Clinical Applications of errors-in-variables methodology

By

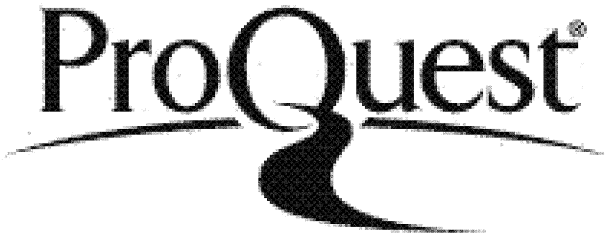
Sophie Gilchrist

A thesis submitted to Cardiff University in accordance with the requirements for the degree of Philosophiæ Doctorate

Department of Epidemiology, Public Health & Statistics

Cardiff University

January 2005



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Summary

In many epidemiological contexts the effects of risk factors on the occurrence of a disease are assessed by means of logistic regression. Many such risk factors are not recorded exactly but are subject to measurement error. This is particularly true for risk factors such as blood pressure and cholesterol levels, which are then related to heart conditions.

Much has been written on the linear regression 'errors-in-variables' problem, of regressing a continuous response on an explanatory variable which is subject to error. It is widely known that in linear regression measurement error acts to weaken the association between variables and, the same is shown to be true when the outcome is binary. This thesis considers the logistic regression errors-in-variables problem where it is assumed that the explanatory variable is measured with error.

The logistic regression measurement error model is more complex than the linear regression measurement error model. This has led to many methods being designed and implemented. This thesis brings together a number of elements and methods of the logistic regression measurement error model, to bring understanding of the effects

of measurement error and clarity as to which methods are best used in practice given certain assumptions. Methods are considered from both a classical as well as Bayesian approach. Though it has been theoretically proven that all the model parameters can be estimated from a single study data set, our investigations have found no practical use of this theory, thus the properties of the logistic regression measurement error likelihood have also been examined. To conclude, the investigated methods are compared with a current method used in the literature showing that there are better correction methods available and therefore, should be more commonly used in practice.

This thesis will enable practical studies to be designed to use the most accurate method for correcting for measurement error, thus leading to the best estimates of the true relationship between an explanatory variable and a disease status.

Chapter 1

1 Introduction

1.1 Background

The causes of most common diseases are not fully understood. Interest lies in both the genetic origins of disease as well as external environmental factors. Both areas are of great concern and have warranted extensive research. For example, what effect does blood pressure have on heart disease? What role does diet have on heart disease? Do certain genotypes affect the risk of different forms of cancer or does living near a source of radiation emission also increase the risk of cancer?

To answer such questions we need to create a model to relate the outcome, in this case that there is a disease present, to the risk factor in question. A very commonly used model for this is logistic regression. This is analogous to the more widely used linear regression, which, in its simplest form, deals with the problem of fitting a line through a scatter plot of data but in more complex situations is concerned with establishing

relationships between a continuous outcome variable and a number of explanatory variables.

These models are well understood and methods have been developed for fitting them to data and for making inferences based on the results. In many situations, however, there is an added complication. The standard model has a vital assumption, namely that the explanatory variables are known exactly and are not subject to error. This is extremely important both in fitting the model and in making inferences based on the model.

In designed studies it is sometimes a plausible assumption that an explanatory variable is measured without error as the values may be determined by the experimenter. In some observational studies it may also be reasonable. For example, if a study is investigating the effect of age and gender on the prevalence of a medical condition, then these should be known exactly.

In many other cases, however, this will not be the case. For example, if we are investigating the effect of the consumption of saturated fat on the risk of a person developing ischaemic heart disease then we need to know the amount consumed by each person in the study. What is relevant is the consumption over a long period of time; is it possible to know this exactly? Clearly the answer is that generally such a value will not be known exactly but will be estimated, and the estimates will be used within the study to determine associations between saturated fat intake and the development of ischaemic heart disease. These estimates will be subject to error and

this may affect the nature of the association between fat consumption and ischaemic heart disease.

Situations of this type are very common in medical contexts and so there is a great need for methods which will take into account this inexact measurement of explanatory variables when modelling relationships. It is widely known that in linear regression measurement error acts to weaken the association between variables and, as we shall see, the same is true when the outcome is binary. It is important, therefore, that corrections are made which will lead to correct inferences being drawn. Before explaining how this thesis will investigate how this may be achieved, a range of more detailed examples are presented.

1.2 Examples of measurement error

The term measurement error will be used to refer to situations in which an explanatory variable is not known exactly. It can arise for many different reasons and in some cases the term 'measurement error' is not entirely appropriate, but the standard convention of using this terminology for all situations in which an explanatory variable is not known precisely will be followed here.

A number of examples follow which illustrate different types of measurement error situations. Sometimes a variable can be measured accurately and precisely by an expensive method but the costs would be too large to do this on a large scale and a cheaper but less accurate method is used. In other situations it may be difficult to measure the quantity of interest exactly and so a proxy measure must be used. In other circumstances it is hard to even define the quantity precisely.

Measurement of gestational age

In antenatal screening, algorithms which estimate the risk of a foetus being affected by Down's Syndrome use a number of markers, such as alpha-fetoprotein and chorionic gonadotrophin. These values need to be adjusted for gestational age, as they change during pregnancy, and this adjustment uses the results of a regression using gestational age as an independent variable. Gestational age is estimated from either the dates of the last menstrual period or from an ultrasound scan which records the size of a particular foetal dimension, such as the crown-rump length. This measurement is converted into a gestational age but the process is based on fitting a model to averages and is subject to error when considering an individual. Methods currently used use linear regression for the adjustment and then relate the pregnancy outcome to adjusted values but it would be possible to relate the measured values directly to outcome with this imprecise measurement of gestational age.

Measurement of blood pressure

Blood pressure is usually measured by a sphygmomanometer although other methods, including electronic ones, are possible. The values may not be exact as studies have shown that there is a certain degree of intra-observer variation. There is another issue here, in that it is well known that blood pressure varies over a time period and a single reading may not reflect that adequately (MacMahon (1990)). If the aim of a study is to study association between, say, blood pressure and the risk of strokes then what matters is the underlying 'normal' level for an individual. Random fluctuations, often considerable in magnitude, mean that a single measurement may be different from this 'normal' level and therefore will estimate it with error. This example will be discussed in greater detail in Chapter 9.

Measurement of fasting plasma glucose

Fasting plasma glucose is used in diagnosing diabetes, with values exceeding 7 mmol/l indicating the presence of the disease. A blood sample is taken after a patient has fasted, usually for 12 hours, and the glucose concentration is measured. In some individuals the level varies considerably from day to day, and if values were measured on a second occasion they may be different from those obtained initially. A study on the reproducibility of fasting plasma glucose (Ollerton et al. (1999)) considered the difference between two values obtained under the same conditions on different days and found that the difference could be as large as 2.65 mmol/l, with a standard deviation of 0.8 mmol/l. This shows that an individual value could easily be different from the underlying true value by at least 1 mmol/l. This obviously causes problems in classifying subjects as diabetic or not. Further, the level of fasting plasma glucose is often used in models assessing the risk of developing complications such as retinopathy and so measurement errors need to be accounted for in this model.

Measurement of dietary intake

There is growing evidence that diet may change the risk of many diseases, including a number of different types of cancer. For example Evans et al (2002) investigated the importance of different types of dietary fibre on the risk of colorectal cancer, and found significant associations. In order to carry out such a study the intake of fibre must be estimated for each person in the study, and many problems arise in attempting to do this.

One approach is to use a food frequency questionnaire, in which subjects are asked to record the number of times per week or month that a certain type of food is eaten. Recall will be imperfect, leading to errors, and the size of portions is often ignored. Further, a subject who knows about the purpose of the study may tend to give the answers expected, or which are more acceptable, thus introducing bias. In a study on heart disease and saturated fat consumption, participants may well under-estimate or under-report the amount of chips consumed.

An alternative approach is to weigh portions of food eaten, which are analysed for their content. It is difficult to do this on a large scale or for a long period. If it is only carried out for a few days, these may not be typical; indeed the use of such a process may well change eating habits, thus introducing more bias.

Therefore, whatever method is used, the reported levels of intake of certain foods are likely to be subject to error.

Measurement of deprivation

Deprivation is used as an explanatory variable in many health studies because, for a large number of diseases, the incidence of the disease, and also the outcome, are different in deprived people compared to affluent ones. Deprivation is not easy to measure on individuals without detailed surveys and often an area-based measure is used. There are a number of these, including the Townsend index (Townsend et al (1988)), derived from census data, and the Index of Multiple Deprivation, derived from a number of data sources. They are usually calculated for an area such as an

electoral division, although sometimes smaller census output areas are used. Using an area-based measure for an individual clearly induces error.

Drug absorption

In studying the effect of a therapeutic drug the outcome may be related to the dosage, via the production of a dose-response curve. In fact different subjects given the same dosage will have different concentrations in their blood and it is probably this concentration that is crucial, rather than the nominal level in the amount given. Thus modelling the outcome in terms of the dosage will need to take account of the measurement error, essentially the difference between the amount ingested and the amount absorbed into the blood.

Laboratory variation

In many situations, the concentration of a chemical in the blood may be regarded as a potential risk factor. Returning to the example of antenatal screening, the levels of alpha-fetoprotein and chorionic gonadotrophin are measured in blood samples from a pregnant woman and are fed into an algorithm which estimates the risk of a foetus being affected by a chromosomal disorder. External quality schemes monitor laboratory performance by sending identical samples to laboratories who have to report the concentrations they measure in these samples. Results can show marked differences, both from each other and from the gold standard. In one instance, laboratories involved in processing blood samples for antenatal screening were asked to report the estimate of risk which would be derived from the blood sample concentrations – a function of maternal age and these concentrations. Figure 1.1 shows the resulting spread of results.

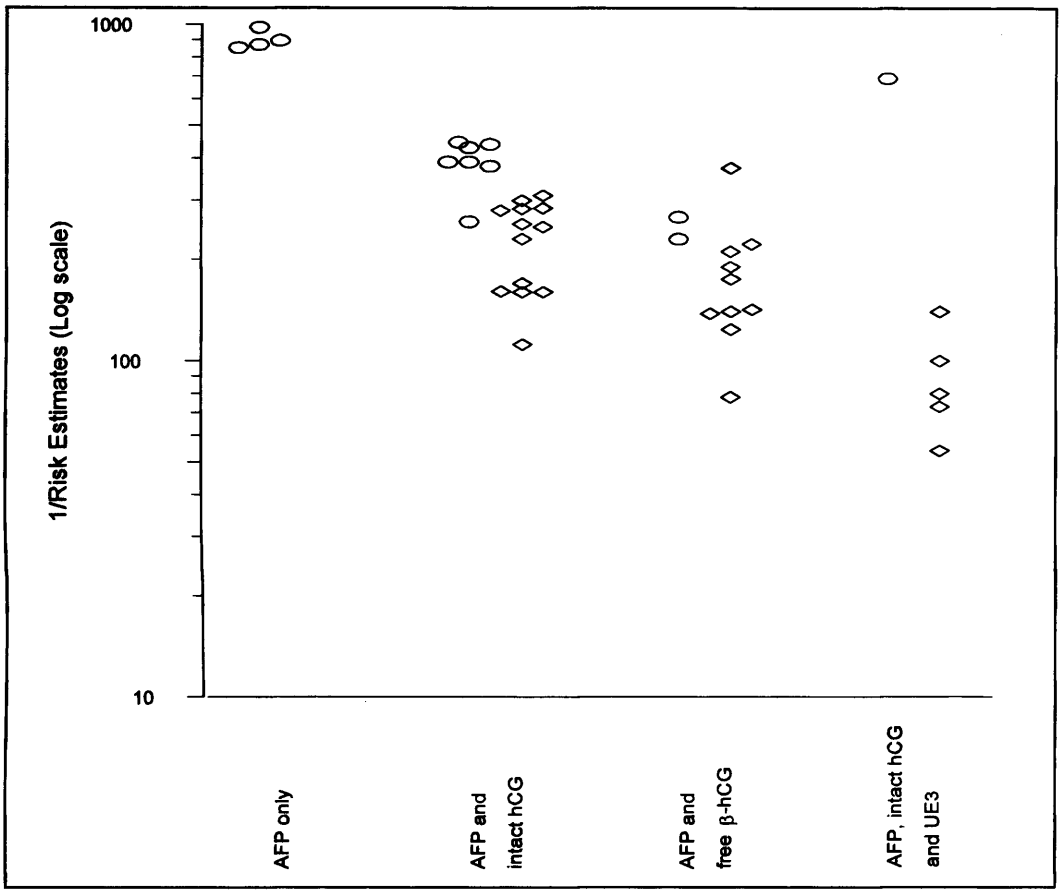


Figure 1.1: Results from antenatal screening laboratories, UK NEQAS Scheme

Differences in reported values of the levels of chorionic gonatotrophin and alpha-fetoprotein have led to very different risk estimates showing that a single measurement from a single laboratory is liable to be subject to considerable measurement error.

1.3 Outline

In section 1.2 a wide range of examples of clinical situations was described. The usual approach to this type of problem is to fit a statistical model relating the outcome variable with a set of explanatory variables. This is usually some form of regression model. Two cases must be distinguished. If the outcome is a continuous variable then the approach usually adopted is that of linear regression. If, however, the outcome is a binary one, such as disease or no disease, then a different method is required. The most widely used method is that of logistic regression and it is this model which is the main focus of this thesis.

In Chapter 2 the basic principles of this approach will be summarized. In linear regression, the estimates of the model parameters have a closed analytical expression and this makes analysis rather easier. In logistic regression, by contrast, iterative methods are required for deriving estimates. This absence of a closed form for the estimators makes the derivation of their properties, including bias and variance, much more difficult. Some simulation results will illustrate this in the absence of measurement error.

A considerable amount of attention has been given to the problem of measurement error in linear regression models and in Chapter 3 some of the key results will be summarized. As will be seen then, situations like the examples in section 1.2 fall into a number of different categories and so an early priority is to classify them into a number of different models for measurement error. Thus the chapter begins with a discussion of these different types of models and presents a unifying structure into which all can be classified. One aspect which will recur throughout the thesis is the

possible knowledge of the statistical distribution of the measurement errors and different study designs which might be employed to estimate this distribution are described there.

Chapter 4 reviews a selection of methods which have been proposed for dealing with the estimation of parameters of logistic regression models subject to measurement error. They use a variety of different approaches, such as deriving a different but related logistic regression model whose parameter estimates can be transformed into estimates for the model under investigation.

Chapters 5 and 6 deal with comparisons of these methods through a simulation exercise. The methods generally make certain assumptions about the nature of the data and the distribution of the measurement errors and the methods will be compared when these assumptions are satisfied. In addition the robustness of the methods to these assumptions will be explored by presenting simulation results for situations in which the assumptions are not satisfied.

The methods presented in Chapter 4 all use a classical, or frequentist, approach. Another statistical approach, which has become increasingly widely used in the last 10 years, is the Bayesian approach. This adopts a different stance, regarding a parameter as having a statistical distribution rather than having a fixed, but unknown, value. An outline of such methods is given in Chapter 7, together with a brief discussion of the main method used for fitting Bayesian models, namely the Markov-Chain Monte-Carlo method. Results will then be presented from the application of

these methods to the measurement error problem for logistic regression. The methods used are computationally intensive and so a smaller range of cases is explored here.

The methods described in Chapter 4 and tested in Chapters 5 and 6 all assume that the distribution of measurement errors is known; in particular, knowledge of the error variance is assumed. In linear regression it is not possible to identify the parameters of the model without an analogous assumption but it has been shown that in the case of logistic regression the parameters are actually identifiable without knowing this error variance. In Chapter 8 this issue of identifiability is discussed further and the results of Kuchenhoff, who first showed this, are presented and extended. It does appear to be difficult to use this in practice to obtain estimates and attempts to do so will be described and discussed in that chapter.

The issues and methods associated with the logistic regression measurement error model have been discussed from a theoretical view point so we conclude this study in Chapter 9 with a practical example returning to the example of inaccurate blood pressure measurements being related to heart disease. This chapter will consider a method currently used by epidemiologists to deal with measurement error in blood pressure measurements when investigating its association with heart disease or strokes. The properties of the method will be derived and compared to those discussed in chapter 4.

The logistic regression measurement error model is more complex than the linear regression measurement error model. This has led to many methods being designed and implemented. This thesis brings together a number of elements and methods of

the logistic regression measurement error model to bring understanding of the effects of measurement error and clarity as to which methods are best used in practice given certain assumptions. This will enable practical studies to be designed to use the most accurate method for correcting for measurement error, thus leading to the best estimates of the true relationship between an explanatory variable and a disease status.

Chapter 2

2 The Linear Logistic Model

2.1 Introduction

This chapter is aimed at providing the background to the linear and logistic regression models referred to in Chapter 1. In particular we look at the logistic regression model and the iterative methods that can be implemented to estimate the model parameters. Further to this, we discuss some of the issues surrounding inference and goodness of fit. This information is provided as a background before introducing the logistic regression measurement error model in Chapter 3, which will take these results a step further.

2.2 The Linear Regression Model

Linear regression is a model for the relationship between a response variable and one or more predictor variables, that is, it provides a way of predicting the value of the response variable from the predictor variables. It is assumed that there is a linear

relationship between a continuous response variable Y and a single explanatory variable X via the following model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1,...,n$$

where n is the sample size, i denotes the ith subject and ε_i is the associated measurement error or random variation in Y_i. The $\{\varepsilon_i\}$ are assumed to be independent with zero mean and constant variance σ^2. An equivalent formulation is

$$E(Y_i) = \alpha + \beta x_i$$

and the variance is

$$Var(Y_i) = \sigma^2.$$

To be able to predict the response from the predictor variable as well as understand the nature of this relationship, values for α and β must be estimated. A widely used method to estimate the model parameters is that of least squares. This method assumes that the predictor variable X is not subject to measurement error and the criterion for choosing estimates of α and β is that S is minimised, where

$$S = \sum \{Y_i - E(Y_i)\}^2 = \sum (Y_i - \alpha - \beta x_i)^2$$

By differentiating S with respect to the model parameters α and β, estimates are obtained using the following formula

$$\hat{\beta} = \frac{\sum (Y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

with estimated variance

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$$

Consider the case of m covariates $x_1, x_2, ..., x_m$ and the model

$$E(Y) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

Let $\boldsymbol{X}$ denote a matrix with n rows, row i being $(1, x_1, ...x_m)$, where x_{ij} is the value of the jth covariate in subject i. Then the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{Y})$$

where $\boldsymbol{\beta}$ denotes all the model parameters, with covariance matrix

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$$

The method of least squares is intuitively appealing and results in simple analytical unbiased estimates. If $\{\varepsilon_i\}$ have a normal distribution then these estimates coincide with the maximum likelihood estimates.

2.3 Binary Data

In many applications in medical statistics the dependent or outcome variable is a disease status. For example, it may simply denote whether the disease is absent or present. This binary variable may be coded as 0 or 1 that is, if Y_i represents the status of the ith person then $Y_i = 1$ if the disease is present and $Y_i = 0$ if not. As the probability of disease will vary between people due to a variety of different factors, we use p_i to denote the probability that person i has the disease. The dependent variable Y is then binomially distributed with $Y_i \sim Bin(1, p_i)$ with mean $E(Y_i) = p_i$ and variance $Var(Y_i) = p_i(1 - p_i)$ where $p_i = P(Y_i = 1)$. The variance in this case, unlike the linear regression case, is not constant but is related to the probability and hence the mean of the distribution.

Suppose x denotes a covariate believed to affect the probability of a person having the disease; we wish to model the relationship between Y and x. A linear regression model would give

$$E(Y_i) = p_i = \alpha + \beta x_i$$

However, this model is flawed due to the fact that it can produce values for the probability p_i greater than 1 and less than 0. Also, the further assumption of constant variance is not satisfied as $Var(Y_i) = p_i(1 - p_i)$ and so varies with the mean. A different model is required in this case. A possible way is to use a transformation of the probability. The aim of this transformation is two fold. The first is to transform p so that the probability values [0,1] are mapped to the real line, the second is so that the transformed values of p are linearly related to the explanatory variables, so that a linear model is appropriate to use in the estimation process. Various methods have been suggested for making this transformation and these are discussed in the following section.

2.4 Transformations

Since the cumulative distribution function $F(w)$ of a continuous random variable w increases from 0 to 1 and has an inverse, a possible general transformation which maps p onto the real line is

$$w = F^{-1}(p)$$

If the domain of w is $(-\infty, \infty)$ then the first objection in 2.3 is removed. Particular examples of this are the Logistic and Normal distributions, which lead to the logit and probit models that are discussed further in the next two sections.

2.4.1 The Logistic Transformation

The logistic transformation is based upon the logistic distribution whose cumulative distribution function is

$$F(w)=\frac{e^{(w-\mu)/\tau}}{1+e^{(w-\mu)/\tau}} \tag{2.1}$$

with mean μ and variance $\frac{\tau^2\pi^2}{3}$. For the logistic transformation the mean is set to zero and $\tau=1$, then

$$F(w)=\frac{e^{w}}{1+e^{w}} \tag{2.2}$$

Inverting this, we define, for a probability p, the logistic transform

$$w=F^{-1}(p)=\ln\left(\frac{p}{1-p}\right)=\text{logit}(p)$$

This transformation, mapping $[0,1]$ onto the real line, is the logistic transformation. Figure 2.1 shows the values of p and the corresponding transformed $\text{logit}(p)$ values. It can clearly be seen that $\text{logit}(p)$ is a sigmoid curve that is approximately linear for p values between 0.2 and 0.8.

A possible model for relating the probability p to an explanatory variable x is therefore

$$\text{logit}(p)=\ln\left(\frac{p}{1-p}\right)=\alpha+\beta x \tag{2.3}$$

or, equivalently,

$$p=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

2.4.2 The Probit Transformation

The probit transformation is based upon the standard Normal distribution whose cumulative distribution function is

$$\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp\left(-\frac{1}{2}w^2\right)dw$$

Defining $w=\Phi^{-1}(p)$ for $p\in[0,1]$ defines another mapping onto the real line and leads to the probit model

$$\text{probit}(p)=\Phi^{-1}(p)=\alpha+\beta x \qquad (2.4)$$

Equivalently,

$$p=\Phi(\alpha+\beta x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha+\beta x}e^{-\frac{1}{2}z^2}dz \qquad (2.5)$$

where $\Phi(.)$ is the Normal cumulative distribution function.

The probit transformation is also a symmetric sigmoid curve around p=0.5, see Figure 2.1. It is worth noting that the probit transformation sigmoid curve is in fact approximately linear between 0.1 and 0.9, which is a larger range than the logistic curve. The comparison between the logit and probit models is discussed further in section 2.4.4.

2.4.3 The Complementary Log-Log Transformation

The complementary log-log transformation is used when the probability of an event is very small or very large. Unlike logit and probit the complementary log-log function is asymmetrical. It is based on the random variable with cumulative distribution function

$$F(w) = 1 - \exp(-\exp(w))$$

where w takes the values from $-\infty$ to $+\infty$.

The inverse transformation gives

$$w = F^{-1}(p) = \ln(-\ln(1-p)) \qquad (2.6)$$

This transformation, which again maps $[0,1]$ onto the real line, is the complementary log-log transformation.

A possible model is then

$$\ln(-\ln(1-p)) = \alpha + \beta x$$

or, equivalently

$$p = 1 - \exp[-\exp(\alpha + \beta x)]$$

These three transformations are shown in Figure 2.1. The complementary log-log transformation is not symmetric about $p = 0.5$, unlike the others, but for small values of p the transformed values under the logistic and complementary log-log models are very similar.

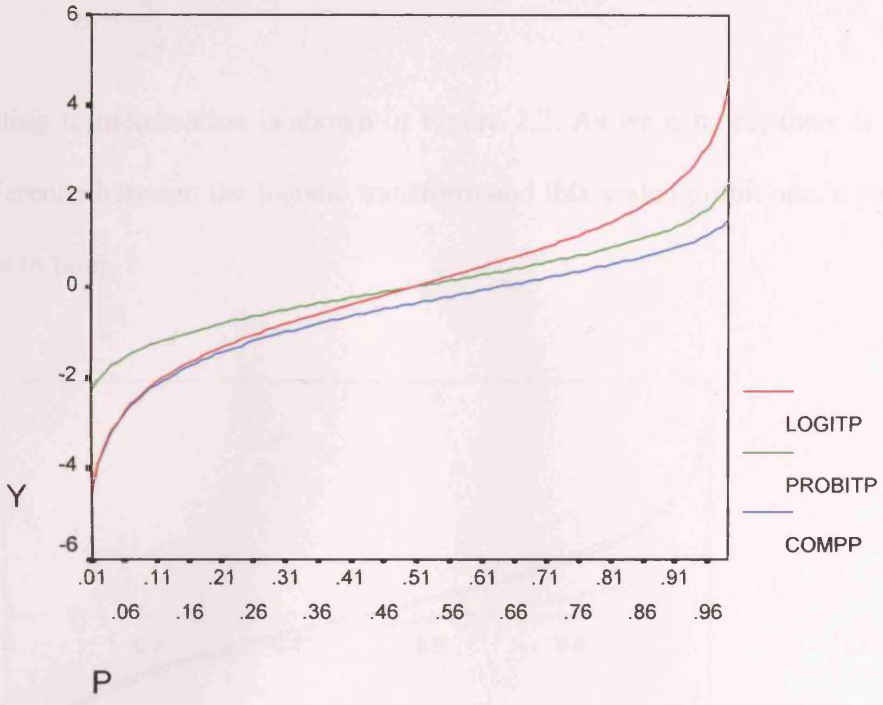


Figure 2.1: Graph to show logistic, probit and complementary log-log mean values across the range [0,1]where Y=transformed values of p from models (2.3), (2.5) and (2.6) respectively.

