature is said to be hotter. The wind speed is a measure of the average speed of movement of the wind at a specific point, and rainfall is that part of precipitation that produces runoff. It is commonly known that solar radiation and temperature are strongly related to one another. Also, humidity is dependent on temperature and rainfall. As a result of this common knowledge, one is inclined to believe that the

possibility of multicollinearity within the Dendrometer data exists. The analysis for multicollinearity diagnostics is therefore presented.

The causes of multicollinearity, its effects on inference, methods of detecting multicollinearity, as well as techniques for dealing with the problem have been discussed in Chapter 3. In the present section, the presence of multicollinearity within the Dendrometer data is checked.

The Pearson correlation coefficients (r) along with the p-values prob $> |r|$ for testing $H_0 : \rho = 0$ are presented in Table 6.2. As can be seen from Table 6.2, all predictor

Table 6.2: Pearson correlation coefficients with Prob $> |r|$ under $H_0 : \rho = 0$

	avg_temperature	sum_rainfall	avg_rh	avg_solarradiation	avg_windspeed
avg_temperature	1	-0.12577	-0.20312	0.60023	0.38517
		0.0161	<0.0001	<0.0001	<0.0001
sum_rainfall	-0.12577	1	0.30338	-0.26482	0.10342
	0.0161		<0.0001	<0.0001	0.0480
avg_rh	-0.20312	0.30338	1	-0.45400	-0.32137
	<0.0001	<0.0001		<0.0001	<0.0001
avg_solarradiation	0.60023	-0.26482	-0.45400	1	0.38343
	<0.0001	<0.0001	<0.0001		<0.0001
avg_windspeed	0.38517	0.10342	-0.32137	0.38343	1
	<0.0001	0.0480	<0.0001	<0.0001	

variables appear to be significantly correlated at alpha level 0.05. The highest correlation is that between *avg_solarradiation* and *avg_temperature*, and here the correlation coefficient is 0.60023.

Examination of the variance inflation factors (VIF) does not reveal the presence of multicollinearity. The VIF values are shown in Table 6.3. Recall the VIF for each predictor in the model measures the combined effect of the dependences among the regressors on the variance of that term. In order for this to be detected, the VIF values need to exceed 5 or 10 (Montgomery et al., 2001). None of the VIFs in Table 6.3 indicates multicollinearity.

By examining at Table 6.2 one can see that there is slight correlation between solar

Table 6.3: Variance inflation factors for the Dendrometer data

Variable	Variance Inflation
Constant	0
avg_temperature	1.67136
sum_rainfall	1.22426
avg_rh	1.43107
avg_solarradiation	1.98464
avg_windspeed	1.39035

radiation and temperature. This correlation is expected, and at the same time it is not severe. Since the variance inflation factors do not indicate any correlation, one can safely conclude that the multicollinearity within the Dendrometer data is at a minimum.

Linear model diagnostics were carried out with the Dendrometer data. The analysis follows analogously to that presented for the Durban South data. However, transformation and the GLM models were not made use of. It was found that the linear regression model does not fit the data well even with the deletion of extreme observations and paired extreme observations.

The conclusion drawn from the multiple linear reg

ression model is that firstly, multicollinearity is not significant. The intercept and the predictor representing solar radiation are significant at an alpha level of 0.05%. All remaining predictor variables are insignificant at an alpha level of 0.05%. Therefore, according to the regression model only relative humidity affects the daily stem increment of the eucalypts.

6.4 Measurement Error analysis

6.4.1 Asymptotic Results

Recall from Chapter 4 that the asymptotic approach can be used to determine the effects of the errors of measurement by assessing the perturbation index which was given as

$$\sum_{i=1}^{n} c_{ii}\sigma_i^2.$$

Here $i = 1, ..., 6$, $c_{ij} = (X'X)^{-1}$, $j = 1, ..., 6$, and σ_i^2 is the variance of measurement error of the ith predictor variable. As mentioned for the Dendrometer data only the predictor variable representing relative humidity is with measurement error, and one assumes the variance of this measurement error equals 1. The remainder of the predictor variables in the data set have a measurement error variance of 0.

Using a PROC IML programme written in SAS, the calculated perturbation index for the Dendrometer data equals 0.0000605. The perturbation index here is essentially equal to the diagonal element of the inverse design matrix corresponding to the predictor *avg_rh*, which represents relative humidity. This is as a result of assuming that the measurement error variance for the relative humidity equals 1 and that the remainder of the predictor variables are error free.

In the absence of any prior information, the choice of variance being equal to 1 is customary. That is why a variance of 1 and a mean of 0 for the measurement error are used.

Recall the perturbation index represents the sum of the relative asymptotic bias which is a distance measure used to assess the effects of errors in measurement. With the Dendrometer data the index is small. One can therefore assume the impact of additive measurement error on the results of the linear regression analysis will be small with regard to the asymptotic approach.

6.4.2 Perturbation results

The perturbation approach for measurement error discussed in Chapter 4 involved calculating a simple bound for the relative error in the coefficients of the predictor variables to try and assess the measurement error. With the Dendrometer data only the relative error in the coefficient estimates of all predictor variables for error in the relative humidity variable is of interest.

From the linear regression analysis, the fitted equation for the Dendrometer data is given in equation (6.1). The measurement error variance for all predictor variables except avg_rh is assumed to be 0. The measurement error variance for avg_rh is assumed to be 1. The bounds for the relative errors in the regression coefficients for error in variable avg_rh are presented in Table 6.4.

Table 6.4: Bounds for relative errors in regression coefficients for measurement error in avg_rh

Regression coefficient of	Relative error
constant	0.0427092
avg_temp	0.018772
sum_rainfall	0.1089675
avg_rh	27.654251
avg_solarradiation	0.0064705
avg_windspeed	0.020419

The estimate for the constant is 113.09526, with a relative error of 0.0427092, as shown in Table 6.4. The percentage error is 4.27092 and therefore the estimate $\pm$ 0.018732391 provides the bounds for the true estimate for the constant. This interval is between 108.26505 and 117.92547. The relative error associated with the average temperature is 0.018772. Therefore, 1.8772 percent of the estimate 0.99789 provides the lower and upper bounds for the coefficient. This results in an interval between 0.97916 and 1.01662. With the rainfall covariate the relative error is 0.1089675, therefore 10.89675 percent of the regression coefficient for rainfall is 0.015983352. From this value the

interval in which the true regression coefficient is between 0.13070 and 0.16266. The relative error for the regression coefficient of the relative humidity variable appears to be the largest with a value of 27.654251. This provides a percentage error of 2766.146908 which is rather large. As a result, the interval in which the true estimate of this coefficient lies between -0.05278 and 0.05674. The intervals for the solar radiation and the wind speed variables are between -102.57895 and -101.26001, and between -9.45214 and -9.07386 respectively.

From the above mentioned one can clearly see that the relative humidity coefficient has with it the largest amount of measurement error when compared to the rest of the coefficients. All the values presented are significant up to five decimal points.

6.4.3 Simulation results

The results for the basic simulation approach are presented. This approach, which was discussed in Chapter 4, involves evaluating the distribution of the regression coefficients obtained from the simulations. Measurement error is simulated and added to the error-prone variable and the result is regressed along with the other predictors against the response. The simulated error is assumed to be normally distributed with mean 0, and variance 1. A thousand simulations are carried out and the regression coefficients are calculated for each simulation.