2.4.4 Comparison between the logistic and probit models

As can be seen from the above graph the logistic and probit transformations have approximately the same shape but give rather different transformed values. The probit transform is derived from the CDF of a standard normal distribution, variance 1, while the logistic distribution implied by (2.3) has variance $\frac{\pi^2}{3}$.

If we use distributions with the same variance then the probit model would be modified to become

$$p = \Phi\left(\frac{x}{\sigma}\right)$$

where

$$\sigma = \frac{\pi}{\sqrt{3}}.$$

The resulting transformation is shown in Figure 2.2. As we can see, there is only a small difference between the logistic transform and this scaled probit one, a point we will return to later.

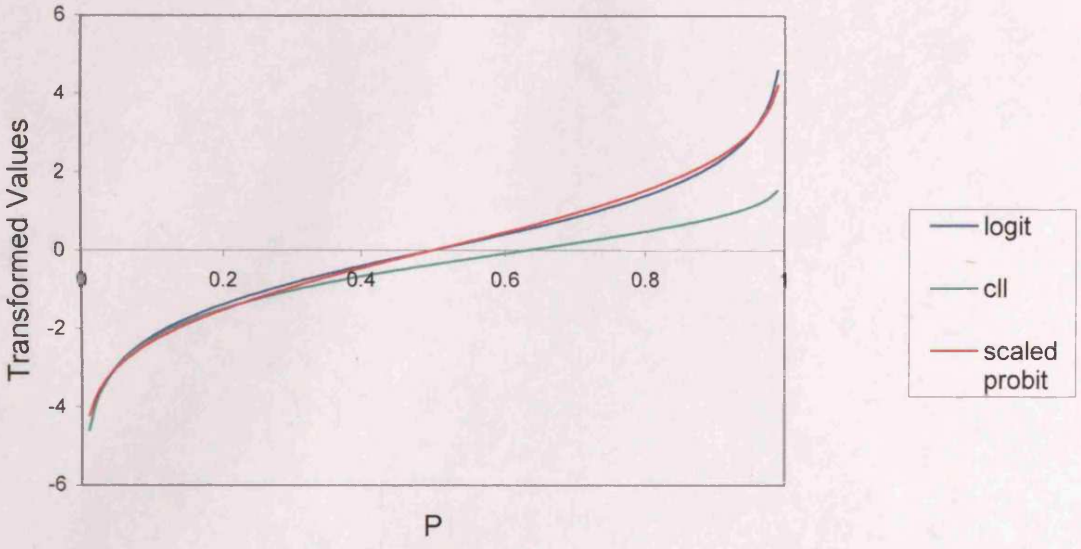


Figure 2.2: Graph to show logistic, scaled probit and complementary log-log mean values across the range [0,1]

2.4.5 Discussion

Despite the similarities between the transformations, and that each of the transformations can be used for certain natural occurring problems, the most commonly used in practice is that of the logistic transformation. The logistic transformation has the advantage that the model parameter associated with the explanatory variable can easily be interpreted as we shall see in section 2.5 and is undoubtedly the most widely used model in this context. Therefore, for the purpose of this study we will restrict our discussion to the logistic transformation although the other two transformations do have their uses and will be described where appropriate.

2.5 The Linear Logistic Model

A single explanatory variable is related to the binary outcome variable Y using the logistic transformation, through the probability p where p is $p(Y=1|X=x)$, such that

$$\log\left(\frac{p}{1-p}\right)=\alpha+\beta x \tag{2.7}$$

and on re-arrangement

$$p=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \tag{2.8}$$

What makes this model so appealing is that the regression coefficient of the explanatory variable has a direct interpretable meaning. That is, the term e^{β} is the Odds Ratio corresponding to a unit change in the explanatory variable. Let

$$p_0 = P(Y_i=1|x_i=x)$$

$$p_1 = P(Y_i=1|x_i=x+1)$$

Then

$$\log\left(\frac{p_1}{1-p_1}\right)=\alpha+\beta(x+1)$$

$$\log\left(\frac{p_0}{1-p_0}\right)=\alpha+\beta x$$

And

$$\log\left(\frac{p_1}{1-p_1}\Big/\frac{p_0}{1-p_0}\right)=\beta$$

Therefore,

$$\log(\text{odds ratio}) = \beta$$

and

e^{β} = odds ratio

Since the odds ratio is widely used in epidemiology, this makes the model very appealing, though this is not in itself a justification for its use.The following sections will show how the above model can be fitted to a set of data using various different techniques and will briefly consider issues of model checking.

2.6 How to fit the linear logistic model

2.6.1 Maximum Likelihood Estimates

As was shown in section 2.2, simple analytical unbiased estimates can be obtained for the linear regression model parameters by the method of least squares. If the errors are normally distributed then the method of least squares is equivalent to the method of maximum likelihood. For the logistic regression model, the method of maximum likelihood is generally used to obtain estimates of the model parameters. In the following sections we will see that parameter estimation for the logistic regression model is more complex than for linear regression and in general there is no closed expression for the maximum likelihood estimates.

For the binary data case where we have n pairs of values (x_i, y_i) with $p_i = P(Y_i = 1 | X = x_i)$ the likelihood function is given by

$$L = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i}$$

For convenience in maximisation we use the log-likelihood

$$\ell = LogL = \sum_{i=1}^{n} Y_i \log p_i + \sum_{i=1}^{n} (1 - Y_i) \log(1 - p_i)$$

On substitution of p_i when a logistic regression model is assumed,

$$\ell = \sum_{i=1}^{n} Y_i \log\left[\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right] + \sum_{i=1}^{n}(1-Y_i)\log\left[1-\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right]$$

$$= \sum_{i=1}^{n} Y_i \log\left[\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right] + \sum_{i=1}^{n}(1-Y_i)\log\left[\frac{1}{1+e^{\alpha+\beta x_i}}\right]$$

$$= \sum_{i=1}^{n} Y_i(\alpha+\beta x_i) + \sum_{i=1}^{n}\log\left[\frac{1}{1+e^{\alpha+\beta x_i}}\right]$$

On differentiating with respect to the individual parameters

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = 0 \qquad (2.9)$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} Y_i x_i - \sum_{i=1}^{n}\frac{x_i e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = 0 \qquad (2.10)$$

It can clearly be seen that these are a set of non-linear equations and cannot be solved easily. Therefore, there are no simple analytical estimates for the model parameters, except for the case where X is binary, unlike with the linear regression model. Hence a numerical method must be used to obtain estimates of the model parameters.

2.6.2 Newton-Raphson's approximation to a root of an equation

The standard method to obtain estimates for the logistic regression model parameters is the iterative Newton-Raphson process. The Newton-Raphson method is an iterative numerical method used to solve an equation in the form $f(x)=0$. Given a starting value x_0, a sequence of approximations is obtained using

$$x_r = x_{r-1} - \frac{f(x_{r-1})}{f'(x_{r-1})} \qquad r = 1,2,... \qquad (2.11)$$

This continues until a sufficient amount of accuracy is achieved, and hence the root of the equation has been found. However, as convergence is not guaranteed then the starting value x_0 should be chosen to be as close to the actual root as possible.

This generalises to a set of equations involving several variables say $f(x)=0$; the resulting iterative process is the Newton-Raphson approximation to a root of an equation. Then

$$x_r = x_{r-1} - J_{r-1}^{-1} f(x_{r-1})$$

where J is the matrix of the first derivatives evaluated at x_{r-1}.

For the case of the simple linear logistic model, two roots from two simultaneous equations must be solved namely (2.9) and (2.10). Therefore, to utilise the Newton-Raphson method their vector forms must replace the elements of equation (2.11).

Let α_r and β_r be the respective starting values for equations (2.9) and (2.10). The two functions are then

$$f_1(\alpha_r, \beta_r) = \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\alpha_r + \beta_r x_i)}{1 + \exp(\alpha_r + \beta_r x_i)} \qquad (2.12)$$

$$f_2(\alpha_r, \beta_r) = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i \exp(\alpha_r + \beta_r x_i)}{1 + \exp(\alpha_r + \beta_r x_i)} \qquad (2.13)$$

respectively. Both functions are then differentiated again with respect to each of the model parameters to be estimated.

$$\frac{\partial f_1}{\partial \alpha} = \frac{\partial^2 \ell}{\partial \alpha^2} = \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} \qquad (2.14)$$

$$\frac{\partial f_1}{\partial \beta} = \frac{\partial^2 \ell}{\partial \alpha \partial \beta} = \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1 + \exp(\alpha + \beta x_i))^2} \qquad (2.15)$$

$$\frac{\partial f_2}{\partial \alpha} = \frac{\partial^2 \ell}{\partial \alpha \partial \beta} = \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{(1+\exp(\alpha + \beta x_i))^2} \tag{2.16}$$

$$\frac{\partial f_2}{\partial \beta} = \frac{\partial^2 \ell}{\partial \beta^2} = \sum_{i=1}^{n} \frac{x_i^2 \exp(\alpha + \beta x_i)}{(1+\exp(\alpha + \beta x_i))^2} \tag{2.17}$$

If the new estimates are said to be α_{r+1} and β_{r+1} respectively, then Newton's formula is

$$\begin{pmatrix} \alpha_r \\ \beta_r \end{pmatrix} = \begin{pmatrix} \alpha_{r-1} \\ \beta_{r-1} \end{pmatrix} - \begin{pmatrix} \frac{\partial f_1(\alpha_{r-1}, \beta_{r-1})}{\partial \alpha} & \frac{\partial f_1(\alpha_{r-1}, \beta_{r-1})}{\partial \beta} \\ \frac{\partial f_2(\alpha_{r-1}, \beta_{r-1})}{\partial \alpha} & \frac{\partial f_2(\alpha_{r-1}, \beta_{r-1})}{\partial \beta} \end{pmatrix}^{-1} \begin{pmatrix} f_1(\alpha_{r-1}, \beta_{r-1}) \\ f_2(\alpha_{r-1}, \beta_{r-1}) \end{pmatrix}$$

This iterative process is continued until a stopping criterion is satisfied; for example this might involve the magnitude of changes between $(\alpha_{r-1}, \beta_{r-1})$ and (α_r, β_r) or the magnitude of the derivatives in (2.12) and (2.13).

2.6.3 The Quasi-Newton Method

The Newton Raphson method found the maximum likelihood estimates by solving the likelihood equations (2.9) and (2.10). This required the first and second derivatives of the log likelihood namely (2.12 –2.13) and (2.14-2.17) respectively. Another approach is to maximise the likelihood, or the log likelihood, directly and there is a whole family of methods available for use which do not require these derivatives. One such method is the Quasi-Newton method.

The basic idea is that the method approximates the given function at each iteration by a quadratic function and then uses the turning point of this quadratic as the next point, this process is repeated until a maximum has been found. The resulting formula is

$$\boldsymbol{x}_r = \boldsymbol{x}_{r-1} - \lambda_{r-1} * \boldsymbol{G}_{r-1}^{-1} \boldsymbol{g}_{r-1}$$

where

x_r is the current point

λ_{r-1} * is determined by a linear search from x_{r-1} in the direction $-G_{r-1}^{-1}g_{r-1}$

G_{r-1}^{-1} is a positive definite symmetric matrix, which is updated in each iteration

g_{r-1} the gradient vector of $f(x)$

Both of the above are evaluated at $x = x_{r-1}$

As with the usual Newton-Raphson method, the iterative process needs starting values for the unknown model parameters.

Convergence is not guaranteed with the Newton-Raphson method, as it is very much dependent on the original starting values of the model parameters. The Quasi-Newton method is always guaranteed to move to points where the function is decreasing in value and so is used in this work. The method that we have used is a NAG routine from the Fortran Library Mark 17, E04JAF. This finds the minimum of a function of several variables using a Quasi-Newton algorithm that has simple bounds on the unknown model parameters and the user has to specify the starting values. To find the maximum of a function of several variables is equivalent to minimising $-f(x)$.

2.7 Properties of estimates of the logistic regression coefficients

For the method of least squares, the Gauss-Markov theorem provides justification that the estimates are unbiased. For the method of maximum likelihood it is not generally true that the estimates are unbiased. However, there are a number of properties associated with maximum likelihood estimators that provide partial justification for the method. These include consistency and large sample efficiency, that is as the

sample size increases the estimates become unbiased minimum variance estimators. To investigate possible bias in small samples a simulation exercise was performed.

An simulation study was conducted to test how well the model parameters were estimated when there is no measurement error in X.

Sample sizes of 100, 500 and 1000 were chosen and the true risk factor X, was taken to follow a Normal distribution with parameters $X \sim N(30,10^2)$. The model parameters were set to $\alpha_t = -3$ and $\beta_t = 0.1$ respectively. The dependent variable Y was then generated according to

$$P(Y=1|X) = \frac{\exp(\alpha_t + \beta_t X)}{1+\exp(\alpha + \beta_t X)}$$

The simulation was run 1000 times. Table 2.1 contains the results that were obtained.

N	$\hat{\alpha}_t$	SE($\hat{\alpha}_t$)	$\hat{\beta}_t$	SE($\hat{\beta}_t$)
100	-3.1125	0.8549	0.1031	0.0275
500	-3.0276	0.3698	0.1007	0.0119
1000	-3.0124	0.2605	0.1004	0.00836

Table 2.1: Tabular form of simulation results

From these results it can be seen that there is a slight bias in the estimation of the model parameters when the sample size is small. This decreases as the sample size increases. Therefore the observations in Table 2.1 confirm that as the sample size increases so the bias in estimating the model parameters reduces.

2.8 Multivariate Case

Within this study we are primarily concerned with the case of a single covariate related to a binary outcome. Thus we have shown the results for this special case. However, these results do generalise to the multivariate case where m explanatory variables $\boldsymbol{X}$ are related to an outcome variable Y through the probability p where $p=p(Y=1 \mid \boldsymbol{X})$, such that

$$\log\left(\frac{p}{1-p}\right) = \alpha + \boldsymbol{\beta_m X_m}$$

where $\boldsymbol{\beta_m}$ is a row vector of the associated model parameters for the m explanatory variables $\boldsymbol{X}$. To estimate the model parameters, the same methods of Newton-Raphson and Quasi-Newton can be used to maximise the log likelihood, finding numerical solutions to the set of $m+1$ simultaneous equations.

2.9 Inference and Goodness of fit

Once a model has been fit to the observed data, it leads to the questions of how well the model actually explains the data and how confident we are in the estimates of the model parameters. Thus estimates of the standard errors associated with the estimates of the model parameters are required.

2.9.1 Confidence intervals for the single covariate coefficient, namely $\hat{\beta}_t$

To be able to construct a confidence interval for the estimate of the model parameter β_t we need an expression for its variance. A proof is given in Appendix A.

$$Var\left(\hat{\beta}_t\right) = \frac{\sum p_i(1-p_i)}{\sum x_i^2 p_i(1-p_i)\sum p_i(1-p_i) - \left(\sum x_i p_i(1-p_i)\right)^2} \quad (2.19)$$

By using the estimates for α and β to determine the estimates for p_i, and assuming approximate normality of the estimators, a confidence interval for the logistic regression coefficient can be constructed.

2.9.2 Assessing the fit of a logistic regression model: Robust Locally Weighted Regression and Smoothing Scatterplots

In section 2.4 we introduced the three transformations that are commonly used to transform the probability. These transformations had two requirements namely, to transform the probability to the real line and create a linear relationship between these transformed values and the explanatory variables. However, these transformations were chosen for the first requirement and do not guarantee that a linear relationship would hold. Therefore, before proceeding with a method such as the logistic regression model, it must be determined whether a linear relationship between the transformed probability and the explanatory variable is a valid model.

There are a number of ways of seeing how well a linear regression model fits the data involved. One way is to use a graphical representation. This is usually in the form of a scatter plot. The data is plotted on a graph with the regression line drawn through the points, judgement is then made on how well this line fits the data. Alternatively, residuals can be plotted. When the dependent variable is binary, such as in our case, the scatter plot can be drawn but the display is uninformative. Figure 2.3 is a plot of this type of data.

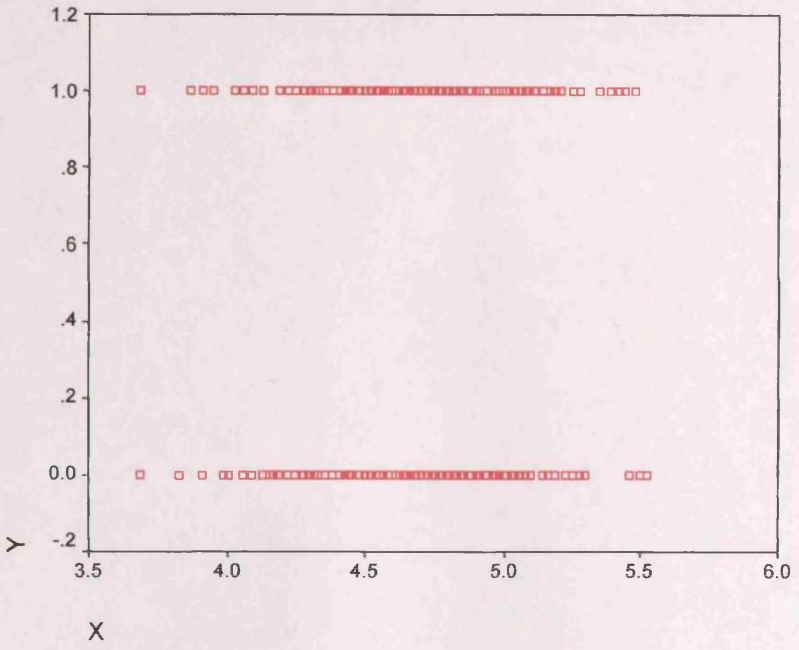


Figure 2.3: Scatter plot of example data

As can be seen from this plot, it visually provides very little information on the relationship between the two variables. As a result, we would like to find a way of representing the data to see what type of model would e appropriate.

One way of fitting a line to this type of data so as to explore the relationship between the variables is to split the range of X values into groups and then calculate in each group the proportion of those values that correspond to a Y value of 1. These proportion values then become points on the scatter plot. Depending on the size of n, this process can be refined further by increasing the number of groups. However, if there are small numbers of X values within each group, or there are some groups that do not have any X values that correspond to a Y value of 1, then this method will not give a smooth curve and therefore, would not show any trend in the data.

For example, if we return to our example from Chapter 1, fasting plasma glucose is used as an indicator for the presence of diabetic retinopathy in type II diabetics. If we plot this relationship on a scatter diagram then the following is observed:

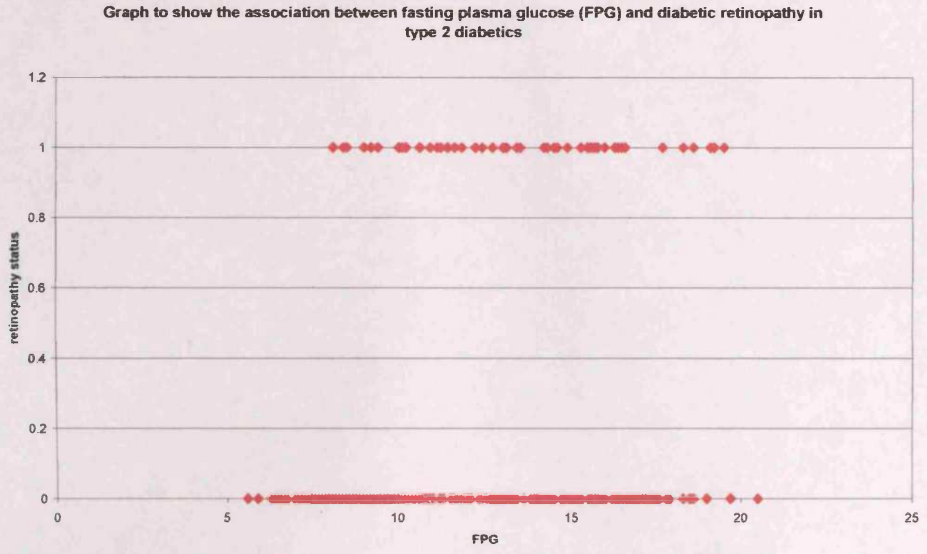


Figure 2.4: Graph to show the association between FPG and diabetic retinopathy in type 2 diabetics

From Figure 2.4 it is not apparent that a linear model would fit this particular set of data. However, if we group the fasting glucose levels into a number of classes and consider the proportion of retinopathy positive status against increasing levels of the grouped fasting plasma glucose values in Figure 2.5, then an upwards linear trend is observed.

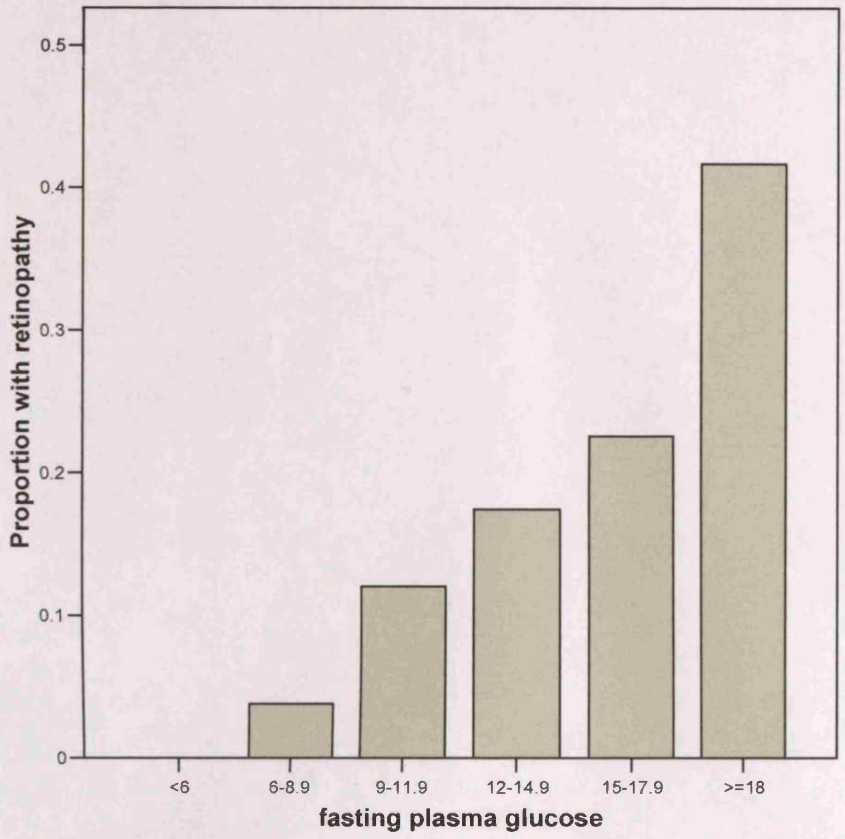


Figure 2.5: FPG against proportion of retinopathy status

A more reliable process that will show the pattern of the data and suggest a plausible model is known as LOWESS, see Cleveland (1979).

For the linear regression model the predicted Y values, $\hat{Y}_i$, were fitted using the model

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$$

LOWESS approaches the process of estimation slightly differently. For each point x_i in turn, a polynomial of degree d is fitted to x_i and its local influential neighbouring points using weighted least squares. This fitted polynomial is then used to predict a value corresponding to the x_i value. This process is repeated for each x_i. These values are then plotted to see if there is any trend in the data.

The author recommends fitting a linear model using the following formulation

$$\hat{Y}_i = \beta_0(x_i) + \beta_1(x_i)x_i$$

where the regression coefficients $\beta_0(x_i)$ and $\beta_1(x_i)$ are dependent on the x_i value and a weighted function of the surrounding influential neighbouring x_k points. Therefore,

$$\hat{\beta}_0 = \frac{\sum_{k=1}^{n} w_k(x_i)Y_k - \hat{\beta}_1 \sum_{k=1}^{n} w_k(x_i)x_k}{\sum_{k=1}^{n} w_k(x_i)}$$

and

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{n} w_k(x_i)Y_k x_k - \frac{\left(\sum_{k=1}^{n} w_k(x_i)Y_k\right)\left(\sum_{k=1}^{n} w_k(x_i)x_k\right)}{\sum_{k=1}^{n} w_k(x_i)}}{\sum_{k=1}^{n} w_k(x_i)x_k^2 - \frac{\left(\sum_{k=1}^{n} w_k(x_i)x_k\right)^2}{\sum_{k=1}^{n} w_k(x_i)}}$$

To calculate the above regression coefficients the author recommends the 'tricube' weight function

$$W(u) = \left(1 - |u|^3\right)^3, \text{ for } |u| < 1$$

$$= 0, \qquad \text{for } |u| \geq 1$$

such that

$$w_k(x_i) = W\left(\frac{x_k - x_i}{h_i}\right)$$

Here x_i is the current explanatory variable value under consideration, x_k is a neighbouring point and h_i is the distance between x_i and the furthest neighbouring influential point, so that the argument of W is less than 1, with a consequent non-zero weight, only for values of x_k sufficiently close to x_i. The value of h_i controls the degree of smoothing, a small value smoothes less than a large value.

Once all the fitted points have been calculated, they can either be plotted or the process can be turned into an iterative one to smooth the points further. This procedure is now standard in a number of statistical packages and can be used to see whether a logistic model will fit the data.

The method described here is not guaranteed to give non-negative predicted values and so may not be ideal when applied to binary data, but is sufficient in most cases to establish the general relationship. Other methods, involving local logistic regression and kernel smoothing methods are described in Bowman & Azzalini (1997).

If we return to the example of relating fasting plasma glucose to retinopathy status then the LOWESS smoothing plot is shown in Figure 2.6:

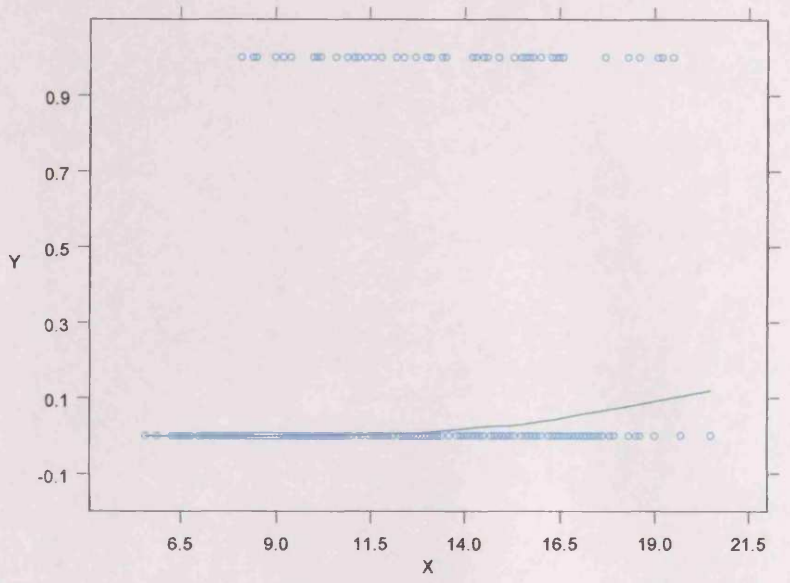


Figure 2.6: Graph to show the relationship between FPG and retinopathy status through a LOWESS plot

This LOWESS plot shows that there is an increasing relationship between fasting plasma glucose and retinopathy status, suggesting that a logistic regression might be an appropriate model to use when examining the relationship between the two factors. A plot on the logit scale would demonstrate this more effectively but has the disadvantage that the observed data cannot be plotted.

Although in this thesis we are concerned with the effect of measurement error in an explanatory variable. However, this discussion on observing whether there is a linear relationship between the explanatory variable and binary outcome raises many issues about the appropriateness of the logistic and other transforms. Current medical literature rarely has cases of model checking and validation and so these issues should be investigated further.

2.10 Discussion

Throughout this chapter we have discussed a model that can be used to relate a binary outcome variable to an explanatory variable, namely the linear logistic model. We have shown how this model can be fitted and how graphical plots such as the LOWESS plot can show how well the logistic model fits the data. We have also shown the linear regression case for relating a continuous variable to an explanatory variable. This information has been provided as background to the measurement error problem, models for which are now discussed in Chapter 3.

Chapter 3

3 The Errors-in-Variables Problem

3.1 Introduction

In Chapter 1 we discussed a number of examples where the measurement of a variable could be subject to measurement error. In Chapter 2 we introduced the modelling technique of logistic regression that relates a disease status to an explanatory variable. In this chapter we extend the logistic regression model to allow for measurement error in an explanatory variable.

As the examples in Chapter 1 showed there is a variety of situations where the observed variable could be subject to measurement error. This variety of situations showed that the measurement error could take different forms for example, machine error or average values being used in place of true values. What is also apparent from these examples is that the true and observed values can be related in different ways. In this chapter we introduce a number of different measurement error models for

describing the relationship between the true and observed values. We also explain the concept of subsidiary studies, and of different forms they may take, and how they can be used to estimate the measurement error distribution parameters. Further models are introduced concerning the explanatory variable and their associated distribution assumptions.

Much work has been conducted on the linear regression errors-in-variables particular form of the errors-in-variables problem, for example see Hill, Nix & Iles (1999) and Chen & Van Ness (1999). Some of the work from Hill, Nix & Iles (1999) is summarized so that contrasts can be made with the logistic case.

We then introduce the logistic regression errors-in-variables problem and its formulation and explain why in comparison to the linear regression errors-in-variables problem that this case is more complex, as with the ordinary logistic regression model described in chapter 2.

3.2 The Errors-in-Variables Models

3.2.1 Introduction

As we have seen, measurement error can occur in a number of different ways and as a result a number of different models can be formulated to represent different situations. In what follows X will denote the true value and Z the observed values. For example, returning to the examples given in chapter 1, consider the measurement of fasting plasma glucose from a single blood measurement. Then Z denotes the measured glucose concentration while X denotes the true glucose concentration. In another case, a food frequency questionnaire is used to measure true fibre intake. Z, the value obtained from the dietary questionnaire, is a surrogate variable for X, the true fibre

intake. The measurement error model expresses the relationship between X and Z. In the case of the measurement of dietary fibre intake, a surrogate is used to describe the true value, but when measuring drug absorption into the blood stream, a clinician knows the amount administered but not how much of the drug has been absorbed by the body. In these two instances, different measurement error models would be needed to describe these situations.

Within this section we introduce the different types of measurement error model as well as explain the role of validation studies and how they help to estimate the measurement error associated with the risk factor.

3.2.2 The Classical Measurement Error Model

The Classical Measurement Error Model is of the form

$$Z = X + V$$

where V, the error incurred in measuring X, has a mean of 0 and is independent of the true value X. This measurement error model corresponds to the situation where the investigator has aimed to measure X directly but has incurred a random error in the process. Returning to the examples given in chapter 1, this type of measurement error model would be used for the case where a measurement of gestational age is required. The true gestational age is calculated from foetal measurements using a model based on expected values and thus the true age is not measured directly and a surrogate value is used instead. Another example was the measurement of blood pressure which can be subject to both machine error as well as variation over time, so that a single blood pressure reading may not truly indicate an individual's long-term blood pressure but would be this true value plus an associated error involving the two

components, one reflecting within-subject fluctuations and the other reflecting genuine measurement error.

3.2.3 The Berkson Measurement Error Model

The Berkson Measurement Error Model is in the form

$$X = Z + V$$

where it is assumed that Z and V are independent. This model is used in situations where a fixed level can be observed or administered but the actual individual level cannot be observed. For example, if we return to the example of measuring drug absorption in the body, the actual level Z administered is known but the level X absorbed by the patient will depend on the physical attributes of the patient. The clinicians are more interested in the true value that is absorbed by the patient than the fixed amount that was administered which is likely to be related to an outcome. In the case of measuring deprivation, a number of area based measures are used but it is the individual level of deprivation that is interesting. Therefore, in these cases the Berkson measurement error model would be used to describe the relationship between the true and observed variables. It is important to note that X and V are now dependent; this has considerable implications for analysis.

3.2.4 Measurement Error Model by Reeves, Cox, Darby & Whitley (1998)

The two models considered appear rather different but in fact can be incorporated into a single formulation. This measurement error model was defined by Reeves, Cox, Darby & Whitley (1998) using the generalisation

$$Z = \tilde{X} + V \qquad X = \tilde{X} + U \tag{3.1}$$

where $(\tilde{X}, V, U)$ are independent variables with corresponding means $(\mu_x, 0, 0)$ and variances $(\sigma_x^2, \sigma_v^2, \sigma_u^2)$. The classical error model case corresponds to the case where

$\sigma_u^2 = 0$, so that $X = \widetilde{X}$ and $Z = X + V$. The Berkson error model corresponds to $\sigma_v^2 = 0$, leading to the measurement error model $X = Z + U$. The usual assumption of $\left(\widetilde{X}, V, U\right)$ being independently normally distributed and therefore mutually uncorrelated, allows for the usual general inferences such as confidence intervals of the model parameters to be drawn up. However, these confidence intervals will not take into account the error involved in estimating the parameters in the validation study. We will return to the subject of validation studies in section 3.2.6

As this formulation can incorporate both the Classical as well as the Berkson Models it can deal with all the medical examples previously given where $E(X) = E(Z)$.

3.2.5 Other Measurement Error Models

If we return to the example of measuring dietary intake, a food frequency questionnaire or a more detailed study of weighing food are used as surrogates for actual fibre intake, but both can be subject to error such as under-reporting. In this case a model in the form

$$X = \lambda Z + V$$

would be required. This form of model may be more appropriate in describing the relationship than assuming that there is no scaling factor.

In this study, most of the work shall consider the Classical, Berkson and Reeves form of measurement error model, this particular measurement error model will be discussed further in section 4.2.1.

3.2.6 Subsidiary and Validation Studies

We will use the term the 'main study' to refer to the study which provides the information on all the variables on the principal set of subjects and which will be used to make inferences about associations between the disease in question and its associated risk factors. The vast majority of the methods we shall consider require making assumptions about the measurement error model and this may involve another study, called a subsidiary or validation study. This is used to estimate the parameters involved in the measurement error model. This subsidiary study may be internal to the main study or external, that is a study that has been conducted independently of the main one. The following then is an explanation of the different types of subsidiary study and their associated assumptions.

The main idea of a subsidiary study is that more detailed measurements can be taken on a sub-sample of individuals. In that way an assessment can be made of the distribution of the error involved in measuring the risk factor under consideration. This can be done in two ways, either by measuring what is known as a 'gold standard' as well as the surrogate risk factor, or through repeated measurements of the surrogate risk factor. The first study is known as a validation study, and it is assumed that there is some procedure, or perhaps a measuring instrument, that can measure what the 'true' value should be; this is known as the 'gold standard'. Therefore, on either a sub sample of the main study, or a completely separate study, both the 'gold standard' and the surrogate exposure variable are measured and compared to derive a model for the measurement error. It can also be concluded whether $E(X) = E(Z)$, as is assumed in both the Classical and Berkson measurement error models. The advantage of this method is that sometimes the 'gold standard' is too expensive to carry out on all the subjects of the main study; however, by just measuring a sub-sample, the cost of the

study is reduced. An indication of how well the surrogate measuring device works at measuring the risk factor is also given.

For example, if we consider the measurement of saturated fat in an individual's diet, this risk factor is measured through the use of a food questionnaire; however it can be subject to measurement error. Therefore, along with the main study a separate validation study can also be conducted. This separate study can involve a more detailed questionnaire, going into greater depth than the questionnaire given to the subjects in the main study. For example, a more detailed questionnaire could involve the weighing of individual portions and noting down each time something is ingested. This would provide far more detail than a general recall questionnaire as well as less biased results, since individuals tend to under-report food intake in general recall questionnaires. This separate study is then used to verify information gathered from the main study.

However, there is not always a 'gold standard' and as a result a reproducibility study is required. Again, an internal or external study can be used, in either case a number of measurements are taken on each individual using the same imperfect device. It is then assumed that by comparing these values for each individual, we can again obtain an estimate of the standard deviation of the measurement error involved.

When assessing the relationship between FPG and retinopathy in type II diabetic subjects, as discussed in chapter 1, the study was designed to look at the variability of measuring FPG. A reproducibility study involving two repeated measurements on the

same subjects could be carried out to estimate the measurement error standard deviation.

To use these subsidiary studies in the assessment of measurement error a number of assumptions have to be made. For an internal subsidiary study it is assumed that the sub-sample used to estimate the measurement error is a representative sample of the population. Otherwise, this selection could introduce a bias into the results, and therefore cloud any inferences that could be made from the study. For an external subsidiary study further checks and assumptions have to be made. Again, the sample must be based on a population similar to that of the main study. However, it also has to be assumed that the measurement error model is the same for both studies. By this we mean that the same model holds with the same parameter values for both the main and validation study. If this assumption cannot be met then the external study should not be used as a subsidiary study for the main study. The result could be an introduction of unnecessary bias into the estimation procedure. Therefore, we issue caution in the use of external studies and advise that an internal subsidiary study should be designed into the original study specifications. Work has been conducted on the effects of measurement error and the requirements for subsidiary studies, see White, Frost, & Tokunaga (2000) and Thurigen, Spiegelman, Blettner, Heuer & Brenner (2000). This thesis will not discuss these issues and requirements but make the assumptions regarding validation studies as stated above.

3.2.7 Summary

In summary, the logistic model linearly relates the probability $p = p(Y = 1|X = x)$ to the true explanatory variable X through the logit transform in the form

$$\text{logit}(p) = \alpha_t + \beta_t x$$

where t denotes the true model parameter values. When the true explanatory variable is known to be subject to error then a model linking the outcome to the observed data is

$$\text{logit}(p) = \alpha_s + \beta_s z$$

where $p = p(Y = 1 | Z = z)$. X and Z are related through a measurement error model such as the Classical or Berkson measurement error models. In the Classical model

$Z = X + V$ where X and V are independent

and in the Berkson model

$X = Z + V$ where Z and V are independent

and V denotes the measurement error term

Standard methods give estimates of α_s and β_s, but we require estimates of α_t and β_t. How are these related?

3.3 Modelling Procedures in the Errors-in-Variables Problem

To completely specify the models, assumptions need to be made about the 'true' unobserved X values. There are two types of modelling procedures which Carroll, Ruppert & Stefanski (1995) have termed 'Functional Modelling' and 'Structural Modelling'. Both procedures assume that the measurement errors are independent and normally distributed as well as assuming that the X_i are independent and identically distributed. They differ, however, in their assumptions about the unobserved X values.

These are defined by:

Functional Model: the X_i are unknown constants

Structural Model: The X_i are a random sample from a distribution with mean μ_x and variance σ_x^2

There is also a third procedure which we shall also introduce that is,

Ultrastructural Model: X_i are independent random variables with means μ_{x_i} and common variance σ_x^2

In the structural modelling approach, it is assumed that the X_i follow a specific parametric distribution which leads to a likelihood based approach. This means that a full specification of the data distributions can be made which can then lead to a set of efficient estimators. For example, blood pressure measurements are believed to follow a Normal Distribution and so in the case of relating blood pressure to Coronary Heart Disease a parametric approach can be applied.

If there is concern over the distribution of the X_i, then a full parametric model cannot be specified and the semi parametric method of functional modelling should be employed. In this approach the X_i can be regarded as either fixed or random; for the latter case no distributional assumptions are required, which is in contrast to the structural approach.

These two different types of modeling procedures allow different modeling assumptions to be made about the X_i. This provides flexibility in the different

designs of medical experiments and provides the ability to model the true situation. Both types of model are a part of the errors-in-variables problem and will be discussed further in Chapter 4 when considering the different types of methods that are available to correct for measurement error.

3.4 The Linear Regression Errors-in-Variables Problem

3.4.1 Introduction

We are now going to consider the linear regression errors-in-variables problem briefly before describing the logistic case, in order to make comparisons between the two techniques when considering methods and identifiability of the models. In the following sections we review some of the key results.

For convenience, we have used the standard notation of the linear regression errors-in-variables that is used throughout the literature, which was devised by Kendall & Stuart (1979).

3.4.2 The Model

The standard linear regression model for relating two variables is

$$Y = mX + c + \varepsilon \qquad (3.2)$$

where the term ε represents the measurement error, or other sources of random variation, in Y. The error-in-variables problem arises when X is known to be subject to measurement error. As a result the linear model is subject to a different formulation. Both the dependent and independent variables Y and X are assumed to be unobservable and hence we observe the two variables ξ and η instead

$$\xi_i = X_i + \delta_i$$

$$\eta_i = Y_i + \varepsilon_i \qquad i = 1,...n \qquad (3.3)$$

where δ_i is the error or random variation, incurred in measuring X_i and ε_i is the result of measurement error or random variation in Y_i. It is also assumed that

$$E(\delta_i) = E(\varepsilon_i) = 0 \qquad Var(\delta_i) = \sigma_\delta^2 \qquad Var(\varepsilon_i) = \sigma_\varepsilon^2 \qquad (3.4)$$

$$Cov(\delta_i, \delta_j) = Cov(\varepsilon_i, \varepsilon_j) = 0 \qquad \text{for all } i \neq j$$

$$Cov(\delta_i, \varepsilon_j) = 0 \qquad \text{for all, } ij$$

Therefore, the linear regression errors-in-variables model is

$$\eta_i = m\xi_i + c + (\varepsilon_i - m\delta_i) \qquad (3.5)$$

The aim is to estimate the model parameters m and c, describing the relationship between X and Y, even though neither is observed. In this model, the model parameters are not identifiable. This means that from a single study data set the model parameters m, c, σ_δ^2 and σ_ε^2 can not all be estimated without further assumptions concerning the error variances.

For the purpose of this study we shall only consider the case of the structural model which has direct applicability to the logistic regression model that we consider. For further specifics we direct the reader to Chen & Van Ness (1999).

From this expanded formulation there are three error variances, namely σ^2, σ_δ^2 and σ_ε^2.

From (3.3) and (3.4)

$$E(\xi_i) = E(X_i) = \mu$$

$$E(\eta_i) = E(mX_i + c) = m\mu + c$$

$$Var(\xi_i) = Var(X_i + \delta_i) = \sigma_x^2 + \sigma_\delta^2$$

$$Var(\eta_i) = Var(Y_i + \varepsilon_i) = Var(mX_i + c) + \sigma_\varepsilon^2 = m^2\sigma_x^2 + \sigma_\varepsilon^2$$

$$Cov(\xi_i, \eta_i) = Cov(X_i + \delta_i, mX_i + c + \varepsilon_i) = m\sigma_x^2$$

$$Cov(X_i, \delta_i) = Cov(Y_i, \varepsilon_i) = 0$$

There are six parameters describing the model namely, μ, c, m, σ_x^2, σ_δ^2 and σ_ε^2. As was described previously, the linear regression errors-in-variables model is not identifiable and therefore, to construct estimates for the model parameters, a further assumption concerning the error variances is required. There are four main cases. These are

Case 1 - Both error variances σ_δ^2 and σ_ε^2 are known

Case 2 - σ_δ^2 is known

Case 3 - σ_ε^2 is known

Case 4 - The ratio $\lambda = \dfrac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ is known

To estimate the model parameters, the method of maximum likelihood is employed. It is assumed that the joint distribution of ξ_i and η_i is bivariate Gaussian that is,

$$\begin{bmatrix} \xi_i \\ \eta_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ m\mu + c \end{bmatrix}, \begin{bmatrix} \sigma_x^2 + \sigma_\delta^2 & m\sigma_x^2 \\ m\sigma_x^2 & m^2\sigma_x^2 + \sigma_\varepsilon^2 \end{bmatrix} \right)$$

and so the likelihood function for a sample of size n is

$$L = \prod \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left\{ -\frac{1}{2}(\xi_i - \mu, \eta_i - (m\mu + c))\Sigma^{-1} \begin{pmatrix} \xi_i - \mu \\ \eta_i - (m\mu + c) \end{pmatrix} \right\}$$

where the variance covariance matrix Σ is

$$\Sigma = \begin{pmatrix} \sigma_x^2 + \sigma_\delta^2 & m\sigma_x^2 \\ m\sigma_x^2 & m^2\sigma_x^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

The log likelihood is therefore

$$l = -n\log(2\pi) - \frac{n}{2}\log\left(\sigma_\delta^2\sigma_\varepsilon^2 + m^2\sigma_x^2\sigma_\delta^2 + \sigma_x^2\sigma_\varepsilon^2\right) - \frac{n}{2}\left(\frac{S_\xi^2\left(m^2\sigma_x^2 + \sigma_\varepsilon^2\right) - 2S_{\xi\eta}m\sigma_x^2 + S_\eta^2\left(\sigma_x^2 + \sigma_\delta^2\right)}{\sigma_\delta^2\sigma_\varepsilon^2 + m^2\sigma_x^2\sigma_\delta^2 + \sigma_x^2\sigma_\varepsilon^2}\right)$$

$$-\frac{n}{2}\left(\frac{\left(\bar{\xi} - \mu\right)^2\left(m^2\sigma_x^2 + \sigma_\varepsilon^2\right) - 2\left(\bar{\xi} - \mu\right)\left(\bar{\eta} - c - m\mu\right)m\sigma_x^2 + \left(\bar{\eta} - c - m\mu\right)^2\left(\sigma_x^2 + \sigma_\delta^2\right)}{\sigma_\delta^2\sigma_\varepsilon^2 + m^2\sigma_x^2\sigma_\delta^2 + \sigma_x^2\sigma_\varepsilon^2}\right)$$

where

$$S_\xi^2 = \frac{1}{n}\sum\left(\xi_i - \bar{\xi}\right)^2$$

$$S_{\xi\eta} = \frac{1}{n}\sum\left(\xi_i - \bar{\xi}\right)\left(\eta_i - \bar{\eta}\right)$$

$$S_\eta^2 = \frac{1}{n}\sum\left(\eta_i - \bar{\eta}\right)^2$$

By differentiating the above log likelihood, estimates for the model parameters can be obtained.

We shall now give the estimates for the model parameters for each of the specified cases in turn , see Hill, Nix & Iles (1999) for further details.

Case 1: σ_δ^2 and σ_ε^2 known

$$\hat{m} = \frac{\left(S_\eta^2 - \lambda S_\xi^2\right) + \sqrt{\left(S_\eta^2 - \lambda S_\xi^2\right)^2 + 4\lambda S_{\xi\eta}^2}}{2S_{\xi\eta}}$$

$$\hat{\sigma}_x^2 = \frac{\lambda^2 S_\xi^2 + 2\lambda m S_{\xi\eta} + m^2 S_\eta^2 - \lambda\left(m^2\sigma_\delta^2 + \sigma_\varepsilon^2\right)}{\left(\lambda + m^2\right)^2}$$

where $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$. These results apply provided

$$S_\eta^2 > \sigma_\varepsilon^2 \qquad \text{or} \qquad S_\xi^2 > \sigma_\delta^2 \qquad \text{or} \qquad S_{\xi\eta}^2 > \left(S_\eta^2 - \sigma_\varepsilon^2\right)\left(S_\xi^2 - \sigma_\delta^2\right)$$

otherwise $\sigma_x^2 = 0$.

Case 2: σ_δ^2 known

$$\hat{m} = \frac{S_{\xi\eta}}{\left(S_\xi^2 - \sigma_\delta^2\right)}$$

$$\hat{\sigma}_x^2 = S_\xi^2 - \sigma_\delta^2$$

$$\hat{\sigma}_\varepsilon^2 = \frac{S_\xi^2 S_\eta^2 - \sigma_\delta^2 S_\eta^2 - S_{\xi\eta}^2}{\left(S_\xi^2 - \sigma_\delta^2\right)}$$

These results are dependent upon

$$S_\xi^2 > \sigma_\delta^2 \qquad \text{and} \qquad S_\eta^2 > \frac{S_{\xi\eta}^2}{\left(S_\xi^2 - \sigma_\delta^2\right)}$$

If the first restriction is broken then set $\sigma_x^2 = 0$ and the model collapses to the functional model, which is explained in more detail in the paper by Hill, Iles & Nix (1999). If the second restriction is broken then set $\sigma_\varepsilon^2 = 0$. This reduces the model to the simple model of X on Y. This is the case where the measurement error distribution associated with the risk factor X is known and is similar to the logistic regression errors-in-variables formulation that is studied within this thesis. It is also worth noting that if the measurement error in X was ignored then the gradient would be estimated as

$$\hat{m}_{ols} = \frac{S_{\xi\eta}}{S_{\xi^2}}$$

Therefore, $|\hat{m}_{ols}| < |\hat{m}|$ unless $\sigma_\delta^2 = 0$.

Case 3: σ_ε^2 known

$$\hat{m} = \frac{S_\eta^2 - \sigma_\varepsilon^2}{S_{\xi\eta}}$$

$$\sigma_x^2 = \frac{S_{\xi\eta}^2}{\left(S_\eta^2 - \sigma_\varepsilon^2\right)}$$

$$\sigma_\delta^2 = \frac{S_\xi^2 S_\eta^2 - \sigma_\varepsilon^2 S_\xi^2 - S_{\xi\eta}^2}{\left(S_\eta^2 - \sigma_\varepsilon^2\right)}$$

These results are dependent upon

$$S_\eta^2 > \sigma_\varepsilon^2 \qquad \text{and} \qquad S_\xi^2 > \frac{S_{\xi\eta}^2}{\left(S_\eta^2 - \sigma_\varepsilon^2\right)}$$

As for case 2, if the first restriction is broken then σ_x^2 must be set to zero and the model collapses to the functional model. If the second restriction is broken then σ_δ^2 must be set to zero and the model reduces to the Y on X least squares regression solution.