The distribution of the relative humidity coefficient does not appear normally distributed. As seen in Figure 6.1 it seems to be negatively skewed. When measurement error is not accounted for, the regression coefficient is 0.00198. The mean of the regression coefficient from the simulations is 0.037927. Because the coefficient appears negatively skewed, as expected in most cases the median = 0.038445 is bigger than the mean.

Measures of variability with the relative humidity coefficient include range = 0.12261, interquartile range = 0.00952, variance = 0.0001789, and the standard deviation = 0.01337.

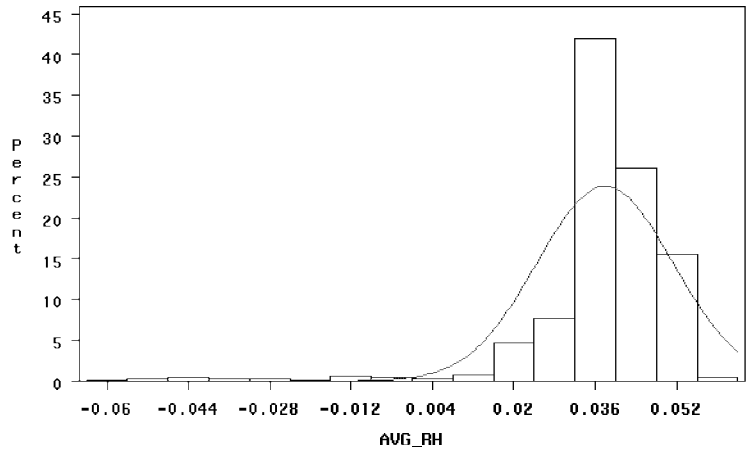

Figure 6.1: Histogram of regression coefficient for relative humidity using the simulation approach.

The skewness of the coefficient is -3.3146828, clearly indicating the possibility of negative skewness. This value also indicates a non-normal distribution since it is less than negative one. The standard error of skewness (ses) can be estimated roughly, using the following formula (Tabachnick and Fidell, 1996): $\sqrt{\frac{6}{n}} = \sqrt{\frac{6}{366}} = 0.128036879$. Since two times the standard error of the skewness is 0.256073758 which is less than the absolute value of skewness 3.3146828, it can be assumed (that the distribution is significantly skewed. Because the 95% confidence interval does not include zero, one can say that there is evidence to reject the hypothesis that the distribution is not skewed.

The kurtosis = 16.7728016, indicating the possibility that the distribution of the coefficient is heavy-tailed relative to the normal distribution, i.e., the extreme portion of the distribution (the part farthest away from the median) spreads out further relative to the width of the center (middle 50%) of the distribution than is the case for the normal distribution. Using the Tabachnick and Fidell (1996) formula of approximation of the standard error of kurtosis it is found to be : $\sqrt{\frac{24}{n}} = \sqrt{\frac{24}{366}} = 0.256073758$. Since two times the standard error of kurtosis is smaller than the absolute value of kurtosis, one can assume the distribution has a significant kurtosis problem.

The existence of skewness and kurtosis indicates violation of the assumption of normality. In Figure 6.1, one can see that the relative humidity coefficient varies from around -0.06 through zero to 0.052. However, more that 96% of the time it is found to be greater than zero, and this implies the significance of the relative humidity coefficient.

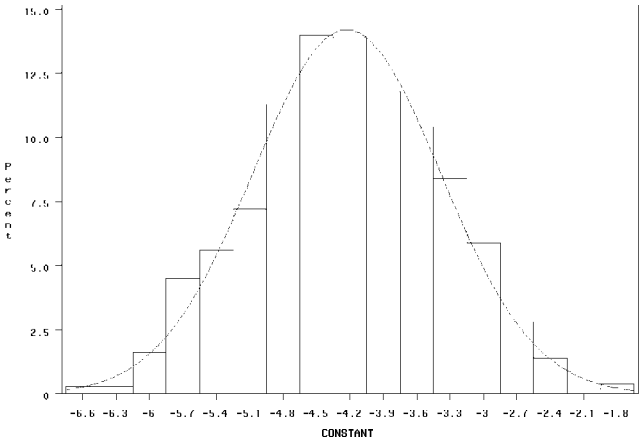

Figure 6.2: Histogram of regression coefficient for the constant using the simulation approach.

In Figure 6.2, the distribution of the regression coefficients for the constant term appears positively skewed. Formal tests for skewness as seen for the relative humidity coefficient indicate positive skewness for the constant term. The kurtosis is also positive, implying a relatively peaked distribution. The various measures of variation and location do not appear unusual and are consistent with a positively skewed distribution. From Figure 6.2, one can see that the constant term does not pass through zero and is therefore significant.

In Figure 6.3, the average temperature appears slightly skewed toward the left. It does not seem to have a normal distribution. The skewness statistic equals -0.3808096 and is bigger than the standard error for skewness, and the 95% confidence interval does imply skewness since the interval does not include zero. The kurtosis value equals

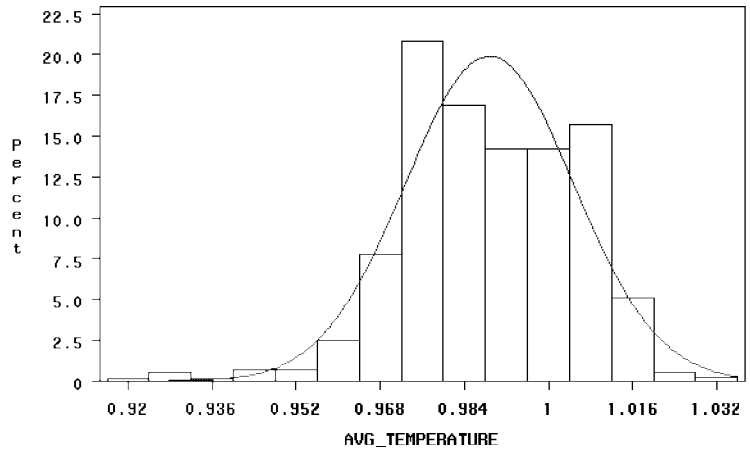

Figure 6.3: Histogram of regression coefficient for average temperature using the simulation approach.

0.44536358, which strictly implies positive kurtosis. Tests also indicate the presence of kurtosis, however, 0.44536358 is close to zero.

The simulations produce a mean value of 0.988792 with a standard deviation of 0.01604 for the average temperature regression coefficient. When measurement error is not accounted for, the average temperature regression coefficient is 0.99789, which lies quite close to the mean value.

With the regressor *sum_rainfall* the mean value of the simulations is -0.0736768, which is much smaller than the 0.14668 obtained from the linear regression analysis. The corresponding standard deviation is 0.07508. Judging from Figure 6.4 one can see that this coefficient is also not normally distributed, and it appears to be strongly positively skewed. Tests for skewness confirm this. Positive kurtosis is also present. The coefficient passes through zero and this implies insignificance, as was seen with the relative humidity coefficient.

The regressions coefficients for the average solar radiation obtained from the simulations also appear to be positively skewed from what is shown in Figure 6.5. The

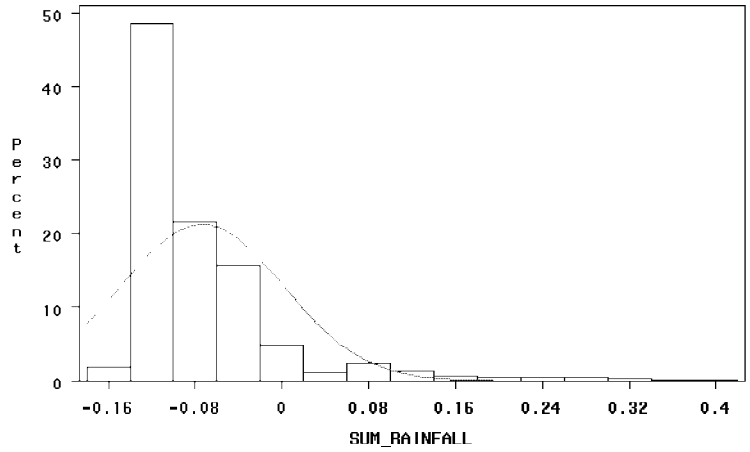

Figure 6.4: Histogram of regression coefficient for rainfall using the simulation approach.

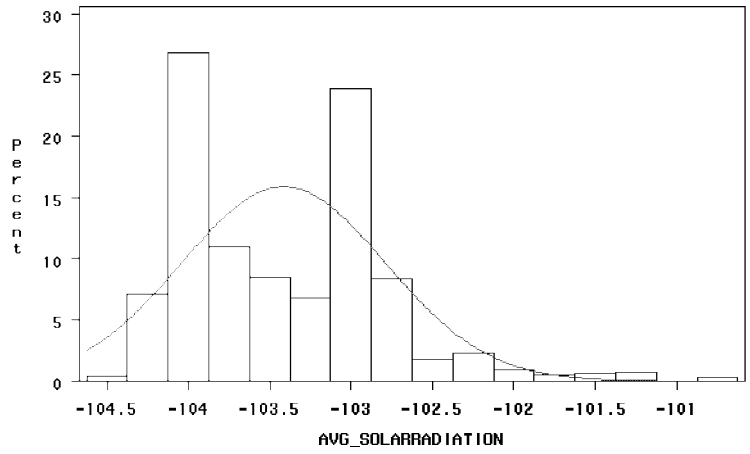

Figure 6.5: Histogram of regression coefficient for average solar radiation using the simulation approach.

corresponding mean value here is -103.416, which is relatively close to the value of -101.91948 from the linear regression analysis. In Figure 6.5, one can clearly see that the histogram is bimodal. The larger modal class is the one with coefficients with -104. The smaller modal class is the one with coefficients with -103. This histogram suggests that there are two groups of coefficients.

Tests for skewness and kurtosis indicate the presence of positive skewness as well as positive kurtosis. In Figure 6.5, one can see that the coefficient does not pass through zero and as a result the coefficient is considered to be significant.

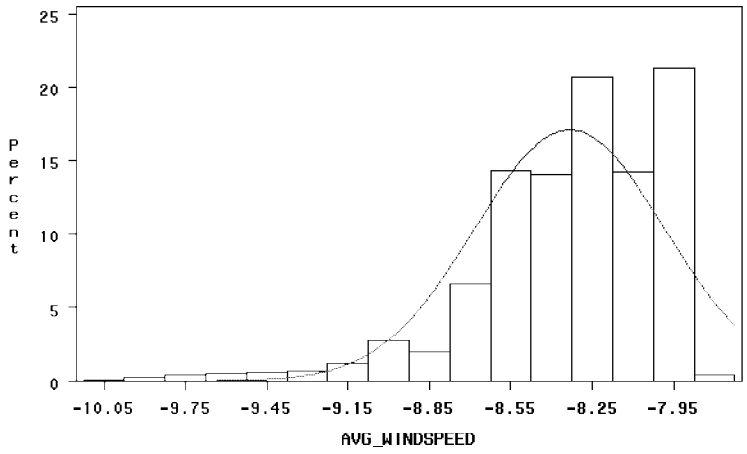

Figure 6.6: Histogram of regression coefficient for average wind speed using the simulation approach .

The average wind speed produces coefficients which are strongly negatively skewed, which is confirmed by the formal tests for skewness. The skewness can clearly be seen in Figure 6.6. Positive kurtosis is also present in the data. The mean value of the simulated regression coefficients here is -8.33285, whereas the regression coefficient for the average wind speed with the linear analysis is -9.26300. The coefficient does not pass through zero and as a result the coefficient remains significant.

In general, the distributions of all the regression coefficients which were obtained via simulations are all skewed. The various formal tests for normality all indicate that the

distributions are not normal. For each regression coefficient estimate, the average value of the 1000 estimates obtained from the simulations does not deviate drastically from the corresponding regression coefficients obtained with the linear regression analysis shown in Table 6.1. The rainfall coefficient is considered to be insignificant according to the simulation approach. This implies that, when measurement error is taken into account by simulating error and adding it to the error-prone variable a large number of times, the resulting coefficients obtained for the relative humidity and the rainfall predictors are sometimes zero.

6.4.4 Bootstrap Results

In this section, the results for the bootstrapping procedures mentioned in Chapter 3 are presented. These include the bootstrapped residuals and the bootstrapped pairs. The distribution of the coefficients are then briefly considered to check if there are any obvious irregularities.

Results for the bootstrapped residuals.

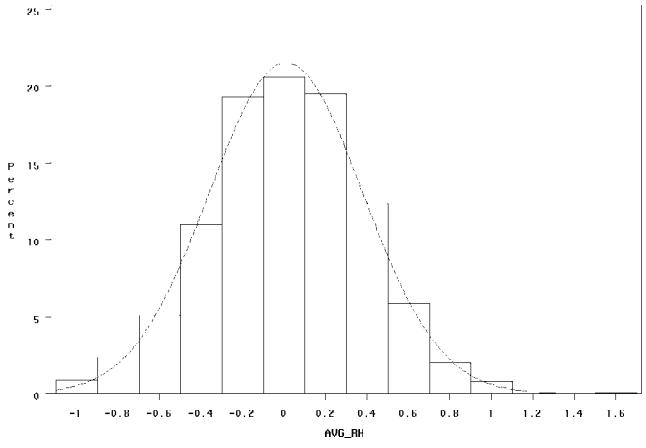

Figure 6.7: Histogram of regression coefficient for relative humidity using the bootstrapped residuals approach.

The histogram for the relative humidity coefficient appears to be reasonably normally distributed in Figure 6.7. There does not seem to be significant skewness and / or kurtosis. However, the histogram clearly passes through zero, implying insignificance of the estimate for the relative humidity coefficient.

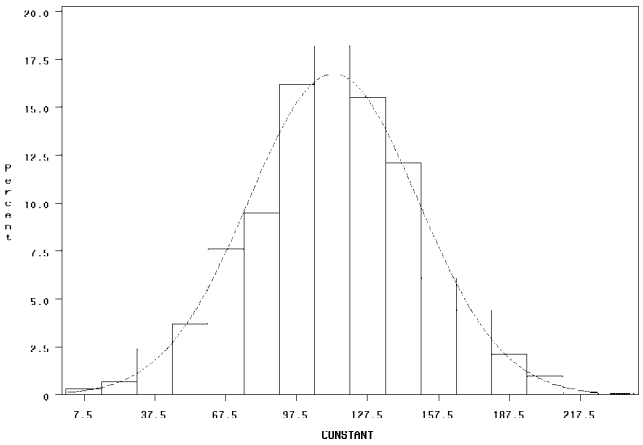

Figure 6.8: Histogram of regression coefficient for constant using the bootstrapped residuals approach.

The distribution of the regression coefficient for the constant term also appears normal with no unusually distributed coefficients.

In Figure 6.9, one can clearly see that certain regression coefficient estimates take on the value of zero. This is an indication of the insignificance of the estimate.