Case 4: $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ known

Eliminating σ_ε^2 from the log-likelihood by writing it in terms of λ and σ_δ^2, the following solution can be obtained.

$$\hat{m} = \frac{\left(S_\eta^2 - \lambda S_\xi^2\right) + \sqrt{\left(S_\eta^2 - \lambda S_\xi^2\right)^2 + 4\lambda S_{\xi\eta}^2}}{2S_{\xi\eta}}$$

$$\hat{\sigma}_x^2 = \frac{S_{\xi\eta}\left(m^2 S_\eta^2 + \lambda^2 S_\xi^2 + 2m\lambda S_{\xi\eta}\right)}{\left(m^2 + \lambda\right)\left(mS_\eta^2 + \lambda S_{\xi\eta}\right)}$$

$$\hat{\sigma}_\delta^2 = \frac{\left(mS_\xi^2 - S_{\xi\eta}\right)\left(m^2 S_\eta^2 + \lambda^2 S_\xi^2 + 2m\lambda S_{\xi\eta}\right)}{\left(m^2 + \lambda\right)\left(mS_\eta^2 + \lambda S_{\xi\eta}\right)}$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{\lambda\left(mS_\xi^2 - S_{\xi\eta}\right)\left(m^2 S_\eta^2 + \lambda^2 S_\xi^2 + 2m\lambda S_{\xi\eta}\right)}{\left(m^2 + \lambda\right)\left(mS_\eta^2 + \lambda S_{\xi\eta}\right)}$$

Asymptotic information and variance-covariance matrices have also been ascertained; see Hill, Iles & Nix (1999). These formulae will not be included here.

Therefore, the linear regression errors-in-variables model is identifiable for the four cases described above, in each of which an assumption has been made concerning the error variances. In that case estimation is through simple analytical formulae that can easily be evaluated. Estimates for the variances of the estimators have also been derived, and therefore measures of the precision of the estimates can also be calculated. These estimates of the variances are dependent on, say, σ_δ^2 or σ_ε^2 being known. If these have been estimated from a validation study then this will have an effect on the estimate of the associated standard errors.

3.4.3 Effects of measurement error in the linear regression errors-in-variables problem

To illustrate the effect of measurement error on the estimation of the model parameters, we consider an example of case 1, where both the error variances σ_ε^2 and σ_δ^2 are assumed known.

For the method of case 1, a single data set was generated. The sample size was taken to be 100 and both X and Y were assumed to be normally distributed with parameters $N(30,10^2)$. The model was generated according to $Y = X + \varepsilon$ with $Z = X + V$ and $\sigma_v^2 = 4$ and $\sigma_\varepsilon^2 = 1$. The gradient and intercept were then calculated using the techniques of ordinary least squares and the method described in case 1. Figure 3.1 shows the results of this investigation.

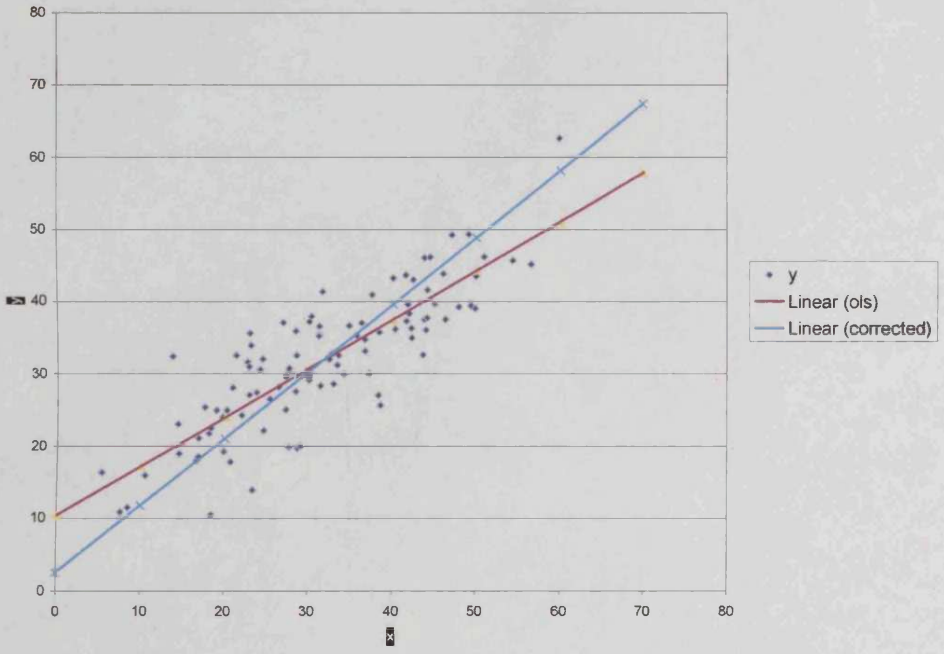


Figure 3.1: Comparison of methods for the linear error-in-variables problem

For this illustrative data set it can be seen by ignoring the measurement error in the X variable and using the standard statistical technique of ordinary least squares, the regression line clearly underestimates the gradient in comparison with the corrected method. For this single data set if both variables are known to be mismeasured then clearly the correction method should be used, otherwise the results produced are seriously attenuated and could result in an underestimation of an effect.

In the above example, we have taken the case where it is assumed that both the error variances are known and found that if there is a large amount of measurement error then this has an effect on estimating the model parameters. If we examine the 2nd case that is, σ_δ^2 is known, then the measurement error associated with the explanatory variable is known. This is analogous to the logistic regression problem.

For the linear regression errors-in-variables problem there are analytical estimates of the model parameters including that of the measurement error variance. A simulation study was conducted for this particular case to see the effect of measurement error on the estimate of the slope as well as seeing how well these new parameter estimates work as the measurement error is increased. Again the method of least squares was utilised to calculate the parameter estimates and the errors-in-variables formulae for case 2 was used to calculate the corrected estimates. For this study, the explanatory variable X was generated from $X \sim N(30,10^2)$ for 500 values, and $\sigma_\varepsilon^2 = 1$, $\alpha = 0$, $\beta = 1$ and σ_δ was varied from 0 to 4. The mean value of 5000 simulations was taken for the parameter β and the associated standard deviation.

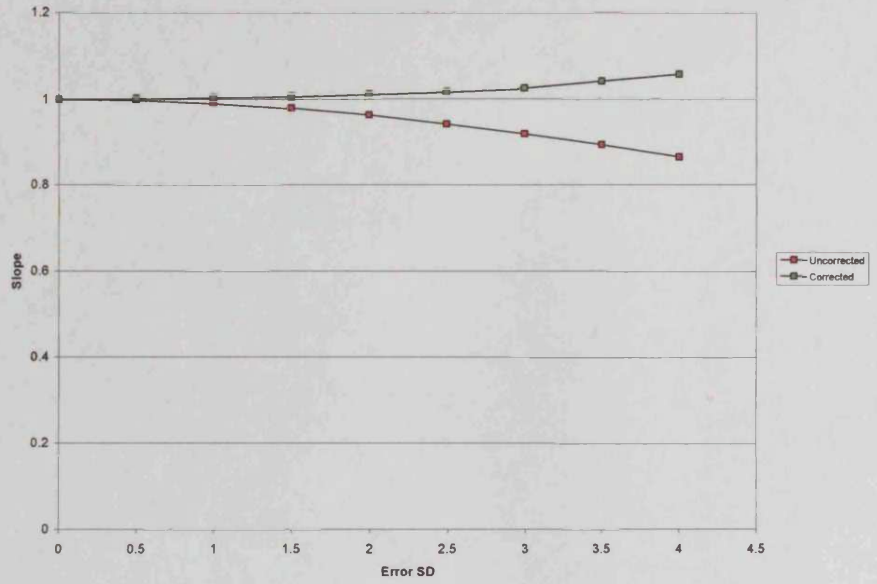


Figure 3.2: Mean estimate for β from 5000 simulations

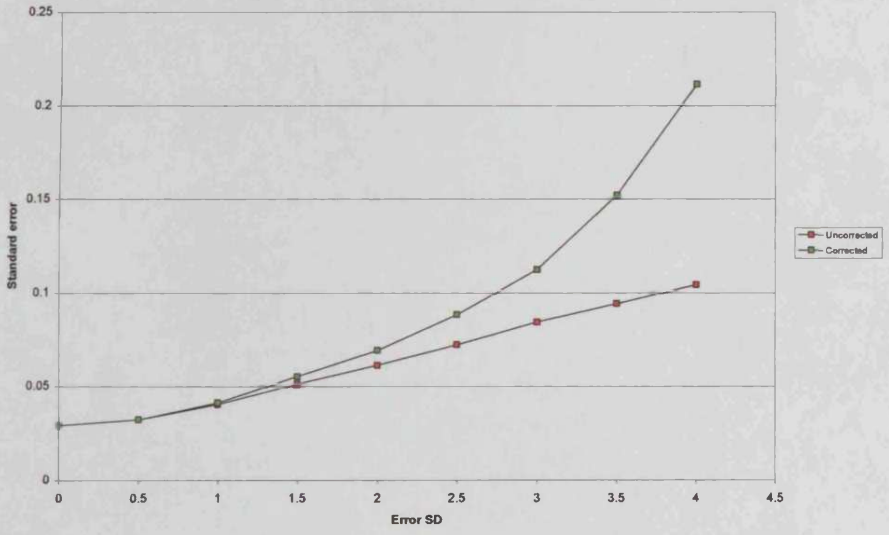


Figure 3.3: Standard Deviation for β from 5000 simulations

Figure 3.2 shows that with no measurement error the method of least squares produced unbiased estimates for the model parameter. As the measurement error is increased a bias is introduced in the estimate of β which increases as the

measurement error is increased. As Figure 3.3 shows, as the measurement error is increased so the precision in estimating β decreases.

The new errors-in-variables estimate for β also produced an unbiased estimate when there was no measurement error. As the measurement error increases a slight positive bias is introduced into the estimate of β. However, this bias is less than that produced by the method of least squares. We shall return to this particular simulation study for the logistic regression model in Chapter 5.

For the linear regression errors-in-variables problem, it can be seen that measurement error in the explanatory variable can cause the strength of association to be underestimated between the explanatory variable and the continuous outcome variable. Thus estimates for the model parameters that correct for this measurement error are required. From these studies it can be seen that the corrected estimates do correct for the measurement error though when the measurement error standard deviation becomes large in relation to the explanatory variable standard deviation then these estimates slightly over-estimate the strength of the relationship.

The work by Hill, Nix & Iles (1999) provided analytical estimates for the linear regression errors-in-variables model parameters depending on certain assumptions concerning the error variances. The model is identifiable when one or more of the error variances is known or at least the ratio between the error variances is known. For the logistic regression errors-in-variables problem, there is only one measurement error variance associated with the model and, as we shall see later, the problem is

identifiable. We will now introduce the Logistic Regression Errors-in-Variables problem.

3.5 The Logistic Regression Errors-in-Variables Problem

3.5.1 Introduction

In Chapter 2 we introduced the logistic regression model and in section 3.2 we defined the measurement error models that can be associated with the observation of the true explanatory variable values. Within this section we explain how these two models come together to produce the logistic regression measurement error model and how in the same way as for the linear regression errors-in-variables model, estimates of the model parameters can be obtained.

3.5.2 Estimation of the Model Parameters

To estimate the model parameters α_t and β_t further assumptions have to be made concerning the distributions of the random variables namely, X, Z and V. Therefore, it is assumed that the true risk factor X is Normally distributed with mean μ_x and variance σ_x^2. The measurement error variable V is also assumed to be Normally distributed with mean 0 and variance σ_v^2. It follows that Z is also Normally distributed. So the full model is

$$\text{logit}\left(P(Y=1|X=x)\right)=\alpha_t+\beta_t x$$

$$X \sim N(\mu_x, \sigma_x^2)$$

and

$$V \sim N(\mu_v, \sigma_v^2)$$

The observed likelihood is

$$L = \prod_{i=1}^{n} P(Y_i = 1|Z_i)^{Y_i} \prod_{i=1}^{n} (1 - P(Y_i = 1|Z_i))^{1-Y_i}$$

If we suppose that X and Z are discrete then

$$P(Y = 1|Z = z) = \frac{P(Y = 1 \cap Z = z)}{P(Z = z)}$$

Now,

$$P(Y = 1 \cap Z = z) = \sum_x P(Y = 1 \cap Z = z \cap X = x)$$

$$= \sum_x P(Y = 1|Z = z \cap X = x) P(Z = z \cap X = x)$$

Hence,

$$P(Y = 1|Z = z) = \frac{P(Y = 1 \cap Z = z)}{P(Z = z)} = \sum_x P(Y = 1|Z = z \cap X = x) \frac{f(z,x)}{f_Z(z)}$$

$$= \sum_x P(Y = 1|Z = z \cap X = x) f(x|z)$$

However, Z is a surrogate value for X and it is assumed that, if X is known, then Z does not contain any 'extra' information concerning the response. Therefore,

$$P(Y = 1|Z) = \sum_x P(Y = 1|X) f(X|Z)$$

The argument can be adapted to deal with the case where X and Z are continuous.

Then

$$P(Y = 1|Z \in (z, z + dz)) = \frac{P(Y = 1 \cap \{Z \in (z, z + dz)\})}{P(Z \in (z, z + dz))} =$$

$$\int P(Y = 1|X \in (x, x + dx) \cap Z \in ((z, z + dz)) P(X \in (x, x + dx) \cap Z \in ((z, z + dz)) dx$$

$$= \int P(Y=1|\ X \in (x, x+dx) \cap Z \in ((z, z+dz)) \frac{P(Y=1|\ X \in (x, x+dx) \cap Z \in ((z, z+dz))}{P(Z \in ((z, z+dz))} dx$$

On taking the limits as dx and $dz \to 0$ then

$$P(Y=1|Z) = \int P(Y=1|X) f(X|Z) dX \qquad (3.6)$$

This expression is not in a closed form due to the fact that $P(Y=1|X)$ is of logistic form and $f(X|Z)$ is Normally distributed. Hence, when the variable distributions are assumed to be Normal, then the integral cannot be evaluated analytically. This is in direct contrast to the linear case, where under the same assumptions there were analytical solutions.

The Logistic regression measurement error likelihood can then be defined by

$$\ell = LogL = \sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - p_i)$$

where $p_i = P(Y_i = 1|Z = z_i)$. In this model there are five parameters to be estimated namely, α_t, β_t, σ_v^2, σ_x^2 and μ_x.

If we return to the different measurement error models discussed in section 3.2 then the choice of measurement error model has an effect on the distribution $f(X|Z)$ and thus the integral in (3.6). Due to the differing nature of the relationship between X and Z, the form of $f(X|Z)$ is generally simpler for the Berkson measurement error model than the classical measurement error model since for the Classical model

$$Z \sim N(\mu_x, \sigma_x^2 + \sigma_v^2)$$

where

$$\mu_{X|z} = \frac{\sigma_x^2}{\sigma_z^2} z + \frac{\sigma_v^2}{\sigma_z^2} \mu_x$$

$$\sigma_{X|z}^2 = \frac{\sigma_x^2 \sigma_v^2}{\sigma_z^2}$$

and for the Berkson model

$$Z \sim N(\mu_z, \sigma_z^2)$$

where

$$\mu_{x|z} = \mu_z$$

$$\sigma_{x|z}^2 = \sigma_v^2 + \sigma_z^2$$

This has later relevance on the methods that are available to correct for measurement error that we discuss in Chapter 4 and the question of identifiability for the logistic regression measurement error model that we return to in Chapter 8.

3.6 The Probit Regression Measurement Error Model

3.6.1 Introduction

In Chapter 2 we introduced three different models for describing the relationship between an explanatory variable and a binary outcome. One of these models was the Probit Model. Though logistic regression is the most popular technique to due its relationship with the odds ratio, the probit model can also be used. For completeness the probit regression measurement error model is described. We shall return to this particular technique in chapter 8 when considering the identifiability of the measurement error models.

3.6.2 Estimation of the Model Parameters

If we return to the expression

$$P(Y = 1|Z) = \int P(Y = 1|X) f(X|Z) dX \qquad (3.6)$$

for the logistic model it was assumed that

$$P(Y=1|X)=\frac{\exp(\alpha_t+\beta_t X)}{1+\exp(\alpha_t+\beta_t X)}$$

For the probit model

$$P(Y=1|X)=\Phi(\alpha_t+\beta_t X)$$

If we assume that the measurement errors follow a classical measurement error model that is

$$\mu_{X|z}=\frac{\sigma_x^2}{\sigma_z^2}z+\frac{\sigma_v^2}{\sigma_z^2}\mu_x$$
$$\sigma_{X|z}^2=\frac{\sigma_x^2\sigma_v^2}{\sigma_z^2}$$

then we can make the substitution

$$u=\frac{x-\mu_{x|z}}{\sigma_{x|z}}$$

to obtain

$$P(Y=1|Z)=\int\Phi\left\{\alpha_t+\beta_t\left(\frac{\sigma_x\sigma_v}{\sigma_z}u+\frac{\sigma_x^2}{\sigma_z^2}z+\frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)\right\}\varphi(u)du$$

This can be re-arranged as

$$P(Y=1|Z)=\Phi\left\{\frac{\alpha_t+\beta_t\left(\frac{\sigma_x^2}{\sigma_z^2}z+\frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)}{\left(1+\beta_t^2\frac{\sigma_v^2\sigma_x^2}{\sigma_z^2}\right)^{\frac{1}{2}}}\right\} \qquad (3.7)$$

See Appendix A.4 for proof of this evaluation. If we assume that the measurement errors follow a Berkson type measurement error model that is

$$\mu_{x|z}=\mu_z$$
$$\sigma_{x|z}^2=\sigma_v^2+\sigma_z^2$$

then we can make the substitution

$$u = \frac{x - z}{\sigma_v}$$

Therefore,

$$P(Y = 1|Z) = \int \Phi\{\alpha_t + \beta_t(\sigma_v u + z)\}\varphi(u)du$$

Hence,

$$P(Y = 1|Z) = \Phi\left\{\frac{\alpha_t + \beta_t z}{(1 + \beta_t^2 \sigma_v^2)^{\frac{1}{2}}}\right\} \tag{3.8}$$

Therefore, there is an analytical expression for $P(Y = 1|Z)$ for both measurement error models. This expression is then substituted into the log likelihood

$$\log L = \sum_{i=1}^{n} y_i \log\{P(Y_i = 1|Z_i)\} + \sum_{i=1}^{n} (1 - y_i)\log\{1 - P(Y_i = 1|Z_i)\}$$

which can then be maximised, as with the logistic model, using numerical procedures to estimate the model parameters.

3.7 Conclusion

In this chapter we have studied the various forms of measurement error model that can be associated with the logistic regression measurement error model as well as introducing the different forms of subsidiary studies that are available to assess this measurement error. Not only does an assessment need to be made of the form of the measurement error but also how the measurement error model parameters will be estimated, both concepts having an effect on the form of the model. Information must also be gathered concerning the explanatory variable and whether it can be defined by a distribution or whether a non-parametric model is required. These concepts were introduced within this chapter as they have an effect on the different methods that are

available to correct for measurement error when estimating the model parameters. For example, methods have been developed for particular cases such as the Berkson measurement model or if a functional model can be assumed. These methods and concepts are further discussed and examined in the following chapters.

We have also introduced both the linear and the logistic regression errors-in-variables models. In Chapter 2 we saw that analytical estimates for the linear regression model parameters could be obtained whereas for the logistic regression model a numerical method had to be employed. When examining the linear regression errors-in-variables model, again analytical estimates of the model parameters could be obtained if certain assumptions were made about the error variances, namely, that if one or more of the error variances are known, or at least the relationship between the two is known, then closed analytical expressions are available. However, the model was not identifiable when both the measurement error variances were unknown. In contrast the logistic regression errors-in-variables model involves an integral where there is no analytical solution so the logistic regression errors-in-variables model and so is more complex than the linear regression model. As a result a number of methods have been designed to estimate the logistic regression measurement error model parameters taking into account the distribution parameters of the measurement error. These are discussed in the following chapter.

Chapter 4

4 An overview of the current methods available to correct for measurement error in logistic regression

4.1 Introduction

In Chapters 2 and 3, we introduced both the logistic regression model and associated measurement error models and in 3.5 an expression for the likelihood was obtained. As we saw, this is complex and practical methods for estimating the model parameters are need to be devised. This chapter will consider a number of methods which have been suggested and which shall be compared in Chapter 5.

In 1989, Willett discussed a number of issues related to correction for measurement error and why such techniques are not commonly employed. Firstly, he stated that

techniques should ideally incorporate any number of covariates and produce confidence intervals for the estimates that take into account the measurement error. Secondly, investigators would perhaps more willingly use correction procedures if the methods were easy to understand and implement. There are so many methods available now that an investigator could easily be confused as to which method is the most appropriate.

Over the past thirty years various errors-in-variables issues and techniques have been designed and discussed. The result is a wide and varied selection of information concerning measurement error in covariates, ranging from the effects of measurement error to particular methods devised for different studies. The following is a list of references discussing the effects that measurement error can have on measures of association. Though not all these papers look specifically at the effect on logistic regression estimates, they give a concise background to the effects of measurement error in general; see Michalek & Tripathi (1980), Kupper (1984), Lagakos (1988), Armstrong (1990), Carroll (1997), Demidenko & Spiegelman (1997), Kronmal & Shemanski (1998) and Oppenheimer & Kher (1999). The following group of references contain methods that are either out of the scope of this work or are too specific in their implementation for our study; see Stefanski & Carroll (1985), Schafer (1987), Fuller (1988), Armstrong, Whittemore & Howe (1989), Carroll (1989), Tosteson, Stefanski & Schafer (1989), Crouch & Spiegelman (1990), Richardson & Gilks (1993), Schmid & Rosner (1993), Zhao, Lee & Van Hui (1994), Bashir & Duffy (1995), Dellaportas & Stephens (1995), Haukka (1995), Mallick & Gelfand (1996), Buzas (1997), Kuchenhoff & Carroll (1997), Muller & Roeder (1997), Richardson & Leblond (1997), Spiegelman & Casella (1997), Srivasta & Shalabh (1997), Wang &

Wang (1997), Carroll, Freedman, Kipnis & Li (1998), Wang, Lin, Guitierrez & Carroll (1998), Palta & Lin (1999) and Thoresen & Laake (1999).

This chapter and the following have been conducted on the basis of Willett's paper, to try to answer some of these questions and provide recommended methods. Throughout the chapter we cover the specific details of some of the relevant methods that are available, showing the various approaches involved, whether confidence intervals can be constructed, as well as showing whether further covariates can be included in the method.

There are four classical main approaches to the logistic regression errors-in-variables problem.

- Ordinary logistic regression, also known as the 'naïve' method. This method is used if the investigator assumes that there is no error in measuring the risk factors involved. For the univariate case, the ordinary logistic regression coefficients are denoted by α_s and β_s respectively.
- Correction Factor Methods
- Structural Modelling
- Functional Modelling

In the following sections we shall discuss each of these approaches in turn.

4.2 Correction Factor Methods

A correction factor method is where the ordinary logistic regression estimator β_s is transformed by an appropriate correction factor, usually a measurement error variance

dependent quantity, to estimate the true parameter β_t. The following is a summary of some of the methods that come under the category of 'correction factor methods'.

4.2.1 Regression Calibration

One of the simplest methods to correct for measurement error in any generalised linear model, or in our specific case logistic regression, is that of Regression calibration; see Carroll, Ruppert & Stefanski (1995). It depends on an internal or external validation study to estimate the measurement error distribution and is based upon the Berkson measurement error model. The true model parameter β_t can then be estimated using the following technique.

1) From the validation study, estimate the relationship between X and Z. If there are any further associated covariates W, which are not subject to measurement error, then these too must be taken into account. For the simple case of the Berkson measurement error model, ordinary linear regression can be performed to estimate the parameters a_0, a_1 and a_2 from the following model

$$X = a_0 + a_1 Z + a_2 W + \varepsilon \qquad (4.1)$$

2) Using the above estimates $\hat{a}_0$, $\hat{a}_1$ and $\hat{a}_2$, estimate the expected true values of X_i from the Z_i and W_i in the main study.

3) Perform an ordinary logistic regression analysis on these transformed values and the dependent Y values.

No variance expression has been devised to calculate the standard errors for the parameters $\hat{\alpha}_t$ and $\hat{\beta}_t$ that take into account the estimation of the parameters $\hat{a}_0$, $\hat{a}_1$ and $\hat{a}_2$. Therefore, the standard errors are calculated using (2.19) and hence will

always underestimate the true variability in estimating $\hat{\beta}_t$. However, this method is simple and can easily be implemented using any standard statistical package.

A follow-on from the above regression calibration method is that of the linear approximation method; see Armstrong (1985) and Rosner, Spiegelman & Willett (1989). Instead of estimating the true variable X through the use of a validation study, the ordinary logistic regression parameter $\hat{\beta}_s$ is corrected by an attenuation factor. The advantage of this method is that an expression for the variance of $\hat{\beta}_t$ has been derived that takes into account the variability in estimating $\hat{a}_0$, $\hat{a}_1$ and $\hat{a}_2$.