In Figure 6.10, there appears to be slight positive skewness, with the skewness statistic equal to 0.28942206, which is greater than two times the standard error for skewness which equals 0.256073759. However, the 95% confidence interval does not include zero and this implies skewness is at a minimum. The histogram here also appears to pass through zero, once again implying insignificance of the estimate.

Kurtosis and skewness are not prevalent in Figure 6.11. The coefficient estimates appear to be normally distributed.

In Figure 6.12, the regression coefficients appear to be normally distributed, with no

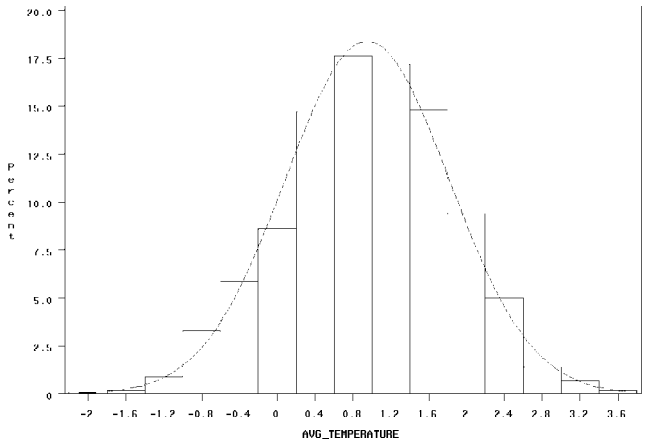

Figure 6.9: Histogram of regression coefficient for average temperature using the bootstrapped residuals approach.

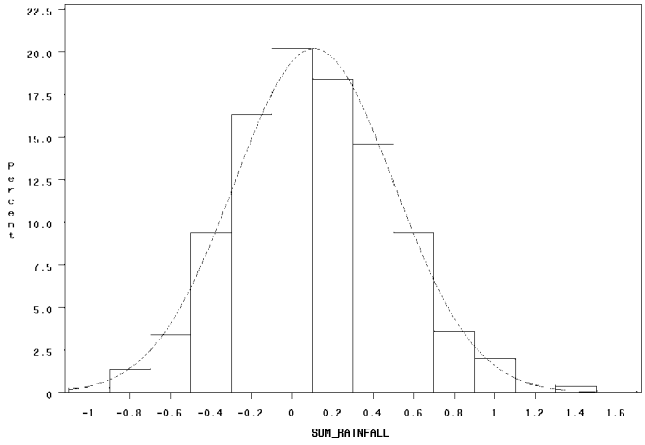

Figure 6.10: Histogram of regression coefficient for sum rainfall using the bootstrapped residuals approach.

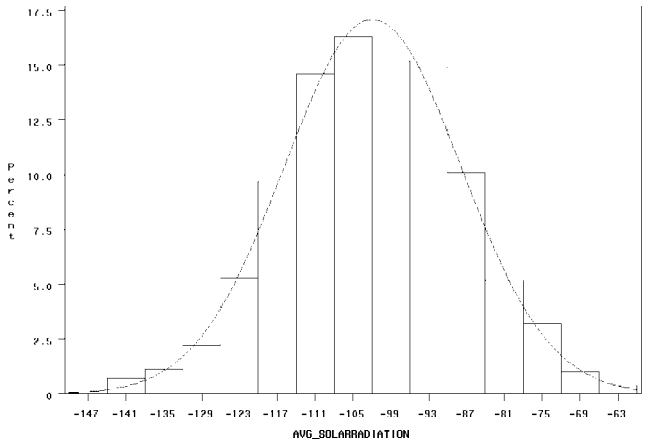

Figure 6.11: Histogram of regression coefficient for average solar radiation using the bootstrapped residuals approach.

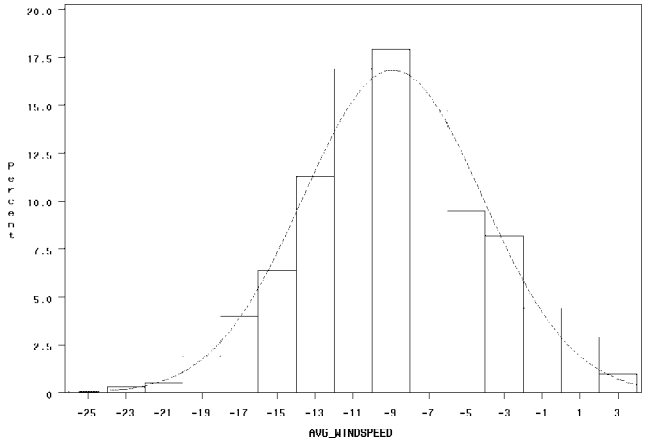

Figure 6.12: Histogram of regression coefficient for average wind speed using the bootstrapped residuals approach.

evident skewness and / or kurtosis .

Results for the regression bootstrap.

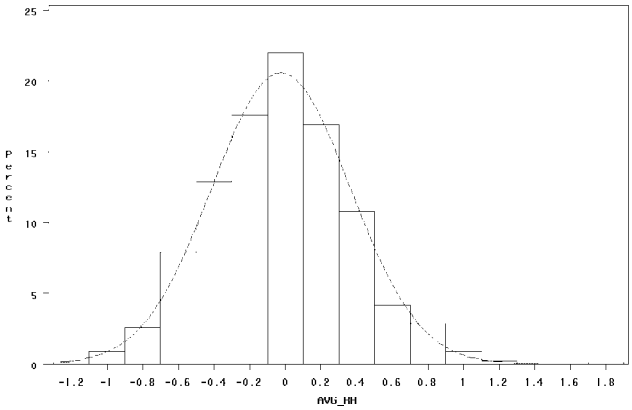

Figure 6.13: Histogram of regression coefficient for relative humidity using the regression bootstrap approach.

The histogram for the relative humidity coefficient appears to be reasonably normally distributed in Figure 6.13. There appears to be slight positive skewness but this is insignificant, with a skewness statistic of 0.1892344. As with the residual bootstrap, the histogram clearly passes through zero thus implying insignificance of the estimate for the relative humidity coefficient.

The distribution of the regression coefficient for the constant term also appears to be normal, with insignificant negative skewness.

In Figure 6.15 as in Figure 6.9 one can clearly see that certain regression coefficient estimates take on the value of zero. As result, the estimate of the regression coefficient for temperature is considered to be insignificant.

Figure 6.16 does appear to have positive skewness, with the skewness statistic equal to 1.21225938, which is greater than two times the standard error for skewness which equals 0.256073759. However, the 95% confidence interval does not include zero and

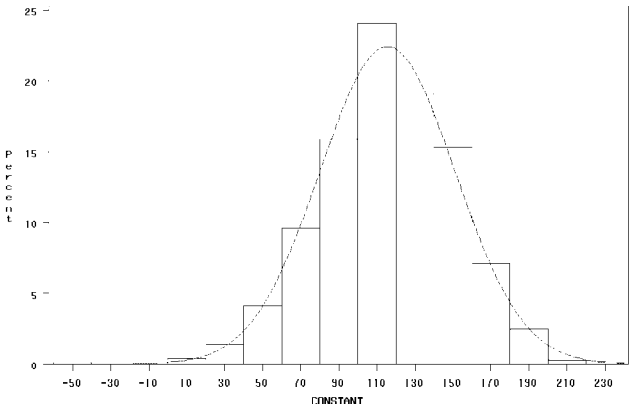

Figure 6.14: Histogram of regression coefficient for constant using the regression bootstrap approach.

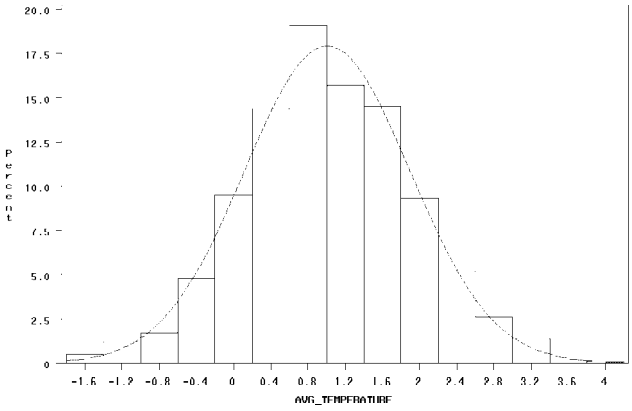

Figure 6.15: Histogram of regression coefficient for average temperature using the regression bootstrap approach.

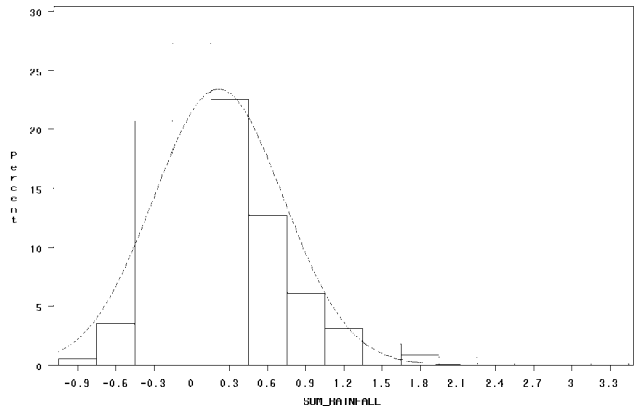

Figure 6.16: Histogram of regression coefficient for sum rainfall using the regression bootstrap approach.

this implies skewness is at a minimum. In the histogram as in Figure 6.9, the regression coefficients also pass through zero, implying insignificance of the estimate.

Kurtosis and skewness are not prevalent in Figure 6.17. The coefficient estimates appear normally distributed, with no evident skewness and / or kurtosis.

The Figure 6.18 appear normally distributed, with no evident skewness and / or kurtosis.

In general, both the approaches to the bootstrapping revealed similar results in that the estimates for the regression coefficients representing the relative humidity, rainfall, and temperature covariates were all insignificant because they took on values of zero when bootstrapped. The remainder of the coefficient estimates all appear to be significant.

6.4.5 SIMEX Results

The SIMEX method from Chapter 4 is used with the Dendrometer data. First, a simple programme written in PROC IML is used to try and assess the measurement error in the data set; thereafter the same analysis is carried out using the SIMEX command in STATA to verify results.

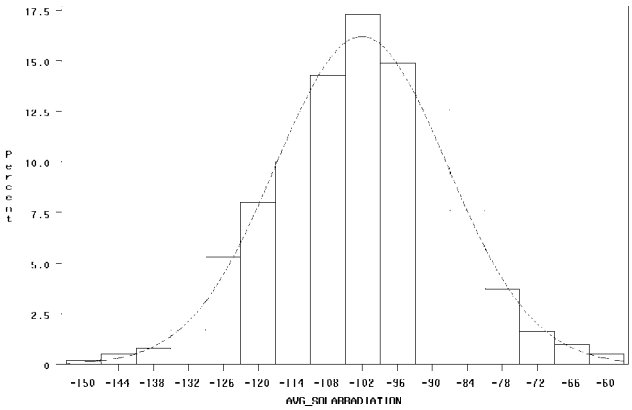

Figure 6.17: Histogram of regression coefficient for average solar radiation using the regression bootstrap approach.

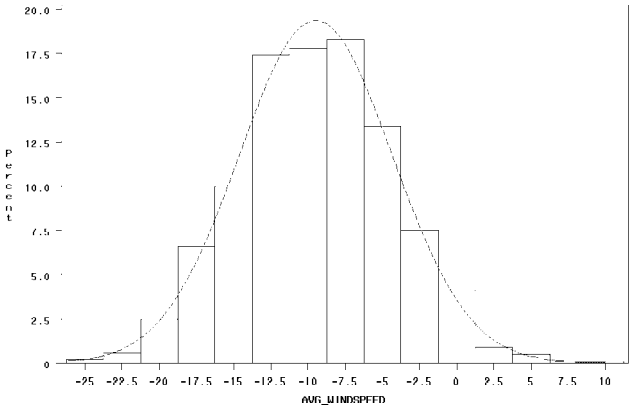

Figure 6.18: Histogram of regression coefficient for average wind speed using the regression bootstrap approach.

Assuming that measurement error is only in the relative humidity variable, and using the approach given by Carroll et al. (1995) for multiple linear regression with additive measurement error, the results obtained from PROC IML are presented in Table 6.5. These results comprise the coefficient estimates for varying values of lambda.

Table 6.5: Measurement error analysis results from PROC IML

lambda	constant	avg temp	sum rainfall	avg solarrad	avg windspeed	w1
-1	113.09525	0.9978914	0.1466752	-101.9195	-9.262998	0.0019776
-0.9	113.09525	0.9978914	0.1466752	-101.9195	-9.262998	0.0019776
-0.8	113.09525	0.9978914	0.1466752	-101.9195	-9.262998	0.0019776
-0.7	113.09525	0.9978914	0.1466752	-101.9195	-9.262998	0.0019776
-0.6	113.09525	0.9978914	0.1466752	-101.9195	-9.262998	0.0019776
-0.5	113.09526	0.9978914	0.1466752	-101.9195	-9.262998	0.0019776
-0.4	113.09526	0.9978914	0.1466752	-101.9195	-9.262998	0.0019775
-0.3	113.09526	0.9978914	0.1466752	-101.9195	-9.262998	0.0019775
-0.2	113.09526	0.9978914	0.1466752	-101.9195	-9.262998	0.0019775
-0.1	113.09526	0.9978914	0.1466753	-101.9195	-9.262998	0.0019775
0	113.09526	0.9978914	0.1466753	-101.9195	-9.262998	0.0019775
0.1	113.09526	0.9978914	0.1466753	-101.9195	-9.262998	0.0019775
0.2	113.09526	0.9978914	0.1466753	-101.9195	-9.262998	0.0019775
0.3	113.09526	0.9978914	0.1466753	-101.9195	-9.262998	0.0019775
0.4	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774
0.5	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774
0.6	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774
0.7	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774
0.8	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774
0.9	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774
1	113.09527	0.9978914	0.1466753	-101.9195	-9.262999	0.0019774

The variable, *w1*, in Table 6.5 represents the error free predictor, *avg_rh*. Recall the SIMEX estimate for each regression coefficient corresponds to that obtained when lambda equals -1. The naive estimate corresponds to that obtained when lambda equals 0. So that the true error-free estimate for the constant is 113.09525, and the estimate with error present in the relative humidity is 113.09526. The remaining coefficient estimates can be interpreted similarly. As a result, the estimates for the regression coefficients from the linear regression analysis in Table 6.1 correspond to the estimates when lambda equals 0. When analysing the results shown in Table 6.5, one can see that measurement error does not have a large impact on the parameter estimates, i.e., estimates when lambda equals -1 do not differ significantly to those estimates when lambda equals 0.