Rosner, Spiegelman & Willett (1989) assume that the measurement error model is of a Berkson form and that X and Z are related by the linear regression model (4.1). As before, the coefficients of the linear regression model (4.1) as well as the ordinary logistic regression coefficients $\hat{\alpha}_s$ and $\hat{\beta}_s$ must be estimated. The correction method is then

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{\hat{a}_1}$$

However, if $\hat{a}_1 = 1$ this correction procedure reduces to the no errors estimate for $\hat{\beta}_t$. For the multiple covariate case see Kuha (1994). The expression for the variance of $\hat{\beta}_t$ is

$$\text{var}(\hat{\beta}_t) = \left(\frac{1}{\hat{a}_1^2}\right)\text{var}(\hat{\beta}_s) + \left(\frac{\hat{\beta}_s^2}{\hat{a}_1^4}\right)\text{var}(\hat{a}_1) \qquad (4.2)$$

See appendix A.3 for the derivation of this expression. It can clearly be seen that this estimate takes into account the variation in estimating $\hat{a}_1$, and therefore producing a less biased estimate of the variance.

The above method assumed that the errors follow a Berkson measurement error model and that $\hat{a}_1 \neq 1$ in (4.1). If it is assumed that the errors follow a classical measurement error model and that $a_0 = 0$ and $a_1 = 1$ then

$$Z = X + V$$

This approach (Rosner, Spiegelman & Willett (1990)) works in the same way as the linear approximation method that is, the ordinary logistic regression coefficient $\hat{\beta}_s$ is corrected by an attenuation factor which is known as the reliability coefficient R and can be used in the following way

1) From the validation study estimate the measurement error variance, σ_v^2.

2) In the main study, calculate the 'naïve' estimator β_s.

3) The true model parameter β_t is then estimated by

$$\hat{\beta}_t = \hat{\beta}_s \frac{\hat{\sigma}_x^2 + \hat{\sigma}_v^2}{\hat{\sigma}_x^2} \qquad (4.3)$$

where $\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_v^2}$ is the reliability coefficient R, and so the above term can be re-written in the form $\hat{\beta}_t = \hat{\beta}_s \hat{R}^{-1}$.

For multiple covariates measured with error $\hat{\beta}_t$ and $\hat{\beta}_s$ are replaced by their respective vectors of regression coefficients for the true and observed situations respectively and R is known as the multivariate reliability coefficient; see Rosner,

Spiegelman & Willett (1992). $\boldsymbol{R}$ is then expressed in terms of the variance-covariance matrices for the true and observed variables as well as the associated measurement error. Hence

$$\boldsymbol{R} = \hat{\Sigma}_x \left(\hat{\Sigma}_x + \hat{\Sigma}_v\right)^{-1}$$

where Σ_x is the variance-covariance matrix for the true vector of covariates and Σ_v is the variance-covariance matrix for the vector of associated measurement errors. Then

$$\hat{\beta}_t = \hat{\beta}_s \hat{\boldsymbol{R}}^{-1}$$

As the method is a correction factor method it is simple and easy to implement through the use of a statistical package and a calculator. However, no expressions for the associated standard errors of the estimates are given and therefore there is no way of assessing the accuracy of the correction estimates taking into account the estimation of the associated variances.

4.2.2 Approximate Methods

The next group of methods revolve around the logistic regression measurement error model integral discussed in section 3.5

$$P(Y=1|Z) = \int P(Y=1|X) f(X|Z) dX \qquad (4.4)$$

If it is assumed that the measurement errors follow a Normal distribution then this integral has been shown to have no analytical solution. By making specific assumptions and measurement error models, the following methods have been devised. The aim of this section is to give an overall picture of these methods with their associated assumptions. However, for specific details we refer the reader to the relevant papers.

Regression calibration is a simple method for correction. However, it is a crude method that does not involve the modelling of the measurement error problem. Therefore, Rosner, Spiegelman & Willett (1989) proposed a further method that is based on the integral (4.4) that produces analytical estimates of the true model parameters and provides an expression for the variance of $\hat{\beta}_t$ that takes into account the estimation of the other model parameters. The method assumes that the errors follow a Berkson measurement error model and that X and Z are related through the equation (4.1). This method is based on the assumption that the probability of the disease under consideration is very rare and hence can be approximated by an exponential function.

Consider then the integral

$$P(Y=1|Z)=\int P(Y=1|X)f(X|Z)dX \qquad (4.4)$$

and assume that the errors follow the Berkson measurement error model. Then $X|Z=z$ is $N(z,\sigma_v^2)$ and so

$$f(X|Z)=\frac{\exp\left(\frac{-(X-Z)^2}{2\sigma_v^2}\right)}{\sqrt{2\pi}\sigma_v}$$

Define $X^*=X-\bar{X}$, where $\bar{X}$ is the sample mean of the true covariate in the validation study. Then the logistic function can be re-written in the following way

$$\ln\left(\frac{P(Y=1|X^*)}{1-P(Y=1|X^*)}\right)=(\alpha_t+\beta_t\bar{X})+\beta_t X^* \qquad (4.5)$$

To be able to approximate the integral (4.4), the expression (4.5) is used to approximate $\ln(P(Y=1|X^*))$ by a second order Taylor series expansion about $X^*=0$ that is,

$$\ln(P(Y=1|X^*)) \cong \alpha_t + \beta_t \bar{X} - \ln(1+\exp(\alpha_t + \beta_t \bar{X})) + \frac{\beta_t X^*}{1+\exp(\alpha_t + \beta_t \bar{X})}$$

$$- \frac{\beta_t^2 X^{*2} \exp(\alpha_t + \beta_t \bar{X})}{2(1+\exp(\alpha_t + \beta_t \bar{X}))^2}$$

Therefore,

$$P(Y=1|X^*) \cong \frac{\exp(\alpha_t + \beta_t \bar{X})}{1+\exp(\alpha_t + \beta_t \bar{X})} \exp\left\{ \frac{\beta_t X^*}{1+\exp(\alpha_t + \beta_t \bar{X})} - \frac{\beta_t^2 X^{*2} \exp(\alpha_t + \beta_t \bar{X})}{2(1+\exp(\alpha_t + \beta_t \bar{X}))^2} \right\}$$

which can be re-written as

$$P(Y=1|X^*) \cong c_0 \exp(c_1 X^* - c_2 X^{*2})$$

where

$$c_0 = \frac{\exp(\alpha_t + \beta_t \bar{X})}{1+\exp(\alpha_t + \beta_t \bar{X})}$$

$$c_1 = \frac{\beta_t}{1+\exp(\alpha_t + \beta_t \bar{X})}$$

$$c_2 = \frac{c_1^2 c_0}{2(1-c_0)}$$

This expression can then be substituted into the integral (4.4) to produce

$$P(Y=1|Z^*) \cong \int c_0 \exp(c_1 X^* - c_2 X^{*2}) \frac{\exp(-(X^*-Z^*)^2/2\sigma_v^2)}{\sqrt{2\pi}\sigma_v} dX^*$$

where $Z^* = Z - \bar{Z}$ and $\bar{Z}$ is the sample mean of the observed values in the validation study. This integral is now in a closed form and therefore, it can be evaluated analytically. Hence, $\hat{P}(Y=1|Z^*)$ can also be expressed as

$$\hat{P}(Y=1|Z^*) \cong c_0^{'} \exp\left(c_1^{'} Z^* - c_2^{'} Z^{*2}\right)$$

where

$$c_0^{'} = \frac{c_0}{\left(1+2\sigma_v^2 c_2\right)^{\frac{1}{2}}} \exp\left(\frac{\sigma_v^2 c_1^2 / 2}{1+2\sigma_v^2 c_2}\right)$$

$$c_1^{'} = \frac{c_1}{\left(1+2\sigma_v^2 c_2\right)}$$

$$c_2^{'} = \frac{c_2}{\left(1+2\sigma_v^2 c_2\right)}$$

However, to produce an analytical estimate of the model parameters a further expression for $\hat{P}(Y=1|Z^*)$ is required. Using a second-order Taylor series the further approximation to $\hat{P}(Y=1|Z^*)$ can be made in the same way as $P(Y=1|X^*)$, resulting in

$$\hat{\hat{P}}(Y=1|Z^*) \cong c_0^{''} \exp\left(c_1^{''} Z^* - c_2^{''} Z^{*2}\right)$$

where

$$c_0^{''} = \frac{\exp\left(\alpha_s + \beta_s \bar{Z}\right)}{1+\exp\left(\alpha_s + \beta_s \bar{Z}\right)}$$

$$c_1^{''} = \frac{\beta_s}{1+\exp\left(\alpha_s + \beta_s \bar{Z}\right)}$$

$$c_2^{''} = \frac{c_1^{''2} c_0^{''}}{2\left(1-c_0^{''}\right)}$$

where α_s and β_s are the ordinary logistic regression estimates, assuming no error, calculated from the main study data set. By equating $\left(c_0^{'}, c_0^{''}\right)$, $\left(c_1^{'}, c_1^{''}\right)$ and $\left(c_2^{'}, c_2^{''}\right)$, the expressions for c_0, c_1 and c_2 can be expressed as

$$\hat{c}_2 = \frac{\hat{\beta}_s^2 \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)}{2\left[1+\exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)\right]^2 - 2\hat{\beta}_s^2 \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)\hat{\sigma}_v^2} \tag{4.6}$$

$$\hat{c}_1 = 2\hat{c}_2\left\{\frac{\left[1+\exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)\right]}{\hat{\beta}_s \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)}\right\} \tag{4.7}$$

$$\hat{c}_0 = \left(1+2\hat{\sigma}_V^2 \hat{c}_2\right)^{\frac{1}{2}}\left\{\frac{\exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)}{1+\exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)}\right\}\exp\left\{\frac{-\hat{\sigma}_V^2 \hat{c}_1^2/2}{1+2\hat{\sigma}_v^2 \hat{c}_2}\right\} \tag{4.8}$$

By solving these equations an estimate for $\hat{\beta}_t$ can be obtained from the equation

$$\hat{\beta}_t = \frac{\hat{c}_1}{1-\hat{c}_0}$$

where the unknown parameters $\overline{Z}$ and $\hat{\sigma}_v^2$ are estimated from the validation study whilst $\hat{\alpha}_s$ and $\hat{\beta}_s$ are the usual 'naïve' estimates.

This method was investigated further to see whether the method reduces to the usual no errors estimate $\hat{\beta}_s$ when there is no error. Let $W = \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)$ then

$$c_2 = \frac{\hat{\beta}_s^2 \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)}{2\left[1+\exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)\right]^2 - 2\hat{\beta}_s^2 \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)\hat{\sigma}_V^2} = \frac{\hat{\beta}_s^2 W}{2(1+W)^2 - 2\hat{\beta}_s^2 W \hat{\sigma}_v^2}$$

For c_1

$$c_1 = 2c_2\left\{\frac{\left[1+\exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)\right]}{\hat{\beta}_s \exp\left(\hat{\alpha}_s + \hat{\beta}_s \overline{Z}\right)}\right\} = 2\left[\frac{1}{2}\left\{\frac{\hat{\beta}_s^2 W}{(1+W)^2 - \hat{\beta}_s^2 W \hat{\sigma}_v^2}\right\}\right]\left\{\frac{1+W}{\hat{\beta}_s W}\right\}$$

Therefore,

$$c_1 = \frac{\hat{\beta}_s(1+W)}{(1+W)^2 - \hat{\beta}_s^2 W \hat{\sigma}_v^2}$$

For c_0

$$c_0 = \left(1+2\hat{\sigma}_V^2\hat{c}_2\right)^{\frac{1}{2}}\left\{\frac{\exp\left(\hat{\alpha}_s+\hat{\beta}_s\bar{Z}\right)}{1+\exp\left(\hat{\alpha}_s+\hat{\beta}_s\bar{Z}\right)}\right\}\exp\left\{\frac{-\hat{\sigma}_V^2\hat{c}_1^2/2}{1+2\hat{\sigma}_V^2\hat{c}_2}\right\}$$

$$= \left(1+\frac{\hat{\beta}_s^2W\hat{\sigma}_v^2}{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}\right)^{\frac{1}{2}}\frac{W}{1+W}\exp\left\{\frac{-\frac{\hat{\sigma}_v^2\hat{\beta}_s^2(1+W)^2}{\left[(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2\right]^2}\Big/2}{1+\frac{\hat{\sigma}_v^2\hat{\beta}_s^2W}{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}}\right\}$$

$$= \frac{W}{\sqrt{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}}\exp\left\{\frac{-\frac{\hat{\sigma}_v^2\hat{\beta}_s^2(1+W)^2}{\left[(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2\right]^2}\Big/2}{1+\frac{\hat{\sigma}_v^2\hat{\beta}_s^2W}{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}}\right\}$$

So

$$\hat{\beta}_t = \frac{\frac{\hat{\beta}_s(1+W)}{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}}{1-\frac{W}{\sqrt{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}}\exp\left\{\frac{-\frac{\hat{\sigma}_v^2\hat{\beta}_s^2(1+W)^2}{\left[(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2\right]^2}\Big/2}{1+\frac{\hat{\sigma}_v^2\hat{\beta}_s^2W}{(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2}}\right\}}$$

Therefore, as $\hat{\sigma}_v \to 0$ then $\hat{\beta}_t \to \frac{\hat{\beta}_s/(1+W)}{1-\frac{W}{1+W}} = \hat{\beta}_s$

Hence, this method reduces to the no errors model, when $\hat{\sigma}_v \to 0$. In addition, this method puts a condition on the estimate $\hat{\sigma}_v$ for this method to give a meaningful estimate,

$$(1+W)^2-\hat{\beta}_s^2W\hat{\sigma}_v^2 > 0$$

Hence,

$$\hat{\sigma}_v^2 < \frac{(1+W)^2}{\hat{\beta}_s^2W}$$

Therefore, using a validation study and the estimates of $\hat{\alpha}_s$ and $\hat{\beta}_s$ produced from any standard statistical package, we have an analytical expression for $\hat{\beta}_t$ and hence $\hat{\alpha}_t$. The other advantage of this method is that an expression for the variance of $\hat{\beta}_t$ can be obtained. As before, this variance expression takes into account the variability from estimating the various parameters in the validation study. The subsequent formulation is the result of using the delta method as given in appendix A.3.

$$\mathrm{var}\left(\hat{\beta}_t\right) \cong \left(\frac{\partial f}{\partial \hat{\alpha}_s}\right)^2 \mathrm{var}\left(\hat{\alpha}_s\right) + \left(\frac{\partial f}{\partial \hat{\beta}_s}\right)^2 \mathrm{var}\left(\hat{\beta}_s\right) + \left(\frac{\partial f}{\partial \overline{Z}}\right)^2 \mathrm{var}\left(\overline{Z}\right)$$

$$+ \left(\frac{\partial \mathrm{f}}{\partial \hat{\sigma}_{\mathrm{v}}^2}\right)^2 \mathrm{var}\left(\hat{\sigma}_{\mathrm{v}}^2\right) + 2\left(\frac{\partial \mathrm{f}}{\partial \hat{\alpha}_{\mathrm{s}}}\right)\left(\frac{\partial \mathrm{f}}{\partial \hat{\beta}_{\mathrm{s}}}\right)\mathrm{cov}\left(\hat{\alpha}_{\mathrm{s}}, \hat{\beta}_{\mathrm{s}}\right)$$

where $f = \dfrac{\hat{c}_1}{1-\hat{c}_0}$ and the other covariances are 0 due to independence.

$$\frac{\partial f}{\partial \hat{\alpha}_s} = \left(\frac{\partial \hat{c}_1}{\partial \hat{\alpha}_s}\right)/\left(1-\hat{c}_0\right) + \left(\frac{\partial \hat{c}_0}{\partial \hat{\alpha}_s}\right)\left[\frac{\hat{c}_1}{\left(1-\hat{c}_0\right)^2}\right]$$

$$\frac{\partial f}{\partial \hat{\beta}_s} = \left(\frac{\partial \hat{c}_1}{\partial \hat{\beta}_s}\right)/\left(1-\hat{c}_0\right) + \left(\frac{\partial \hat{c}_0}{\partial \hat{\beta}_s}\right)\left[\frac{\hat{c}_1}{\left(1-\hat{c}_0\right)^2}\right]$$

$$\frac{\partial f}{\partial \hat{\sigma}_v^2} = \left(\frac{\partial \hat{c}_1}{\partial \hat{\sigma}_v^2}\right)/\left(1-\hat{c}_0\right) + \left(\frac{\partial \hat{c}_0}{\partial \hat{\sigma}_v^2}\right)\left[\frac{\hat{c}_1}{\left(1-\hat{c}_0\right)^2}\right]$$

$$\mathrm{var}\left(\overline{Z}\right) = \sum_{i=1}^{n} \frac{\left(Z_i - \overline{Z}\right)^2}{n(n-1)}$$

$$\mathrm{var}\left(\hat{\sigma}_v^2\right) = \frac{2\hat{\sigma}_v^4}{n-2}$$

As before, $\overline{Z}$ and $\hat{\sigma}_v^2$ are estimated from the validation study as are the associated model parameter variances. The derivatives, such as, $\frac{\partial \hat{c}_1}{\partial \hat{\alpha}_s}$, are calculated from equations (4.6), (4.7) and (4.8). The rest of the parameters are estimated from the main study data set.

This simple analytical formula has one drawback, the method cannot be extended to include further covariates. Therefore, the reader is directed to the multivariate version of the simple correction factor method discussed in section 4.2.1.

One further approximate method was devised by Reeves, Cox, Darby & Whitley (1998). The advantages of this method is that it can incorporate either a classical or Berkson measurement error model and is also a correction factor method that can be implemented using any statistical package. The disadvantage of this paper compared to that of Rosner, Spiegelman & Willett (1989), is that no formula for the variance of $\hat{\beta}_t$ incorporating the error of estimating the parameters in the validation study is included. However, as the examples and simulation study will show, this method works extremely well at correcting for measurement error. This method uses the measurement error model described in section 3.2.4.

So that the Reeves error structure can be utilised, a different form of the conditional distribution of $X|Z$ is used. This form can be found in Cox and Hinkley (1974).

$$X = \mu_x + \gamma_{x.z}(Z - \mu_x) + \xi_{x.z} \tag{4.9}$$

where

$$\mu_x = E(X) = E(\tilde{X}) = E(Z) \tag{4.10}$$

and

$$\gamma_{x.z} = \frac{Cov(X,Z)}{Var(Z)} \tag{4.11}$$

Now

$$Cov(X,Z) = Cov(\tilde{X}+U, \tilde{X}+V)$$

$$= Var(\tilde{X}) \qquad \text{by independence}$$

$$= \sigma_x^2$$

Therefore,

$$\gamma_{x.z} = \frac{Cov(X,Z)}{Var(Z)} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$$

As $(\tilde{X}, V, U)$ are all normally distributed then (4.9) is a normal linear regression model with $\xi_{x.z}$ as the error term independent of Z with mean zero and variance $\sigma_{x.z}^2$. If we return to (4.9), it can be re-written as

$$\xi_{x.z} = (X - \mu_x) - \gamma_{x.z}(Z - \mu_x)$$

By substituting into (3.1)

$$X - \mu_x = \tilde{X} - \mu_x + U$$

and

$$Z - \mu_x = \tilde{X} - \mu_x + V$$

we have

$$\xi_{x.z} = (\tilde{X} - \mu_x)(1 - \gamma_{x.z}) + U - \gamma_{x.z} V$$

Hence $E(\xi_{x.z}) = 0$ and

$$Cov(\xi_{x.Z}, Z) = Cov(Z, X) - \gamma_{x.z}\, Cov(Z, Z)$$

$$= Var(Z)\gamma_{x.z} - \gamma_{x.z}Var(Z)$$

$$= 0$$

As $E(\gamma_{x.z}) = 0$ and $(\tilde{X}, V, U)$ are independent the variance term can be written as

$$Var(\xi_{x.z}) = (1-\gamma_{x.z})^2 Var(\tilde{X} - \mu_x) + Var(U) + \gamma_{x.z}^2 Var(V)$$

$$= \left(\frac{\sigma_v^2}{\sigma_x^2 + \sigma_v^2}\right)^2 \sigma_x^2 + \sigma_u^2 + \frac{\sigma_x^4}{(\sigma_x^2 + \sigma_v^2)^2}\sigma_v^2$$

$$= \sigma_u^2 + \frac{\sigma_x^2 \sigma_v^2}{\sigma_x^2 + \sigma_v^2}$$

$$= \sigma_u^2 + \sigma_x^2 - \frac{\sigma_x^4}{\sigma_x^2 + \sigma_v^2}$$

$$= \sigma_u^2 + \sigma_x^2 - \frac{\sigma_x^4}{(\sigma_x^2 + \sigma_v^2)^2}(\sigma_x^2 + \sigma_v^2)$$

Hence,

$$Var(\xi_{x.z}) = \sigma_{x.z}^2 = \sigma_u^2 + \sigma_x^2 - \gamma_{x.z}^2(\sigma_x^2 + \sigma_v^2) \qquad (4.12)$$

Using this new error structure and the reparameterized conditional distribution of $X|Z$, Reeves et al produced a simple analytical method that corrects the model parameters for measurement error providing the error model satisfies (3.1).

The actual methodology of this approach revolves around the close relationship between the logistic and probit models that was described earlier in chapter 2. The true logistic model is assumed to be approximately probit which allows for the new error structure model to be utilised. This model is then arranged so that it is in the form of a logistic model and so can be transformed back into the logistic domain. Therefore, the observed model parameter β_s is expressed in terms of β_t and the measurement error parameters σ_v^2 and σ_u^2. After re-arranging, the true model parameter β_t is in fact estimated by using a correction term associated with β_s.

Consider the usual simple linear logistic model as defined before

$$P(Y=1|X)=\frac{\exp(\alpha_t+\beta_t X)}{1+\exp(\alpha_t+\beta_t X)} \tag{4.13}$$

Reeves et al define the relationship between logistic and probit as

$$\frac{e^t}{1+e^t}\approx\Phi(kt) \tag{4.14}$$

where $\Phi(.)$ is the cumulative standard normal distribution function and $k\approx 0.588$.

Therefore, (4.14) can be re-written in the form

$$P(Y=1|X)=\frac{\exp(\alpha_t+\beta_t X)}{1+\exp(\alpha_t+\beta_t X)}\approx\Phi\{k(\alpha_t+\beta_t X)\}$$

Now

$$\Phi\{k(\alpha_t+\beta_t X)\}=P(W\le k(\alpha_t+\beta_t X))$$

where W is a standardised normal random variable. Conditioning on Z and using (4.9)

$$P(Y=1|Z)\approx P(W\le k(\alpha_t+\beta_t X)|Z)=P(W\le k(\alpha_t+\beta_t\{\mu_x+\gamma_{x.z}(Z-\mu_x)+\xi_{x.z}\}))$$

Hence,

$$P(Y=1|Z)\approx P[W-k\beta_t\xi_{x.z}\le k\{\alpha_t+\beta_t\mu_x(1-\gamma_{x.z})+\beta_t\gamma_{x.z}Z\}]$$

where μ_x and $\gamma_{x.z}$ are as defined in (4.10) and (4.11) respectively. Now

$$E(W-k\beta_t\xi_{x.z})=0$$

and

$$Var(W-k\beta_t\xi_{x.z})=1+k^2\beta_t^2\sigma_{x.z}^2$$

As $(\tilde{X},V,U)$ are all normally distributed, $W-k\beta_t\xi_{x.z}$ is Normally distributed. Hence,

$$P(Y=1|Z)\approx\Phi\left[\frac{k(\alpha_t+\beta_t\mu_x(1-\gamma_{x.z})+\beta_t\gamma_{x.z}Z)}{(1+k^2\beta_t^2\sigma_{x.z}^2)^{\frac{1}{2}}}\right]$$

using (4.14)

$$P(Y=1|Z) \approx \frac{\exp\left(\frac{(\alpha_t + \beta_t \mu_x (1-\gamma_{x.z}) + \beta_t \gamma_{x.z} Z)}{(1+k^2\beta_t^2\sigma_{x.z}^2)^{\frac{1}{2}}}\right)}{1+\exp\left(\frac{k(\alpha_t + \beta_t \mu_x (1-\gamma_{x.z}) + \beta_t \gamma_{x.z} Z)}{(1+k^2\beta_t^2\sigma_{x.z}^2)^{\frac{1}{2}}}\right)}$$

This is a logistic model based on the observed data Z and hence,

$$\alpha_s = \frac{\alpha_t + \beta_t \mu_x (1-\gamma_{x.z})}{(1+k^2\beta_t^2\sigma_{x.z}^2)^{\frac{1}{2}}}$$

and

$$\beta_s = \frac{\beta_t \gamma_{x.z}}{(1+k^2\beta_t^2\sigma_{x.z}^2)^{\frac{1}{2}}}$$

On re-arrangement the true model parameter β_t can be estimated in terms of β_s, $\gamma_{x.z}$ and $\sigma_{x.z}^2$. Hence,

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{(\hat{\gamma}_{x.z}^2 - k^2\hat{\sigma}_{x.z}^2\hat{\beta}_s^2)^{\frac{1}{2}}} \qquad (4.15)$$

where $\hat{\beta}_s$ is the logistic model estimate from the main study using the observed data values and

$$\hat{\gamma}_{x.z}^2 = \frac{\hat{\sigma}_x^4}{(\hat{\sigma}_z^2 + \hat{\sigma}_v^2)^2}$$

$$\hat{\sigma}_{x.z}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_x^2 - \hat{\gamma}_{x.z}^2(\hat{\sigma}_x^2 + \hat{\sigma}_v^2)$$

can be estimated from both the validation study and the main study. No variance estimate for $\hat{\beta}_t$ taking into account the estimation of β_s, $\gamma_{x.z}$ and $\sigma_{x.z}^2$ is given. Hence, the accuracy of this estimate is determined by the usual variance estimator (2.19).