To verify the results of the programme written in PROC IML, the SIMEX analysis using the STATA built-in SIMEX commands for additive measurement error (Hardin et al. 2003) was repeated. The results for the analysis are given in Tables 6.6 and 6.7.

Table 6.6: Results for measurement analysis in STATA

Simulation Extrapolation — No. of obs = 366
Bootstrap reps = 199
Residual df = 360 — Wald F(5,360) = 26.75
Prob > F = 0.0000

Variance Function : V(u) = 1 [Gaussian]
Link Function : g(u) = u [Identity]

avg_stemra~s	Coef	Std. Err	t	P>|t|	[95% Confidence	Interval]
avg_temp	.9964338	.7934304	1.26	0.210	-.5639071	2.556775
sum_rain	.1455442	.5399203	0.27	0.788	-.9162499	1.207338
avg_solarr	-101.8837	14.19389	-7.18	0.000	-129.7971	-73.97034
avg_windspeed	-9.253349	4.796409	-1.93	0.054	-18.68585	.1791515
w1	.0051179	.3835802	0.01	0.989	-.7492216	.7594573
_cons	112.8285	34.95834	3.23	0.001	44.08024	181.5767

As can be seen in Table 6.6, average solar radiation is significant, and average wind speed is almost significant. The entire analysis is significant with F-prob = 0.0000. The error-free predictor for the relative humidity is insignificant with a p-value of 0.989. This value has not changed much from the p-value in the linear regression analysis which was 0.9958. In general, the p-values are more or less the same for the linear regression analysis and the measurement error analysis.

When comparing the parameter estimates obtained from PROC IML one sees that those estimates strongly resemble the linear regression estimates shown in Table 6.1. The SIMEX estimate for *w1* with STATA is 0.00511788, and the same variable estimate with PROC IML is 0.0019776. This is a difference of 0.00314038.

With STATA the coefficient estimates for varying values of lambda are also given, and these values are presented in Table 6.7. The differences in the estimates here are bigger than those obtained from the PROC IML programme shown in Table 6.5, especially with the relative humidity predictor.

A graphical representation of the SIMEX estimates from STATA are presented in

Table 6.7: The lambda matrix for measurement error variance in STATA

lambda	intercept	avg temp	sum rainfall	avg solarrad	avg wind speed	w1
-1	112.82846	.99643376	.14554424	-101.8837	-9.2533491	0.00511788
0	113.09526	.99789142	.14667524	-101.91948	-9.2629981	.0019775
.1	113.16246	.99816037	.14690213	-101.92898	-9.2655892	.00121548
.2	112.98452	.99748598	.1462997	-101.90476	-9.2586023	.00322973
.3	112.80668	.99676985	.14574031	-101.87963	-9.2521712	.00525467
.4	114.49591	1.0032668	.15131525	-102.10991	-9.317525	-0.01390335
.5	113.55027	.99970302	.14821503	-101.98117	-9.2809751	-0.00319159
.6	113.08521	.99791711	.14669131	-101.91937	-9.2624322	0.00208247
.7	112.92795	.99729624	.14608752	-101.89668	-9.2572478	0.00387841
.8	111.76835	.99281524	.14225722	-101.73878	-9.2104912	0.01699888
.9	113.22286	.99845548	.14713442	-101.93621	-9.2692477	0.00052839
1	113.73726	1.000304	.14875607	-102.00733	-9.2882965	-0.00527852

Figure 6.19. In Figure 6.19 the SIMEX graph for *avg_temperature* depicts how the naive estimate differs from the SIMEX estimate. For lambda equal to 0 the naive estimate is 0.99789142; as one extrapolates back to lambda equal to -1, the SIMEX estimate equals 0.99643376. The the remaining graphs in Figure 6.19 are interpreted in a similar way. The naive estimate for *sum_rainfall* equals 0.14667524, with the SIMEX estimate being 0.14554424. For *avg_solarradiation* the naive estimate equals -101.91948, and the SIMEX estimate is -101.8837. Notice the individual plots in Figure 6.19 give a clear representation of how the estimates for each regression coefficient change over the lambda values. With *avg_windspeed* the naive estimate equals -9.2629981, and the SIMEX estimate equals -9.2533491. The *w1* has a naive estimate of 0.0019775, and a SIMEX estimate of 0.00511788. Lastly, the *_cons* has a naive estimate of 113.09526, and a SIMEX estimate of 112.82846. Recall that the naive estimates represent those estimates for the regression coefficients when measurement error is unaccounted for, and the naive estimates correspond to those regression estimates presented in Table 6.1. By carefully examining the naive and SIMEX estimates it can be seen that the naive estimates for *_cons*, *avg_temperature*, *sum_rainfall*, and *avg_windspeed* are reliable up to two significant figures. The naive estimate for *avg_solarradiation* is reliable up to three significant figures. However, the naive estimate for *w1* is not reliable: with a naive estimate of 0.0019775, and a SIMEX estimate of 0.00511788. The naive estimate is only 38.64% of the SIMEX estimate and the remaining 61.36% is unaccounted for.

Therefore the conclusion is reached that there exists a large amount of error within the relative humidity covariate.

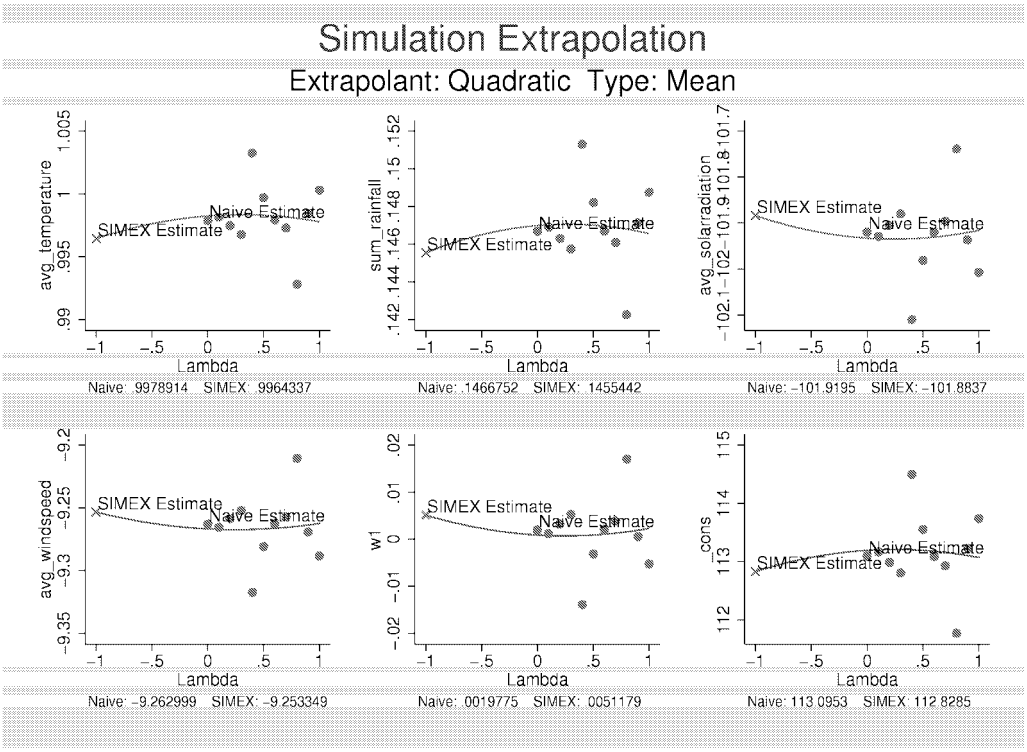

Figure 6.19: SIMEX graphs.

6.5 Summary

The results for the linear regression analysis of the Dendrometer data show that only solar radiation are significant. In general, the results are not satisfactory.

The method to check for the presence of multicollinearity was explained in Section 6.3.1, and it was found that multicollinearity plays a small role in the data and this multicollinearity is therefore considered to be insignificant.

In an effort to try and improve model results, measurement error analysis was carried out. This analysis was performed primarily because of the evident presence of measurement error with the predictor for relative humidity. Recall the reading for relative

humidity could not exceed a value of 100 and as a result the data had been modified to satisfy this condition. However, in actual reality, readings which exceeded a 100 was observed.

The results for the asymptotic approach with implied measurement error does not severely affect the data.

The perturbation approach, which is favoured by numerical analysts and is said to be more attractive, is essentially studies the effect of measurement error on the regression coefficients when the data is perturbed by small amounts. Recall that the perturbations correspond to errors of measurement, and therefore with the Dendrometer data only the predictor representing relative humidity was perturbed. For this analysis the relative bounds are presented. The relative humidity showed a percentage error of 2766.15% which is very large. This is the largest percentage error in the perturbation analysis.

The simulation results in which the error-prone predictor is added to the simulated error and analysed as usual with the rest of the data a large number of times are presented. The predictor for rainfall takes on coefficients equal to zero when simulated. This predictor is therefore considered to be insignificant with regard to the simulation approach.

The two bootstrapping approaches agree with each other in that the regression estimates for relative humidity, rainfall, and temperature are considered to be insignificant. However, these results somewhat disagree with the simulation approach.

The SIMEX method for assessing measurement error with the Dendrometer data was performed twice: the first time, by means of a written programme in PROC IML, and the second time, by means of SIMEX commands provided in STATA.

The SIMEX programme in PROC IML follows the approach for additive measurement error with multiple linear regression, when only one predictor is exposed to measurement error. The SIMEX estimate from the PROC IML programme equals 0.0019776, and the naive estimate equals 0.0019775 for relative humidity. This indicates that the measurement error has a very small effect on the relative humidity regression co-

efficient. In general, the SIMEX estimates for all predictor variables do not differ drastically when compared to the naive estimates with the PROC IML programme.

The SIMEX results from STATA are different to those obtained in PROC IML. As mentioned, it is found in STATA that the relative humidity has with it a large amount of measurement error.

The simulation and bootstrapping approaches somewhat agree with the linear regression analysis in that the relative humidity, temperature, and rainfall are insignificant. The intercept and solar radiation are significant, and the wind speed is almost significant. In general, the perturbation and SIMEX approach show that relative humidity has measurement error that is significant.

Table 6.8: Summary results for measurement error analysis

Estimates for regression coefficient	Linear model	Perturbation approach	Simulation approach	Bootstrapped residuals approach	Regression bootstrap approach	SIMEX with PROC IML	SIMEX using STATA
Constant	113.09526	108.26505 - 117.92547	110.147293	113.015583	115.571302	113.09525	112.8285
avg_temperature	0.99789	0.97916 - 1.01662	0.98879209	0.95014609	1.0073554	0.9978914	0.9964338
sum_rainfall	0.14668	0.13070 - 0.16266	-0.0736768	0.1084463	0.21154049	0.1466752	0.1455442
avg_rh	0.00198	-0.05278 - 0.05674	0.0379273	0.00877197	-0.0241982	0.0019776	0.0051179
avg_solarradiation	-101.91948	-102.57895 - -101.26001	-103.41564	-101.88646	-102.18377	-101.9195	-101.8837
avg_windspeed	-9.26300	-9.45214 - -9.07386	-8.3328478	-8.8901055	-9.4809563	-9.262998	-9.253349

A summary of the measurement error results is presented in Table 6.8. The second column contains the regression coeffcients for the linear analysis. The third approach column contains the bounds for the estimated regression coefficients when the perturbation approach is used. The simulation, bootstrap residuals, and regression bootstrap approach columns, contain the average of the 1000 regression coefficients simulated for each of these approaches. Lastly, the regression coefficient estimates obtained from the SIMEX method when PROC IML is used, and those obtained from the SIMEX method when STATA is used, are given in the last two columns in Table 6.8. From

the results shown in Table 6.8, it can clearly be seen that measurement error has an effect on the estimates of the regression coefficients, especially those obtained for the relative humidity coefficient.

Chapter 7

Conclusion

A valuable tool in statistics is fitting models to data and drawing vital conclusions from the results obtained. It is crucial in any statistical analysis that the model being used is correct. In order to assess the correctness of a model, diagnostic measures must be taken. An important aspect of modeling that has recently been developed, but is not rigorously made use of in the standard modeling environment, is the detection of the presence of measurement error in data. The main aims of this thesis are to illustrate theoretically and practically, linear model diagnostics, and to assess the effect measurement error has on the linear model.

At first the general linear model is presented. Specifically the multiple linear model, the weighted linear model, and the generalized linear model are considered. The general linear model theory for estimation, hypothesis testing, and confidence intervals are given. For the weighted linear model, the weighted least squares are used. The basic theory which includes the link and variance functions, the exponential family, and the estimation are given for the generalized linear model.

Detailed diagnostics for these models are presented. With the linear, and the weighted linear model the diagnostic measures are the same. These measures include residual analysis, plotting methods, measures of influence, and multicollinearity diagnostics. With the generalized linear model, the most common diagnostic measures are pre-

sented. These measures are somewhat similar to those used for the general linear model. They include the deviance measure, leverage, residual analysis, Cook's distance, diagnostic plots, model statistics, and assessment of the link function.

The theory for measurement error with linear models is discussed. This section provides an overview of some of the basic as well as the well known methods used to assess measurement error. The basic measurement error techniques which apply strictly to the case of additive measurement error include the asymptotic approach, the perturbation approach, the simulation approach, and the bootstrap approach. Also, the SIMEX method is considered in detail.

The above mentioned covers the theoretical aspect of the thesis, which provides the basis for application.

The application of linear model diagnostics is demonstrated by using the Durban South data set. The basic biological background information is first presented. The multiple linear regression model when applied to the data, does not account for the majority of explanatory variables. The diagnostic analysis performed further revealed a number of extreme observations existing within the data set, and the model deviated from normality. In general, the multiple linear model was not an appropriate choice. In an effort to find a model that would better fit the data set, the log transformation, the inverse transformation, and the generalized linear model were made use of.

With the log transformation there was an improvement with the extreme observations and the normality assumption when certain extreme observations were deleted. However, the amount of variation that the explanatory variables accounted for was still too small. The inverse transformation did not improve the model fit. Both transformations were unsuccessful.

It is seen that the data appear to follow a distinct gamma distribution. This is shown in Figure 5.6. Since the figure does not appear normal, and careful scrutiny reveals that the residuals behaves like a S-shaped. As a result of this, a generalized linear model with gamma distribution was considered. Furthermore, since the response variable lies

between 0 and 1, the generalized linear model with beta distribution is also looked at. Basic diagnostics have been considered. The residual plots and, in particular, tests of the link function were carried out to try and ensure a good model was chosen. According to the link function tests, it was discovered that only the generalized linear model with beta distribution and logit link appeared to be the most suitable. However, the explanatory variables were found to be insignificant.