For the multivariate case, this formulation translates to

$$\boldsymbol{\beta}_s' = \boldsymbol{\beta}_t' \boldsymbol{\Gamma}_{X.Z}(1+k^2 \boldsymbol{\beta}_t' \boldsymbol{\Sigma}_{X.Z} \boldsymbol{\beta}_t')^{-\frac{1}{2}}$$

where

$$\Sigma_{X.Z} = \Sigma_X \Sigma_V (\Sigma_X + \Sigma_V)^{-1}$$
$$\Gamma_{X.Z} = \Sigma_X (\Sigma_X + \Sigma_V)^{-1}$$

For the univariate case, if the measurement error is known to follow a classical measurement error model then $\sigma_u^2 = 0$. Hence,

$$\hat{\gamma}_{x.z}^2 = \frac{\hat{\sigma}_x^4}{(\hat{\sigma}_x^2 + \hat{\sigma}_v^2)^2} \tag{4.16}$$

and

$$\hat{\sigma}_{x.z}^2 = \hat{\sigma}_x^2 - \hat{\gamma}_{x.z}^2 (\hat{\sigma}_x^2 + \hat{\sigma}_v^2) \tag{4.17}$$

which can be re-written in the form

$$\hat{\sigma}_{x.z}^2 = \hat{\sigma}_x^2 - \left(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_v^2}\right)^2 (\hat{\sigma}_x^2 + \hat{\sigma}_v^2)$$

$$= \frac{\hat{\sigma}_v^2 \hat{\sigma}_x^2}{\hat{\sigma}_v^2 + \hat{\sigma}_x^2} \tag{4.18}$$

Therefore, after the substitution of (4.16) and (4.18) the estimate for β_t, (4.15) is

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{\left(\frac{\hat{\sigma}_x^4}{(\hat{\sigma}_x^2 + \hat{\sigma}_v^2)^2} - k^2 \hat{\beta}_s^2 \frac{\hat{\sigma}_v^2 \hat{\sigma}_x^2}{\hat{\sigma}_x^2 + \hat{\sigma}_v^2}\right)^{\frac{1}{2}}} \tag{4.19}$$

If we consider this Reeves estimate (4.19) in comparison to the attenuation factor method (4.3) presented in section 4.2.1, then the Reeves estimator for $\hat{\beta}_t$ can be re-written in terms of the reliability coefficient such that

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{\sqrt{\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^2+\hat{\sigma}_v^2}\left(\frac{\hat{\sigma}_x^2}{\left(\hat{\sigma}_x^2+\hat{\sigma}_v^2\right)}-k^2\hat{\beta}_s^2\hat{\sigma}_v^2\right)^{\frac{1}{2}}}}$$

and so

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{\sqrt{R}\left(R-k^2\hat{\beta}_s^2\hat{\sigma}_v^2\right)^{\frac{1}{2}}}$$

where R is the reliability coefficient in (4.3). The comparison between these two methods will be studied further in Chapter 9.

If the measurement error follows a Berkson type measurement error model then $\sigma_v^2 = 0$. Hence,

$$\hat{\gamma}_{x.z}^2 = 1$$

and

$$\hat{\sigma}_{x.z}^2 = \hat{\sigma}_u^2 + \hat{\sigma}_x^2 - 1.\left(\hat{\sigma}_x^2\right) = \hat{\sigma}_u^2$$

Therefore, the estimate of β_t for the case of the Berkson measurement error model is

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{\left(1-k^2\hat{\beta}_s^2\hat{\sigma}_u^2\right)^{\frac{1}{2}}}$$

For both measurement error models $\left|\hat{\beta}_s\right| \leq \left|\hat{\beta}_t\right|$ and the bias in $\hat{\beta}_s$ increases as σ_v^2 and σ_u^2 increase respectively.

All the model parameters can therefore be estimated using standard techniques which makes this method extremely attractive due to its simplicity and ease of use.

The above review has given an overall picture of the current correction factor method procedures. The result is a number of methods for different types of situation and measurement error models, providing simple, easy-to-use approximations to the true model. However, these methods are all approximations, and so exact methods should also be considered.

4.3 Modelling Methods

In the previous section we described simple approximate analytical methods using standard statistical techniques. However, as the methods were approximating the true situation, so the resulting estimates of the model parameters were only approximate. These methods were devised out of a need for simplicity as opposed to the need for the most efficient estimate. Therefore, techniques have been devised to find consistent estimators for the model parameters. These techniques are in the more general form of modelling procedures. The concept of structural and functional modelling was introduced in Chapter 3 however it inevitably leads to the question of which is the most appropriate approach to use. As the literature shows, there is no clean-cut answer to this question. The structural modelling approach leads to a likelihood based approach. Full specification of the data distributions can then lead to a set of efficient estimators and applies to far more general problems. However, if there is concern over the distribution of the X_i then the semi parametric method of functional modelling should be employed. In a medical study context, a structural model is the most likely to arise naturally.

The following review will show how these approaches can be applied to the logistic regression errors-in-variables problem and will discuss their associated advantages and disadvantages. In an attempt to produce a clearer answer to the question proposed above, the two approaches will be compared in the simulation study. This comparison, as far as we know, has not been made anywhere else. Special attention will also be given to the question of robustness, that is, how well each approach works when the distributions of both the unobserved variable and the measurement errors have been miss-specified. Thus a comparison will be given for the overall efficiency of each of the approaches.

4.3.1 Structural Modelling

The main problem with the logistic regression errors-in-variables problem was there was no analytical solution to

$$P(Y=1|Z)=\int P(Y=1|X)f(X|Z)dX \qquad (4.4)$$

However, this does not mean that the problem is unsolvable. By fully specifying the distributions of the data, the likelihood can be constructed and maximum likelihood techniques can be employed to estimate the model parameters, using numerical procedures. The disadvantage of this approach is that it involves extensive programming. The advantage is that the model does not rely on an approximation and any conditions that may go with the approximation plus the model can include any number of covariates. Estimates of the standard errors can also be obtained using the information matrix. There are two such approaches, numerical optimisation of the

logistic log likelihood and numerical optimisation of the probit log likelihood where the probit model is assumed to be an approximation to the logistic model.

4.3.2 The Logistic Log likelihood

For the logistic log likelihood it is assumed that the true covariate is related to the dependent disease status by the logistic model, that is

$$P(Y=1|X)=\frac{e^{\alpha_t+\beta_t X}}{1+e^{\alpha_t+\beta_t X}}$$

If it is assumed that the measurement errors follow a classical measurement error model, as was shown in Chapter 3, then $P(Y=1|Z)$ can be expressed as

$$P(Y=1|z)=\int\frac{e^{\alpha_t+\beta_t z}}{1+e^{\alpha_t+\beta_t z}}\varphi\left(\frac{x-\mu_{x|z}}{\sigma_{x|z}}\right)dz \qquad (4.20)$$

where

$$\mu_{x|z}=\frac{\sigma_x^2}{\sigma_z^2}z+\frac{\sigma_v^2}{\sigma_z^2}\mu_x$$

$$\sigma_{x|z}^2=\frac{\sigma_x^2\sigma_v^2}{\sigma_z^2}$$

and $\varphi(.)$ is the standard normal density function. If it is assumed that the measurement errors follow a Berkson type measurement error model, then $P(Y=1|Z)$ can be expressed as

$$P(Y=1|z)=\int\frac{e^{\alpha_t+\beta_t x}}{1+e^{\alpha_t+\beta_t x}}\frac{1}{\sigma_v}\varphi\left(\frac{x-z}{\sigma_v}\right)dx$$

In either case the log likelihood is then defined to be

$$LogL=\sum_{i=1}^{n}y_i\log\{P(Y=1|Z)\}+\sum_{i=1}^{n}(1-y_i)\log\{1-P(Y=1|Z)\}$$

To obtain estimates of the regression coefficients a numerical procedure such as optimisation is used to maximise the log likelihood.

A program was written in Fortran to implement this method. Hence, Nag routines were used for the numerical integration and optimisation procedures, see section 2.6. Details of the Nag routines can be found at their web site. These two Nag routines used 1-D quadrature to approximate the integral and a Quasi-Newton algorithm to optimise the log likelihood. However, other numerical procedures can be utilised.

Standard errors of the regression coefficients can be obtained using the usual variance function. Again, this statistic does not take into account the estimation process of the other model parameters and hence will always over-estimate the accuracy of these estimates of the regression coefficients. This method can also be extended to include further covariates, by substituting the multivariate equivalent parameters from section 2.6 in the above formulae.

4.3.3 The Probit log likelihood

For the logistic log likelihood there was no analytical solution to $P(Y=1|Z)$. However, it was shown by Carroll, Spiegelman, Lan, Bailey & Abbott (1984) that by replacing the logistic model by the probit model, then an analytical solution could be found. Unlike the Reeves method that assumed the relationship

$$G(t) = \Phi(0.588t)$$

where $\Phi(t)$ is the cumulative standard normal density function, Carroll et al assume that the two models are close enough in approximation that they are interchangeable and hence the probit model can be used in place of the logistic model.

As we saw in Chapter 3, the two measurement error models, Classical and Berkson, produce the following expressions for $P(Y=1|Z)$

$$P(Y=1|Z)=\Phi\left\{\frac{\alpha_t+\beta_t\left(\frac{\sigma_x^2}{\sigma_z^2}z+\frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)}{\left(1+\beta_t^2\frac{\sigma_v^2\sigma_x^2}{\sigma_z^2}\right)^{\frac{1}{2}}}\right\} \tag{4.21}$$

and

$$P(Y=1|Z)=\Phi\left\{\frac{\alpha_t+\beta_t z}{\left(1+\beta_t^2\sigma_v^2\right)^{\frac{1}{2}}}\right\} \tag{4.22}$$

respectively , which can be substituted into the log likelihood

$$\log L=\sum_{i=1}^{n}y_i\log\{P(Y_i=1|Z_i)\}+\sum_{i=1}^{n}(1-y_i)\log\{1-P(Y_i=1|Z_i)\}$$

to be maximised using numerical procedures to estimate the model parameters.

This method eliminates the need to approximate the integral using numerical methods however, an optimisation procedure is still needed to maximise the log likelihood. Therefore, this method still has some of the same computational problems as the logistic likelihood.

The accuracy of the estimates of the regression coefficients can be obtained from the observed information matrix, which is calculated from the matrix of the second

derivatives of the log likelihood. Also, it can easily be extended to the multivariate case by substituting the multivariate parameters in the expressions (4.21) and (4.22) respectively.

4.3.4 Conclusion

To be able to perform structural modelling it is assumed that every component of the data can be fully specified by a parametric model. If such assumptions can be made, the benefits of a likelihood-based approach are asymptotically unbiased minimum variance estimators and inferences that are not based on approximate techniques such as bootstrapping. As there is no analytical solution to this problem, to perform a likelihood based approach numerical procedures are required. Maximisation of the logistic log likelihood involves numerical integration to evaluate $P(Y=1|Z)$ and then numerically optimising the log likelihood. If it is assumed that the probit model can be interchanged with the logistic model then numerically integration is not required. Standard errors are then based on the usual variance expression. Therefore, not only will these methods provide the standard advantages of using a likelihood based approach but they can also be adapted to include any number of covariates. Though these methods require extensive programming to implement them, the estimates produced have very attractive properties.

4.3.5 Functional Modelling: Conditional Score Method

The conditional score method approaches the logistic regression errors-in-variables problem slightly differently from that of the likelihood approach. It still assumes that the measurement errors are independently and normally distributed but no

distributional specification of the unobserved variable X need be made. This is in direct contrast with the likelihood based approach and therefore, fewer assumptions need be made concerning the data in order to implement this method. Though this approach is also algebraically more complex than previous methods, it produces theoretically consistent estimators. Formulae are also given for the variance of the estimators and any number of covariates can be included in the model.

4.3.6 The Model

To explain how the method works we will begin by explaining the asymptotic theory behind unbiased estimating functions, of which the conditional score function is a specific version. For the multivariate case an unbiased estimating equation is defined to be

$$\sum_{i=1}^{n} \Psi(Y_i, \beta) = 0 \qquad (4.23)$$

where n is the sample size, β is the vector of parameters to be estimated , $\mathbf{Y}_i$ are the observed data, and $\psi(.)$ is the score function. The solution to these set of equations, $\hat{\beta}$, are also known as M-estimators. These equations are said to be conditionally unbiased if

$$E\{\sum_{i=1}^{n} \Psi(Y_i, \beta)\}=0$$

Therefore, if these equations are unbiased then under certain regularity conditions $\hat{\beta}$ is a consistent estimator of β. A proof of this statement can be found in Huber(1967).

To estimate the univariate logistic regression errors-in-variables model parameters through unbiased estimating equations a special form of the unbiased estimating function (4.23) must be derived. The conditional score is calculated as

$$\psi\left(\mathrm{Y},\mathrm{Z},\boldsymbol{\beta}\,\sigma_v^2\right)=\left[Y-G\left(\alpha_t+\beta_t\delta_i-\frac{1}{2}\beta_t^2\sigma_v^2\right)\right]\begin{bmatrix}1\\ \delta_i\end{bmatrix} \qquad (4.24)$$

where

$$\delta_i = z_i + y_i\sigma_v^2\beta_t$$

Therefore, to use this score function, only the distribution of the measurement errors need be specified; no information is required concerning the unknown variable X. This is in direct contrast with the likelihood approach. The score function (4.24) is then substituted back into the formula

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathrm{Y},\mathrm{Z},\boldsymbol{\beta},\sigma_v^2)=0$$

which can then be solved for $\boldsymbol{\beta}$ by an iterative procedure such as the Newton-Raphson method. This is known as the conditional-score method.

To implement the conditional score method, the iterative procedure of Newton-Raphson is employed. For the univariate case (4.24) is defined by the two functions namely

$$f_1 = \sum_{i=1}^{n}\left(y_i - G\left(\alpha_t+\beta_t\delta_i-\frac{1}{2}\beta_t^2\sigma_v^2\right)\right)=0$$

$$f_2 = \sum_{i=1}^{n}\left(y_i - G\left(\alpha_t+\beta_t\delta_i-\frac{1}{2}\beta_t^2\sigma_v^2\right)\right)\delta_i=0$$

where $G(.)$ is the usual logistic model and

$$\delta_i = z_i + y_i\sigma_v^2\beta_t$$

The first derivatives are calculated to be

$$\frac{\partial f_1}{\partial \alpha_t} = -\sum_{i=1}^{n} \frac{\exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)}{\left[1 + \exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)\right]^2}$$

$$\frac{\partial f_1}{\partial \beta_t} = -\sum_{i=1}^{n} \frac{\exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)}{\left[1 + \exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)\right]^2} \left(\delta_i + \beta_t \sigma_v^2 y_i - \beta_t \sigma_v^2\right)$$

$$\frac{\partial f_2}{\partial \alpha_t} = -\sum_{i=1}^{n} \frac{\exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)}{\left[1 + \exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)\right]^2} \delta_i$$

$$\frac{\partial f_2}{\partial \beta_t} = \sum_{i=1}^{n} y_i^2 \sigma_v^2 -$$

$$\sum_{i=1}^{n} \frac{\exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)}{\left[1 + \exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)\right]^2} \left\{ \delta_i \left(\delta_i + \beta_t \sigma_v^2 y_i - \beta_t \sigma_v^2\right) + y_i \sigma_v^2 \left(1 + \exp\left(\alpha_t + \beta_t \delta_i - \frac{1}{2}\beta_t^2 \sigma_v^2\right)\right)\right\}$$

These equations are then substituted into the Newton-Raphson formula that can be found in section 2.6.2 and hence estimates for α_t and β_t can be obtained. The asymptotic standard errors of these estimates can be calculated using the bootstrap or the sandwich formula, both of which are discussed in Carroll, Ruppert & Stefanski (1995).

The asymptotic theory for unbiased estimating equations was given in terms of the multivariate case. Therefore, the conditional score method can be adapted to include

any number of continuous covariates by substituting the multivariate parameters into the above formulae (4.24).

4.3.7 Conclusion

We have shown that by using the results of unbiased estimating equations, consistent estimators can be obtained for the model parameters. No information is required concerning the unobserved X values, so reducing the number of assumptions that are needed to be made from the data. Therefore, the unbiased estimating equations method is theoretically sound without making extreme assumptions concerning the data. However, the simulation study will show how well this method works in comparison to the more simple approximate methods.

4.4 Conclusion

This Chapter has been designed to provide a summary of the various types of logistic regression measurement error methods, explaining the method and the associated model assumptions. Each of the methods discussed has been shown to have advantages and disadvantages related to them, which leads to the questions of how well the methods work in practice and in comparison to each other. In Chapter 5, we investigate into the performance of these methods studying how the methods react to increasing levels of measurement error as well as a decreasing relationship between the risk factor and the disease status. In chapter 6 we will examine their robustness to some of the assumptions which have been made here.

Chapter 5

5 Comparison of methods

5.1 Introduction

In Chapter 4 we introduced various methods that have been devised for correcting for measurement error when estimating the model parameters in the logistic regression measurement error problem. These methods were accompanied by the associated assumptions that need to be satisfied in order for the methods to be used. This leads to the question of which method is best to use given the assumptions and ease of implementation. In order to answer this question a comparison of the methods is required. However, this comparison cannot be done theoretically as there are no standard error calculations to compare and so a simulation study has been conducted.

For the basis of this comparison and simulation study we have chosen one method from each section within chapter 4 namely

- Ordinary logistic regression assuming no measurement error – 'Olr'
- Correction Factor method – 'Reeves'

- Structural Modelling Logistic Log Likelihood – 'Optimisation'
- Functional Modelling – 'Conditional Score'

The methods are compared in a number of situations in which the logistic regression model is varied and, for each of these, a range of values is taken for the measurement error standard deviation. This chapter considers the cases where the associated model assumptions hold for each method, we will go on to look at the effect of departures from these assumptions in chapter 6. Before presenting the results of the simulation study, however, two specific examples of real situations are presented and contrasted.

5.2 Framingham Heart Study

The Framingham study was a large cohort study conducted in the self-contained town of Framingham in Masschusetts, USA, to determine the epidemiology of a number of clinical illnesses, particularly coronary heart disease (CHD) and hypertension. The study was set up in 1949 and the subjects were then followed-up over a number of examinations. At each examination the values of a number of variables, including systolic blood pressure and cholesterol, were measured as well as information as to the presence of certain diseases. Our interest lies in the status of blood pressure as a risk factor for coronary heart disease. However, it was well noted when the study was conducted, see Dawber (1980), that measurements of blood pressure were not always accurate and therefore, could be assumed to be subject to measurement error.

Our data set consists of 1406 subjects from the first examination consisting of both male and female subjects within an age range of 45-62. The presence of coronary heart disease was recorded as well as the subject's systolic blood pressure according to the following protocol. The subject was not allowed to consume food or alcohol, or

smoke within the hour previous to the examination. Whilst the subject was sitting down, the systolic blood pressure was recorded at the first appearance of sound as the pressure was lowered.

We will consider a similar example in more detail in Chapter 9; for now a summary of results will be presented. It was found that the systolic blood pressure measurements did not follow a Normal distribution. Hence, the transformation to $\log(SBP-50)$, see Dawber (1980), was suggested to give a normally distributed variable so that methods discussed in Chapter 4 can be applied.

In a univariate logistic regression model, Y corresponds to coronary heart disease, Z is the transformed observed systolic blood pressure, X is the transformed true systolic blood pressure and V is the error incurred in measuring X. By assumption, that

$$\ln(Z'-50)=\ln(X'-50)+V$$

where X' and Z' are systolic blood pressures.

Therefore, a validation study is used to calculate the error in measuring the systolic blood pressure. For this study, the validation study is based on the repeated measurements of the systolic blood pressure from examinations 2 and 3, two and four years after the initial examination respectively. Therefore, the reproducibility study consisted of 1,615 subjects, see Carroll, Ruppert & Stefanski (1995). From these examinations the following statistics were obtained.

	Mean	Standard Deviation
Examination 2	4.374	0.226
Examination 3	4.355	0.229
Difference	0.019	0.159

Table 5.1: Table of statistics for each examination from the Framingham Heart Study

It is assumed that the measurement error follows a classical measurement error model and that the measurement error has the same distribution for all three examinations. Therefore,

Examination 2

$$Z_2 = X + V_2$$

Examination 3

$$Z_3 = X + V_3$$

Now,

$$\hat{SD}(Z_3 - Z_2) = 0.159$$

Since

$$\hat{SD}(Z_3 - Z_2) = \sqrt{2}\sigma_v$$

$$\hat{\sigma}_v = \frac{\hat{SD}(Z_3 - Z_2)}{\sqrt{2}} = 0.1124$$

To determine whether there is any evidence for systolic blood pressure being a risk factor for coronary heart disease, a logistic model was fitted using the methods compared in the simulation study and using the above estimate of the standard deviation of the measurement error distribution.

	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$
Naïve	-8.823	1.098	1.687	0.238
Reeves	-10.193	1.210	2.070	0.290
Optimisation	-10.578	1.338	2.070	0.290
Conditional Score	-10.471	1.240	2.042	0.267

Table 5.2: Table of results

To present these results graphically, the resulting values of $\text{logit}(\hat{p}) = \hat{\alpha} + \hat{\beta}z$ are

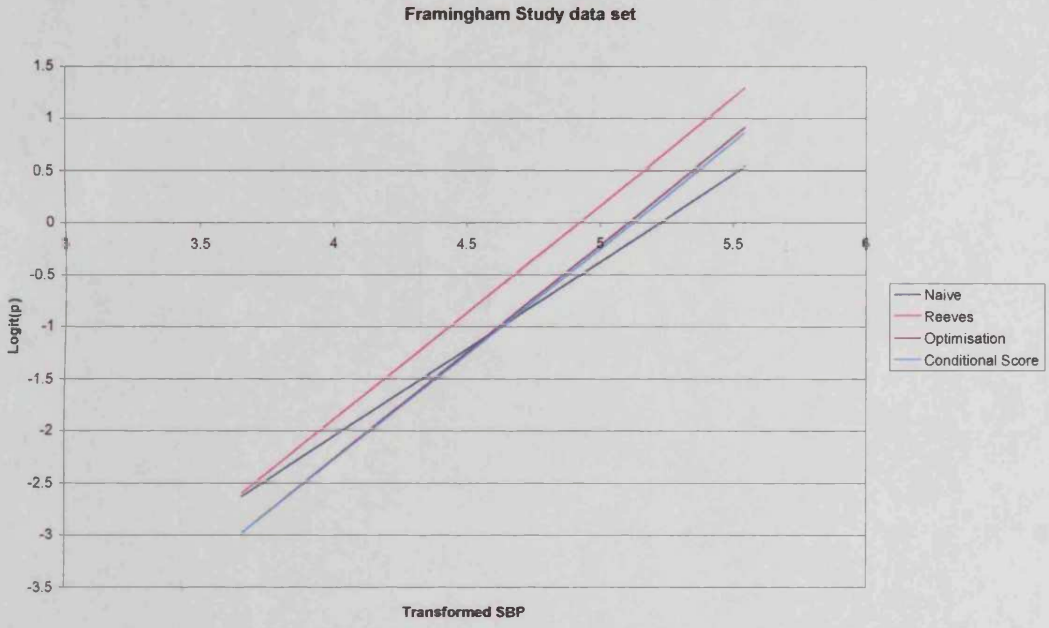


Figure 5.1: Comparison of methods for the Framingham study data set

Table 5.2 and figure 5.1 show that the Reeves, Conditional Score and Optimisation methods give approximately the same estimates of β_t, with rather larger absolute, but comparable relative, differences in $\hat{\alpha}_t$. The Conditional Score method gives a slightly lower estimate than the method by Reeves which, when as can be seen from Figure 5.1, shows values for logit(p) very different from those of the Conditional Score

method due to the different $\hat{\alpha}_1$. The ordinary logistic regression estimates, obtained by ignoring the measurement error, are negatively biased compared to those from the conditional score method and hence would give a much lower estimate of the association between systolic blood pressure and coronary heart disease.

In Chapter 4 the reliability coefficient $R=\frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}=\frac{\sigma_z^2-\sigma_v^2}{\sigma_z^2}$ was discussed and played a key role in some of the correction factor methods. In this case calculations show that the value of R is approximately 0.75.

5.3 Retinopathy example

A study was conducted to test the strong association between fasting plasma glucose (FPG) and the development of diabetic retinopathy in type II diabetics.

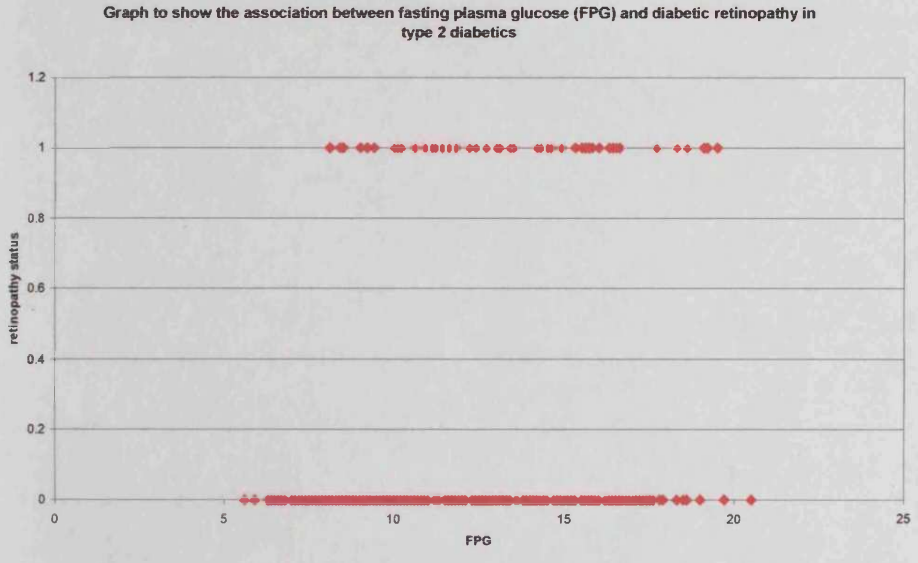


Figure 5.2: Graph to show the association between FPG and diabetic retinopathy in type 2 diabetics

Type II, or non-insulin-dependent, diabetes is diagnosed by having a fasting plasma glucose (FPG) level exceeding 7mmol/l. The disease is usually asymptomatic and patients presenting with it may have suffered from the disease for some time. As a result complications such as retinopathy may have already occurred by the time of diagnosis. It is possible that such complications could arise after diagnosis too. It is believed that the risk of developing retinopathy increases with the level of FPG. Retinopathy itself requires special equipment for diagnosis and so it is of interest to see if it is possible to assess the risk by measuring FPG. Measurement of FPG is subject to error. This has led to the question of how variable the measurement of FPG can actually be.

An investigation into the reproducibility of estimating the standard deviation of the measurement error was performed. It was assumed that the measurement error was the result of within-person variability and followed the classical measurement error model. Therefore, the results of a reproducibility study that studied the day-to-day variability of FPG levels was utilised, see Ollerton, Dunstan, Playle, Luzio, Ahmed & Owens (1999). This study consisted of a total of 193 newly diagnosed type 2 diabetics where the FPG was recorded on two consecutive days FPG_1 and FPG_2 respectively. The FPG level was recorded after the subject had fasted for 12 hours and rested for 30 minutes beforehand

	Mean	Standard Deviation
FPG_1	12.2	3.4
FPG_2	12.1	3.3
Difference	-0.1	0.9

Table 5.3: Table to show results of reproducibility study

It was then assumed that the underlying true value X is the same for both recordings and that the classical measurement error model is $Z_1 = X + V$ and $Z_2 = X + V$. Therefore $SD(Z_1 - Z_2) = \sqrt{2}\sigma_v$ and so the measurement error standard deviation is estimated to be $\hat{\sigma}_v = 0.6364$.

A main study data set consisted of a total of 351 newly diagnosed type II diabetics who were measured for FPG and retinopathy status. The FPG level was recorded after the subject had fasted for 12 hours and rested for 30 minutes beforehand, resulting in a mean of 11.772 and standard deviation of 3.393. A crude scatterplot is shown in Figure 5.2 and a smoothed plot in figure 5.3, suggesting that a logistic regression might be an appropriate model to use when examining the relationship between FPG and retinopathy status.

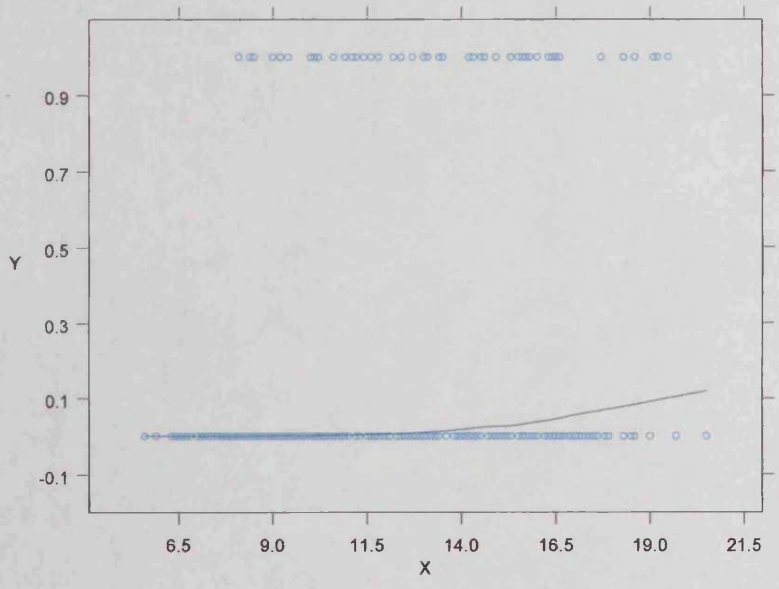


Figure 5.3: Graph to show the relationship between FPG and retinopathy status through a LOWESS plot

The same 4 methods as in section 5.2 were used to fit the logistic regression model/

To summarise the methods results the following table has been drawn.

	$\hat{\alpha}_t$	SE($\hat{\alpha}_t$)	$\hat{\beta}_t$	SE($\hat{\beta}_t$)
Naive	-4.1952	0.6478	0.1936	0.0472
Reeves	-4.2763	0.6587	0.2012	0.0490
Optimisation	-4.2914	0.6693	0.2012	0.0490
Conditional Score	-4.2910	0.6668	0.2005	0.0483

Table 5.4: Table to show method estimates for $\hat{\alpha}_t$ and $\hat{\beta}_t$ with associated standard errors

To present these results graphically, the values of logit(p) were plotted for all methods; values are shown in figure 5.4.

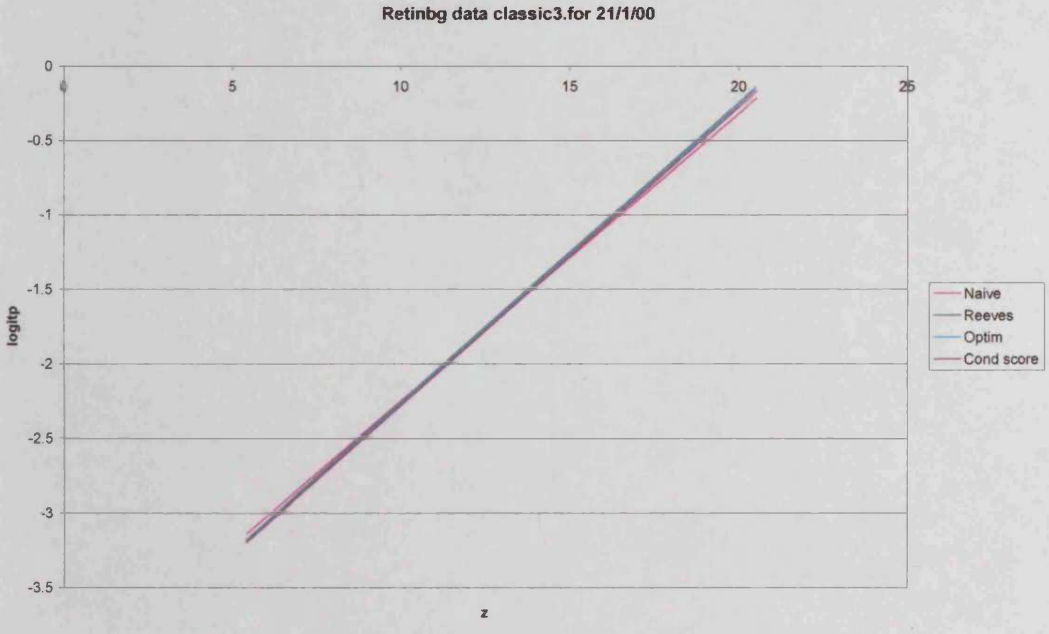


Figure 5.4: Comparison of methods for the retinopathy data set

As can be seen from this example there is very little difference between the methods. In this case the value of R is 0.965, compared to 0.75 in the Framingham study. For such a small measurement error standard deviation the correction factor is close to 1 and the method of ordinary logistic regression gives results not very different from the more sophisticated methods.

5.4 Simulation Study

In the above examples we could see that the various correction methods gave quite similar results, all of which were quite different from those produced by ordinary logistic regression. While these results are very suggestive, they do not prove that the methods are definitely better since we do not know the correct answer. To make this comparison we use a simulation study.

5.4.1 X follows a normal distribution

To compare the methods described in Chapter 4 in a number of realistic situations a simulation study was conducted according to the following model.

1) Sample size N, was chosen to be 100, 500 and 1000 respectively.
2) The true X values were assumed to follow a Normal Distribution with parameters
   $X \sim N(30,10^2)$
3) The observed Z values were generated according to a classical measurement error model, $Z = X + V$, where V was also assumed to be Normally distributed
   $V \sim N(0,\sigma_v^2)$
4) The disease status Y was generated conditional on X for the four cases

1) Case A where $\alpha_t = -3$ and $\beta_t = 0.1$, prevalence of disease 0.5

2) Case B where $\alpha_t = -4.5$ and $\beta_t = 0.1$, prevalence of disease 0.23

2) Case C where $\alpha_t = -6$ and $\beta_t = 0.1$, prevalence of disease 0.07

3) Case D where $\alpha_t = 0$ and $\beta_t = -0.1$, prevalence of disease 0.07

and $Y = 1$ with probability

$$P(Y = 1|X) = \frac{\exp(\alpha_t + \beta_t X)}{1 + \exp(\alpha_t + \beta_t X)}$$

5) The measurement error standard deviation σ_v ranged from 0 to 4, with steps of 0.5.

6) Each simulation was run 13000 times with the optimisation program running for the first 1000 runs. The optimisation method requires a prohibitively large amount of computing power and so a smaller number of runs was used in that case. This enabled the optimisation method to be included in the comparison study.

When studying the effects of measurement error in the explanatory variable the parameter of interest is the associated regression coefficient; in the simple case this corresponds to the parameter β_t. Therefore, this study looked at the effects of measurement error on β_t, that is how well each method estimates β_t as σ_v is increased. A number of scenarios have been simulated so that the following is a summary of the simulation results. Estimates will be given for the mean value of $\hat{\beta}_t$ for the 13000 simulation runs for the ordinary logistic regression, Reeves and Conditional Score methods and the mean value for 1000 simulation runs for the optimisation method. Further summary statistics are given in the form of two standard errors. The first standard error is calculated from the likelihood $\left(SE_L\left(\hat{\beta}_t\right)\right)$, i.e.

ignoring measurement error, and the second is the empirical sample standard deviation calculated from the values of $\hat{\beta}_t$ for the 13000 simulation runs $\left(SE_E\left(\hat{\beta}_t\right)\right)$. These values will be given in tabular form and the resulting confidence intervals, CI1 and CI2 respectively, will be displayed in graphical form. Further to these summaries, a coverage term, that is the probability a 95% Confidence Interval contains the true value, will also be shown. For each simulation run the odds ratio comparing the upper quartile of X to the lower quartile of X that is, $\exp\left(1.348981 * \sigma_x * \hat{\beta}_t\right)$ is calculated, and the mean and standard deviation of these values from the 13000 and 1000 simulation runs respectively are tabulated.

The results for each case and sample size within each case are displayed in the following form:

- Table displaying mean value for $\hat{\alpha}_t$ and the likelihood standard deviations for each method.
- Table displaying mean value for $\hat{\beta}_t$ and the two associated standard deviations, likelihood and empirical, for each method.
- Graph comparing the mean estimates for $\hat{\alpha}_t$ for each method
- Graph comparing the mean estimates for $\hat{\beta}_t$ for each method

The following graphs are displayed with different axis scales so that detail can be seen; comparisons must therefore be made with caution.

- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the ordinary logistic regression method
- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the Reeves method.

- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the optimisation method.
- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the conditional score method.
- Table displaying the Odds Ratio and associated standard deviation for each method.
- Table displaying the coverage of each method based on a 95% Confidence Interval.

Case A $\alpha_t = -3$ and $\beta_t = 0.1$

N=100

The results for this sample size are displayed in Figure 5.5 to Figure 5.10 and Table 5.5 to Table 5.8.

From Table 5.6 it can be seen that when there is no measurement error, that is $\sigma_v = 0$, each of the methods' expected value of $\hat{\beta}_t$ has a slight positive bias compared with the true value of β_t. For the ordinary logistic regression method, as the measurement error standard deviation is increased, the bias decreases until $\sigma_v = 2$, when it becomes negative. When $\sigma_v = 4$ this negative bias has increased so that the mean of $\hat{\beta}_t$ is 0.08679. In contrast, for the Reeves and conditional score methods, the expected values of $\hat{\beta}_t$ had a slight positive bias compared to the true value of β_t, that increases slightly as σ_v is increased. For the optimization method, the pattern of the expected values of $\hat{\beta}_t$ as σ_v is increased follows the same pattern of the ordinary logistic regression method. That is, when $\sigma_v = 0$, the expected value of $\hat{\beta}_t$ has a slight positive bias which decreases as σ_v increases and becomes negative when $\sigma_v = 2$. This negative bias then increases as σ_v is increased to 4.

Figure 5.6 shows the mean values of $\hat{\beta}_t$ for each method on the same scale so that they can be compared. From this graph it can be seen that the Reeves and conditional score methods produced approximately the same expected values for $\hat{\beta}_t$ and so the methods are comparable with respect to bias.

Figure 5.7 to Figure 5.10 display the expected values of $\hat{\beta}_t$ for each method, together with associated confidence intervals. These graphs have been set with different axes to display the detail for each individual method and so comparisons must be made carefully. In all four methods, the widths of the confidence intervals for both standard errors are approximately the same size as well as being the same width for all σ_v values. These results show that the bias referred to above is real and not merely due to sampling variation.

For each of the methods, the mean of the standard errors calculated from the likelihood are slightly smaller than the empirical standard errors. In the case of the ordinary logistic regression method, both standard errors decrease in size as σ_v increases. For the Reeves and conditional score methods, both of the standard errors increase as σ_v increases, though for all three methods the changes are small as the standard deviation of the measurement errors increases. For the optimization method, the likelihood based standard error increases as σ_v increases, whereas the empirical standard error decreases in size as σ_v increases. As with the other three methods the respective increases and decreases are very slight.

In comparison, if we take $\sigma_v = 3$, comparing the Reeves, conditional score and optimization methods, the optimization method has produced the smallest standard errors by a fraction for the likelihood-based standard errors and by a slightly larger value for the empirical standard error. However, across the three methods there is little difference between the methods when comparing the standard errors.

Table 5.7 displays the mean of the odds ratios calculated from each simulation value of $\hat{\beta}_t$ and an associated standard error. In this case, the true value for a change in value from the

25th percentile to the 75th percentile is 3.853. As the odds ratio is simply exp(β) there is no new information here, but it is the odds ratio which is used as a summary in this context, rather than the regression coefficient. The sampling distribution of the odds ratio will generally be positively skewed and so the mean must be interpreted carefully. Positive bias will be expected in general.

The positive bias in the expected values of $\hat{\beta}_t$ are reflected in the odds ratio values for each of the methods. All four methods give an odds ratio value of 4.43 when $\sigma_v = 0$, a positive bias of 0.58, which is 0.58 above the true value of the odds ratio. For the ordinary logistic regression method, the negative bias in the expected values of $\hat{\beta}_t$ when $\sigma_v = 4$ results in a reduction of 0.43 in the odds ratio. For the Reeves method, the positive bias when $\sigma_v = 4$ results in a bias of 0.71. Therefore, the Reeves method produced a more positively biased estimate of the odds ratio than the negative bias associated with the ordinary logistic regression. This is a result of the non-linear relationship of the exponential in the odds ratio. In this case, when comparing the methods, the optimization method produced the least biased estimate of the odds ratio for all σ_v values.

When considering the associated odds ratio standard error, the standard errors of the odds ratios using all four methods follow the same pattern as those for the expected values of $\hat{\beta}_t$. In this case, the optimization method produced the smallest standard errors when comparing the Reeves, conditional score and optimization methods.

Table 5.8 displays the coverage terms as σ_v increases for all four methods. The coverage displays the probability a 95% confidence interval contains the true value. So for the

Reeves, optimization and conditional score methods the coverage stays at approximately 0.95. For the ordinary logistic regression method this probability has reduced to approximately 0.88 which reflects the decrease in the standard errors and the bias associated with the expected values of $\hat{\beta}_t$.

A further pattern that can be seen from Graphs 5.8 to 5.10, is an oscillating effect in the mean estimates of β_t as the measurement error standard deviation is increased. This effect only seems to appear for the correction methods and not for the ordinary logistic regression method. The reason for this effect does not seem obvious and though it is a curious pattern the reason behind it remains unresolved and could be the subject of future work.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.1332	0.85589	-3.13316	0.85592	-3.1146	0.85275	-3.13316	0.85589
0.5	-3.1219	0.85333	-3.12998	0.8549	-3.0884	0.85268	-3.13369	0.85665
1	-3.0877	0.8475	-3.11941	0.85364	-3.091	0.85732	-3.13442	0.86055
1.5	-3.039	0.83929	-3.10854	0.85286	-3.1038	0.86697	-3.14304	0.86846
2	-2.9648	0.82672	-3.0833	0.85028	-2.9937	0.86342	-3.14553	0.87806
2.5	-2.9174	0.81695	-3.09575	0.85337	-2.9622	0.87414	-3.1985	0.89841
3	-2.8067	0.7998	-3.04769	0.85065	-2.8509	0.88124	-3.19928	0.91627
3.5	-2.7207	0.78387	-3.02876	0.85133	-2.7949	0.88978	-3.24505	0.94452
4	-2.6001	0.7634	-2.97284	0.84841				

Table 5.5: Case A Sample Size 100 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10443	0.02754	0.02858	0.10443	0.02754	0.02858	0.10358	0.02742	0.02896	0.10443	0.02754	0.02858
0.5	0.1041	0.02746	0.0283	0.10443	0.02755	0.02842	0.10325	0.02748	0.02774	0.10445	0.02755	0.02843
1	0.10295	0.02726	0.02838	0.10426	0.02763	0.02887	0.10301	0.02757	0.02856	0.10431	0.0276	0.02891
1.5	0.10129	0.02697	0.02778	0.10419	0.0278	0.02888	0.10327	0.02793	0.02779	0.10432	0.02774	0.02897
2	0.09884	0.02655	0.0271	0.10387	0.028	0.02901	0.09991	0.02785	0.02787	0.10409	0.0279	0.02917
2.5	0.09717	0.0262	0.02695	0.10495	0.02846	0.02996	0.09899	0.02822	0.02763	0.10529	0.02833	0.03022
3	0.0936	0.02562	0.02634	0.10439	0.0288	0.03055	0.0957	0.02847	0.0291	0.10489	0.02865	0.03097
3.5	0.09063	0.02506	0.02586	0.10489	0.02932	0.03159	0.09393	0.02881	0.02763	0.10561	0.02921	0.03232
4	0.08679	0.02437	0.02502	0.10467	0.0298	0.03238						

Table 5.6:Case A Sample Size 100 $\hat{\beta}_t$ summary results

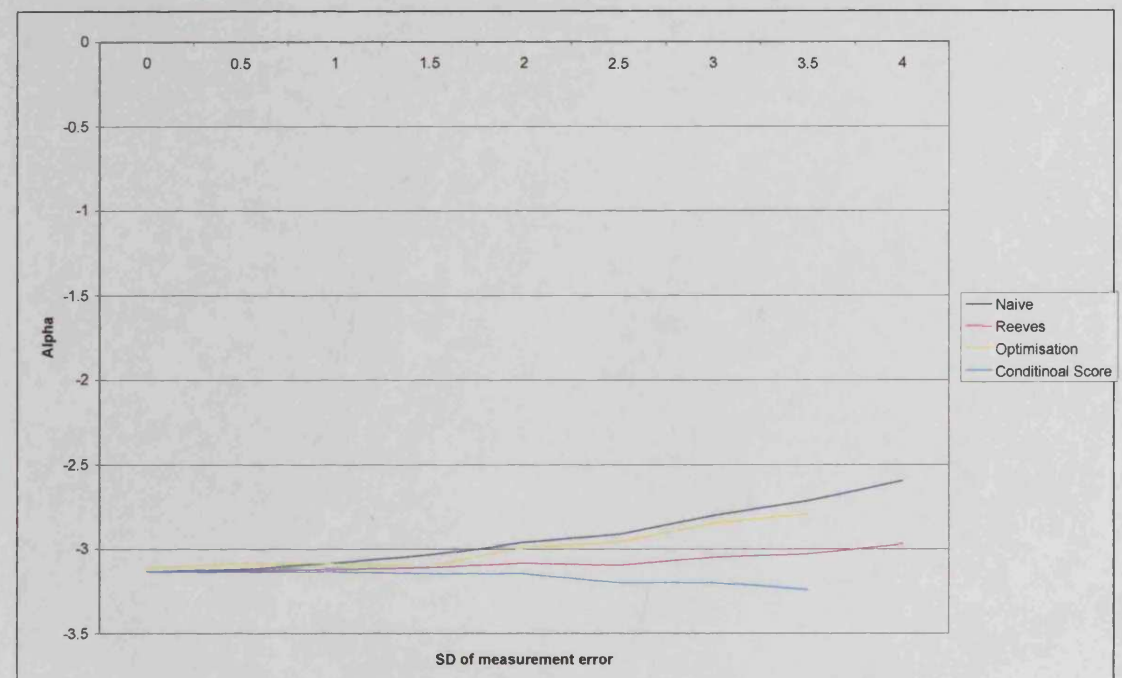


Figure 5.5: Case A Sample size 100 $\hat{\alpha}_t$ method comparison

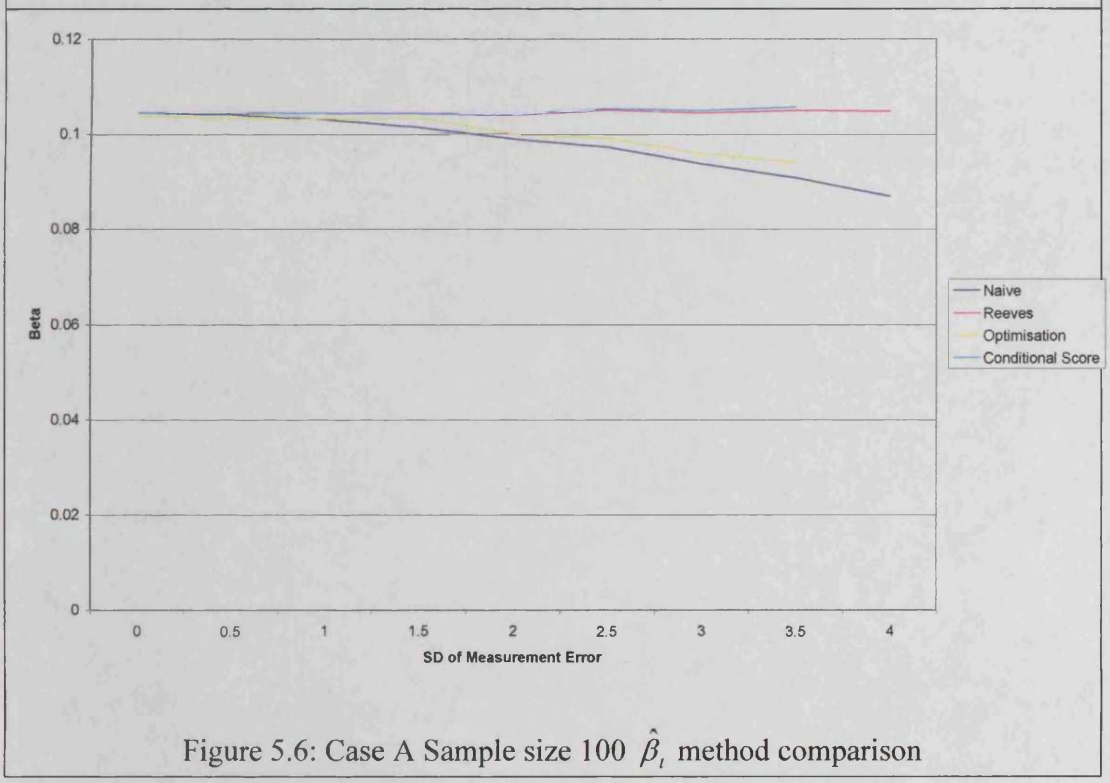


Figure 5.6: Case A Sample size 100 $\hat{\beta}_t$ method comparison

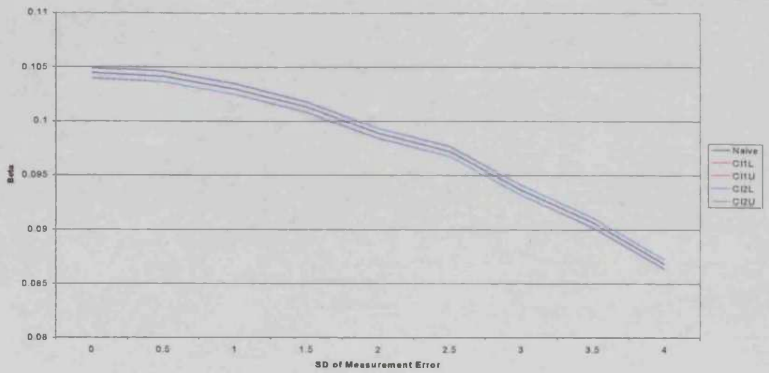


Figure 5.7: Case A Sample size 100 $\hat{\beta}_t$ Olr Confidence Interval comparison