As was seen with the Durban South data, the normality plot indicated that the data set was gamma distributed. However, the generalized linear model with gamma distribution was ruled out when the test for the link function showed that all links were inappropriate. The generalized linear model with beta distribution and logit link was the most appropriate, since the test for the logit link appeared somewhat linear, as previously mentioned. Many extreme observations were identified during the diagnostic check. As argued in this thesis, it is vital that diagnostic checks are performed.

In many situations, measurement error plays a significant role in analysis. When a model is used to fit data, either the model fits the data well or it does not fit the data well. If the effect measurement error has on the data has not been considered then the results obtained may be inaccurate. The Dendrometer data were used to illustrate the effects of measurement error.

Multicollinearity diagnostics were looked at, and as expected collinearity did exist. This was as a result of the nature of the data set. However, the collinearity was at a minimum and the VIF did not reveal that the multicollinearity was in any way severe.

With the Dendrometer data, a multiple linear regression analysis was carried out. It was found that the model did not represent the data. Because of the evident presence of measurement error within the Dendrometer data, the various measurement error techniques were applied to the data. The perturbation approach revealed a large amount of error within the error-prone variable. This result is expected. With the simulation approach it was found that two of the explanatory variables were insignificant when measurement error was accounted for. The bootstrapping approaches showed

that three of the explanatory variables were insignificant when measurement error was accounted for. Lastly, the SIMEX approach revealed a large amount of measurement error with the error-prone variable. This measurement error, as can be seen in Table 6.8, has an effect on the estimates for the regression coefficients. The estimate of the error-prone variable when measurement error is not accounted for and the estimate obtained when measurement error is accounted for differ drastically.

The Dendrometer data analysis clearly illustrates the effect measurement error has on the model. If the measurement error is ignored, the estimates of the regression coefficients obtained are sometimes inaccurate.

One can see that carrying out data analysis does not involve blindly applying a model. Once model results are obtained, it must be confirmed that the assumptions associated with the model are valid. The other important aspect of statistical modeling is carefully studying the data and checking for the presence of measurement error. If measurement error exists within the data, measurement error models must be looked at.

With regard to the Durban South data set, an appropriate model that represents the data has not been found. The linear modelling which included the linear models, as well as the general linear model does not fit the data. One can safely say that the data cannot be modeled using these techniques. It is suggested that nonlinear analysis, and nonlinear generalized linear modeling be used to try and fit the Durban South data. These techniques are beyond the scope of this thesis. However, it must be noted that nonlinear diagnostic tools for the nonlinear, and nonlinear generalized linear modeling has not yet been well developed. The future direction of this study is therefore nonlinear regression and nonlinear generalized linear models.

Bibliography

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1989). *Statistical modelling in GLIM.* Oxford University Press : New York.

Beers, H., and Berkow, R. (1999) *The Mark Manual of Diagnosis and Therapy.* Gary Zelko : Merck and Co., Inc.

Behrman, R.E., and Kliegman, R.M. (1990) *The Respiratory System.* W.B Saunders : Philadelphia.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics : Identifying data and sources of collinearity.* Wiley : New York.

Berhane, K., McConnell, R., Gilliland, F., Islam, T., Gauderman, W.J., Avol, E., London, S.T., Rappaport, E., Margolis, H.G., and Peters, J.M. (2000). Sex-specific Effects of Asthma on Pulmonary Function in Children. *American Journal of Respiratory and Critical Care Medicine,* 162(5), 1723-1730.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models.* Chapman and Hall : London.

Carroll, R.J., and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall : United States of America.

Chatterjee, S., and Hadi, A.S. (1988). *Sensitivity Analysis In Linear Regression.*

Wiley : United States of America.

Cook, J.R., and Stefanski, L.A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association,*, 89, 1314-1328.

Coovadia, H.M., and Wittenberg, D.F., (2000). *Paediatrics and child health.* Oxford University Press : Southern Africa.

Clark, N., Wynne, R., and Schmoldt, D. (2000). A review of past research on dendrometer. *Forest Science,* 46(4), 570-576.

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models.* Chapman and Hall : Great Britian.

Draper, N.R., and Smith, H. (1998). *Applied Regression Analysis.* John Wiley and Sons : United States of America.

Ellis, K.J., Abrams S.A., and Wong W.W. (1999). Monitoring childhood obesity: assessment of the weight/height index. *American Journal of Epidemiology,* 150(9), 939-946.

Ellis, K.J., Abrams S.A., and Wong W.W. (1999). Monitoring childhood obesity: assessment of the weight/height index. *American Journal of Epidemiology,* 150(9), 939-946.

Fox, J. (2005). Lecture Notes : Unusual and Influential Data.

Fuller, W.A. (1987). *Measurement Error Models.* Wiley : New York.

Guyton, A.C. (2000). *Textbook of Medical Physiology.* W.B Saunders : Philadelphia.

Hardin, J.W, and Carroll, R.J (2003). Measurement Error, GLMS, and Notational Conventions. *The Stata Journal,*1-12.

Hardin, J., and Hilbe, J. (2001). *Generalized Linear Models and Extensions.* Stata Press : United States of America.

Hardin, J.W, Schmiediche, H., and Carroll, R.J. (2003). The Simulation Extrapolation Method for Fitting Generalized Linear Models with Additive Measurement Error. *The Stata Journal,*1-12.

Harik-Khan, R.I., Wis, R.A., and Fleg, J.L.(2001). The effect of gender on the relationship between body fat distribution and lung function. *Journal of Clinical Epidemiology,* 54(4), 399-406.

Hogg, R.V., and Craig, A.T. (1995). *Introduction to Mathematical Statistics.* Prentice-Hall, Inc : United States of America.

Lederer, W., and Kuchenhoff, H. (2006). The simex Package.

McCullagh, P., and Nelder J.A. (1989). *Generalized Linear Models.* Chapman and Hall : Great Britain.

McCulloch, C.E. (2001). *Generalized, Linear, and Mixed Models.* John Wiley and Sons : New York.

Montgomery, D.C., Peck E.A., and Vinning G.G. (2001). *Introduction to Linear Regression Analysis.* John Wiley and Sons : United States of America.

Myers, R.H., Montgomery, D.C., and Vinning, G.G. (2002). *Generalized Linear Models : With Applications in Engineering and the Sciences.* John Wiley and Sons: New York.

Nasirumbi, P.O. (2006). *Longitudinal survey data analysis.*

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society,* 135, 370-384.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied Linear Statistical Models. Fourth Edition.* McGraw-Hill Companies : United States of America.

Osborne, J. (2002). *Notes on use of data transformation.*

Pelosi, P., Caironi, P., and Gattinoni, L. (2002). *Perioperative respiratory management of obese patient.*

Piegersch, W.W. (2003). *Notes and Extensions for a course in generalized linear models.*

Chapman, R.S., Hadden W.C., and Perlin, S.A. (2003). Influences of Asthma and Household Environment on Lung Function in Children and Adolescents. *American Journal od Epidemiology,* 158, 175-189.

Stefanski, L.A. (2000). Measurement Error Models. *Journal of the American Statistical Association,* 452(95), 1353-1358.

Tabachnick, B.G., and Fidell, L.S. (1996). *Using Multivariate Statistics.* New York: Harper Collin.

Ulger. Z., Demir, E., Tanac, R., Goksen, D., Gulen, F., Darcan, S., Can, D., and Coker, M. (2006). *The effect of childhood obesity on respiratory function.* McGraw-Hill Companies : United States of America.

Zewotir, T., and Galpin, J.S. (2004). The behaviour of normal plots under non-normality for mixed models. *South African Statistical Journal,* 38, 115-138.