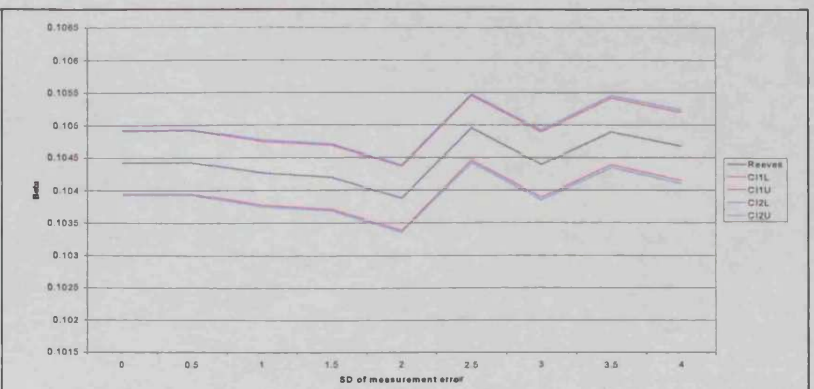


Figure 5.8: Case A Sample size 100 $\hat{\beta}_t$ Reeves Confidence Interval comparison

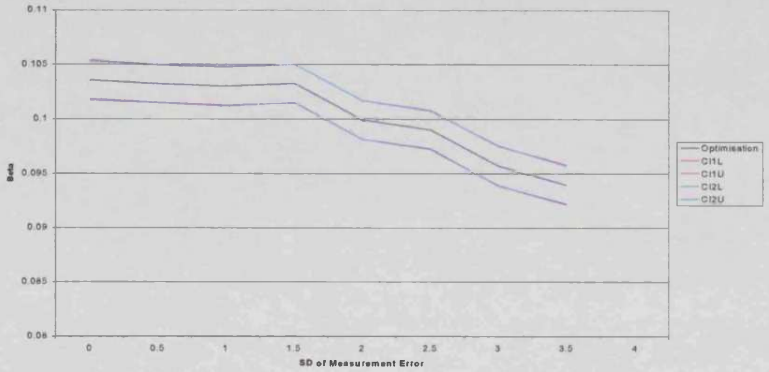


Figure 5.9: Case A Sample size 100 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

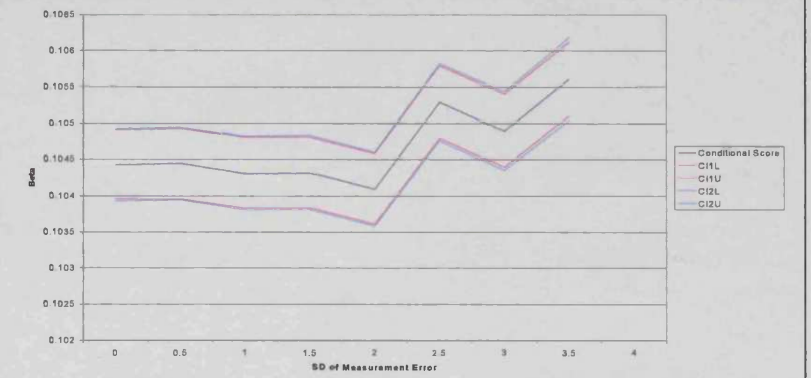


Figure 5.10: Case A Sample size 100 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	4.43018	2.00318	4.43018	2.00317	4.39788	2.09017	4.43018	2.00318
0.5	4.40487	1.99057	4.42796	2.01407	4.3425	1.95433	4.42904	2.01574
1	4.33823	2.00582	4.42924	2.11729	4.33913	1.89137	4.43364	2.12548
1.5	4.2283	1.88244	4.42719	2.10241	4.33552	1.81306	4.43745	2.11822
2	4.07507	1.75137	4.4123	2.11427	4.15261	1.86435	4.4301	2.14123
2.5	3.97966	1.67926	4.50271	2.24462	4.09426	1.76746	4.53274	2.32262
3	3.77932	1.54303	4.48634	2.31718	3.95038	1.8163	4.53146	2.40503
3.5	3.62222	1.44582	4.55326	2.5009	3.82859	1.71497	4.63113	2.83601
4	3.42349	1.29537	4.56431	2.57996				

Table 5.7:Case A Sample Size 100 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	95.47	95.45	95.46	95.45	94.3	94.1	95.46	95.45
0.5	95.69	95.42	95.76	95.45	95.2	95.3	95.55	95.47
1	95.16	95.23	95.26	95.3	95.1	94.6	95.26	95.22
1.5	95.45	95.45	95.68	95.75	95.8	96.1	95.42	95.36
2	94.81	94.68	95.28	95.27	96	95.1	95.44	95.18
2.5	94.16	93.98	95.25	95.27	96.1	95.9	95.3	94.95
3	92.75	92.69	94.52	95.04	94.3	93.7	95.42	95.12
3.5	91.62	91.1	94.69	94.96	94.4	94.8	95.13	94.78
4	88.58	88.25	93.98	94.28				

Table 5.8:Case A Sample Size 100 Coverage summary results

N=500

The results for this sample size are displayed in Figure 5.11 to Figure 5.16 and Table 5.9 to Table 5.12.

With the increase in sample size, Table 5.10 shows that the associated biases and standard errors of the expected values of $\hat{\beta}_t$ have reduced for all four methods. The same patterns of positive and negative bias, increase and decrease in standard errors and confidence intervals patterns (Figure 5.13 to Figure 5.16), observed for n=100, have not changed with the increase in sample size.

Figure 5.12 graphically compares the four methods for the expected values of $\hat{\beta}_t$. In this case, the Reeves method produced the most consistent and least biased expected values of $\hat{\beta}_t$ as σ_v increased. However, for $\sigma_v < 1.5$, the optimization method produced the least biased results. The conditional score method produced only slightly more positively biased results in comparison with the Reeves method. These results are reflected in Table 5.11 with the mean of the estimated odds ratio values. For the Reeves method, the means of the estimated odds ratio values are approximately between 0.09 and 0.1 above the true value whereas for ordinary logistic regression when $\sigma_v = 4$, the method under-estimates the odds ratio by approximately 0.7.

With respect to the coverage terms in Table 5.12, the ordinary logistic regression method only provided a 65% coverage compared to a 93% coverage for the Reeves method. So with the increased sample size, the ordinary logistic regression method

has had a significant reduction in the probability that the 95% confidence interval contains the true value of β_t. In this case, the conditional score method provided the best coverage with approximately 94%.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.0197	0.3694	-3.01972	0.36942	-3.01105	0.36896	-3.01972	0.3694
0.5	-3.0186	0.36924	-3.02619	0.36988	-3.0291	0.37083	-3.02923	0.37053
1	-2.9859	0.36682	-3.01568	0.3693	-2.9927	0.37126	-3.02794	0.37191
1.5	-2.9472	0.36362	-3.01253	0.36911	-2.9762	0.3737	-3.04086	0.37501
2	-2.886	0.35904	-2.9979	0.36861	-2.8937	0.37422	-3.04954	0.3791
2.5	-2.8097	0.35319	-2.97629	0.36777	-2.8495	0.3769	-3.05922	0.38433
3	-2.7263	0.34678	-2.95313	0.36723	-2.7704	0.37956	-3.07663	0.39129
3.5	-2.6359	0.33975	-2.92571	0.36675	-2.6895	0.3822	-3.09976	0.39998
4	-2.539	0.3321	-2.89169	0.36618	-2.60643	0.38526	-3.12618	0.41024

Table 5.9: Case A Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10068	0.01187	0.01189	0.10068	0.01187	0.01189	0.10041	0.01186	0.01154	0.10068	0.01187	0.01189
0.5	0.10056	0.01186	0.0118	0.10086	0.0119	0.01185	0.10088	0.01191	0.01124	0.10087	0.01189	0.01185
1	0.0995	0.01178	0.01193	0.10068	0.01193	0.01212	0.09991	0.01194	0.01224	0.10073	0.01191	0.01213
1.5	0.09824	0.01167	0.01183	0.10087	0.01201	0.01225	0.09938	0.01202	0.01187	0.10098	0.01197	0.01228
2	0.09618	0.01151	0.01159	0.10076	0.0121	0.01231	0.09688	0.01205	0.01188	0.10094	0.01203	0.01236
2.5	0.09371	0.01131	0.01138	0.10068	0.01222	0.01249	0.09558	0.01216	0.01266	0.10095	0.01212	0.01257
3	0.09089	0.01109	0.01113	0.10063	0.01237	0.01271	0.09341	0.01226	0.01192	0.10102	0.01224	0.01283
3.5	0.08785	0.01085	0.011	0.10067	0.01255	0.01313	0.09071	0.01236	0.01197	0.10118	0.01239	0.01329
4	0.08459	0.01058	0.01061	0.10072	0.01275	0.01332	0.08834	0.01248	0.01165	0.10138	0.01257	0.01354

Table 5.10:Case A Sample Size 500 $\hat{\beta}_t$ summary results

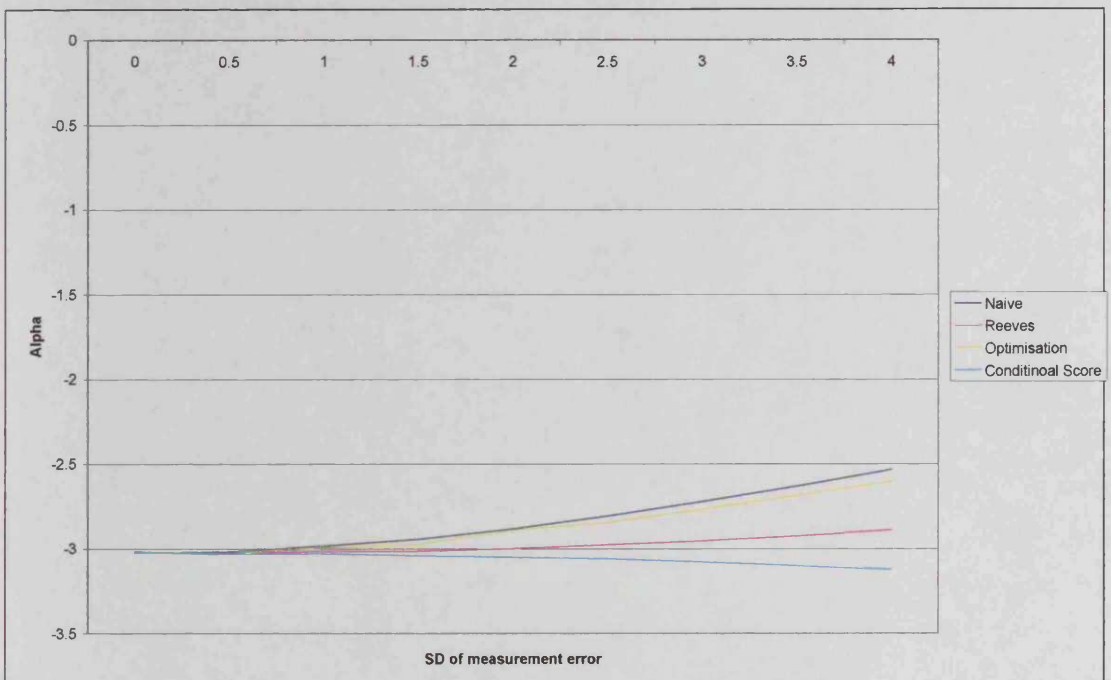


Figure 5.11: Case A Sample size 500 $\hat{\alpha}_i$ method comparison

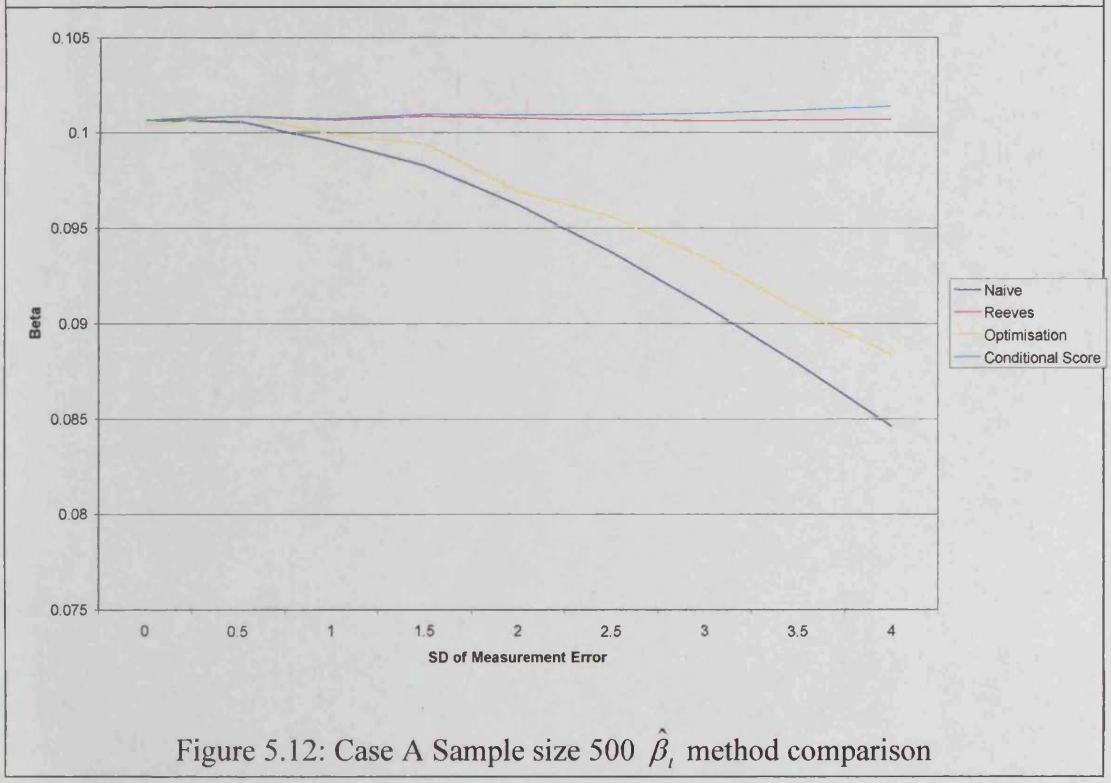


Figure 5.12: Case A Sample size 500 $\hat{\beta}_i$ method comparison

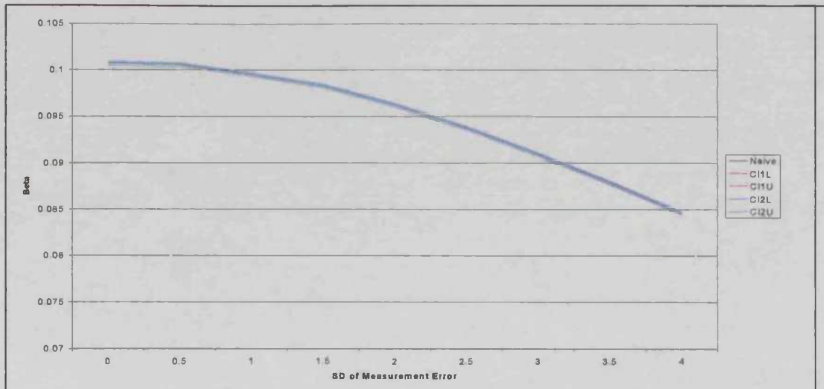


Figure 5.13: Case A Sample size 500 $\hat{\beta}_t$ Olr Confidence Interval comparison

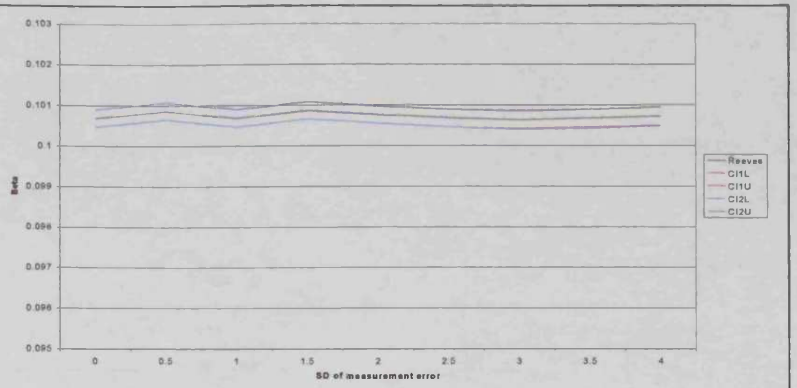


Figure 5.14: Case A Sample size 500 $\hat{\beta}_t$ Reeves Confidence Interval comparison

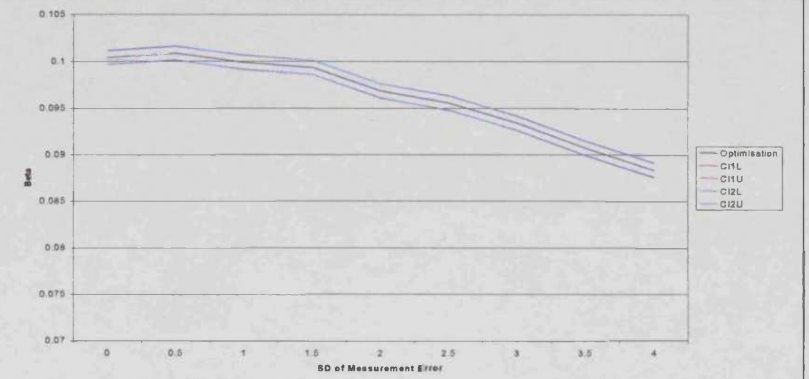


Figure 5.15: Case A Sample size 500 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

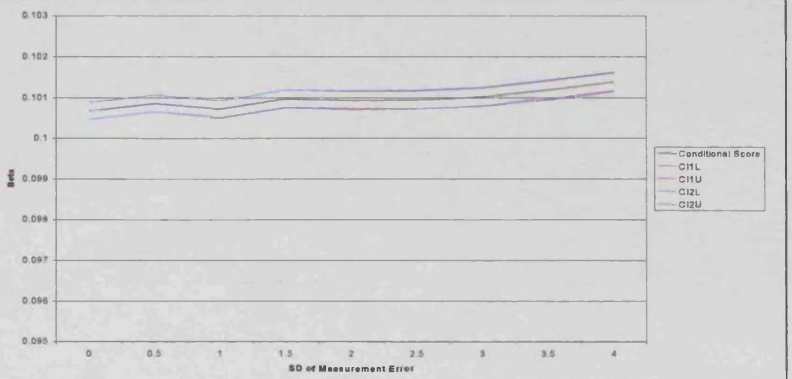


Figure 5.16: Case A Sample size 500 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.93997	0.64782	3.93995	0.64789	3.92212	0.61964	3.93997	0.64782
0.5	3.93271	0.63895	3.94904	0.64426	3.94488	0.60815	3.94968	0.64451
1	3.87799	0.64046	3.94213	0.6619	3.90258	0.66411	3.94464	0.66296
1.5	3.81195	0.62383	3.95351	0.67118	3.87152	0.63647	3.95918	0.67392
2	3.70548	0.59413	3.94817	0.67541	3.74344	0.62308	3.95808	0.6799
2.5	3.58217	0.56138	3.94523	0.68218	3.68486	0.65135	3.96023	0.68907
3	3.44697	0.52834	3.94487	0.69573	3.57204	0.58876	3.9665	0.70651
3.5	3.30732	0.50076	3.95079	0.7217	3.44496	0.57179	3.97993	0.73618
4	3.16295	0.46173	3.9555	0.73559	3.33424	0.54052	3.99296	0.75679

Table 5.11:Case A Sample Size 500 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	94.88	94.92	94.88	94.92	95.8	96.2	94.88	94.92
0.5	95.25	95.28	95.25	95.37	95.8	96.2	95.48	95.25
1	95.13	94.75	95.17	95.02	94.9	95.5	94.95	95.09
1.5	94.68	94.48	94.99	95.12	95.6	95.6	94.88	94.77
2	93.2	92.96	94.79	94.91	94.5	93.9	94.6	94.55
2.5	90.12	89.96	94.23	94.75	91.3	91.4	94.48	94.43
3	85.58	84.6	94.27	94.68	90.3	91	94.17	94.02
3.5	78.75	77.32	93.61	94.43	86.2	88.2	93.32	93.39
4	68.08	65.82	92.21	93.92	82.1	83.7	93.67	93.44

Table 5.12:Case A Sample Size 500 Coverage summary results

N=1000

The results for this sample size are displayed in Figure 5.17 to Figure 5.22 and Table 5.13 to Table 5.16.

For the increase in sample size from n=500 to 1000, the bias in the expected values of $\hat{\beta}_t$ (Table 5.14 and Figure 5.18) stays at relatively the same level for each of the methods; any change is in the fourth decimal place. The increase in the sample size is reflected in the standard errors. For each of the methods, the standard error from the likelihood is slightly smaller than the empirical standard error, but both have reduced in comparison with the same results from n=500. For example, for the Reeves method for $\sigma_v = 4$ the standard errors have reduced from (0.01275, 0.01332) to (0.00896, 0.0094) respectively.

With respect to the coverage terms, Table 5.16, the probability that the true value of β_t lies in the 95% confidence interval for the ordinary logistic regression method has now significantly reduced to approximately 42%, showing the effect of the negative bias and small standard error. For the Reeves method, the coverage has only reduced to approximately 94%.

As previously observed, the Reeves method produced the least biased expected value for $\hat{\beta}_t$ and as a result the mean estimate of the odds ratio. However, there is very little difference between the Reeves and conditional score methods when considering the bias in the estimates and the associated standard errors of those estimates.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.0158	0.26044	-3.01582	0.26045	-3.00587	0.2602	-3.01582	0.26044
0.5	-3.0037	0.25984	-3.01122	0.26028	-3.007	0.2608	-3.01418	0.26074
1	-2.9741	0.25835	-3.00366	0.26007	-2.9851	0.26126	-3.01563	0.26189
1.5	-2.9308	0.25601	-2.99556	0.25984	-2.9406	0.26232	-3.02308	0.26392
2	-2.8691	0.25269	-2.97992	0.25936	-2.8985	0.26395	-3.03009	0.26659
2.5	-2.8005	0.24894	-2.966	0.25913	-2.8454	0.26609	-3.04725	0.27063
3	-2.7124	0.24423	-2.93736	0.25849	-2.7663	0.26722	-3.05757	0.27509
3.5	-2.6242	0.23931	-2.91168	0.25814	-2.6886	0.26934	-3.0812	0.28103
4	-2.52	0.23374	-2.86916	0.25747	-2.58453	0.2707	-3.09658	0.28767

Table 5.13: Case A Sample Size 1000 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10051	0.00836	0.00847	0.10051	0.00837	0.00847	0.10026	0.00836	0.00837	0.10051	0.00836	0.00847
0.5	0.10009	0.00834	0.0084	0.10039	0.00837	0.00844	0.10021	0.00837	0.00866	0.1004	0.00837	0.00844
1	0.09916	0.00829	0.00829	0.10033	0.0084	0.00842	0.09946	0.00839	0.00794	0.10038	0.00839	0.00843
1.5	0.09768	0.00821	0.00826	0.10027	0.00845	0.00855	0.09829	0.00844	0.0082	0.10037	0.00842	0.00857
2	0.09565	0.0081	0.00803	0.10017	0.00851	0.00853	0.09713	0.0085	0.00854	0.10034	0.00846	0.00856
2.5	0.0934	0.00797	0.00797	0.1003	0.0086	0.00874	0.09558	0.00858	0.00833	0.10057	0.00853	0.0088
3	0.09043	0.00781	0.00787	0.10004	0.0087	0.00897	0.09313	0.00862	0.00809	0.10041	0.0086	0.00905
3.5	0.08744	0.00764	0.00769	0.10008	0.00883	0.00915	0.09074	0.00871	0.00838	0.10059	0.00871	0.00926
4	0.08402	0.00745	0.00752	0.09989	0.00896	0.0094	0.0878	0.00877	0.00839	0.10052	0.00882	0.00955

Table 5.14:Case A Sample Size 1000 $\hat{\beta}_t$ summary results

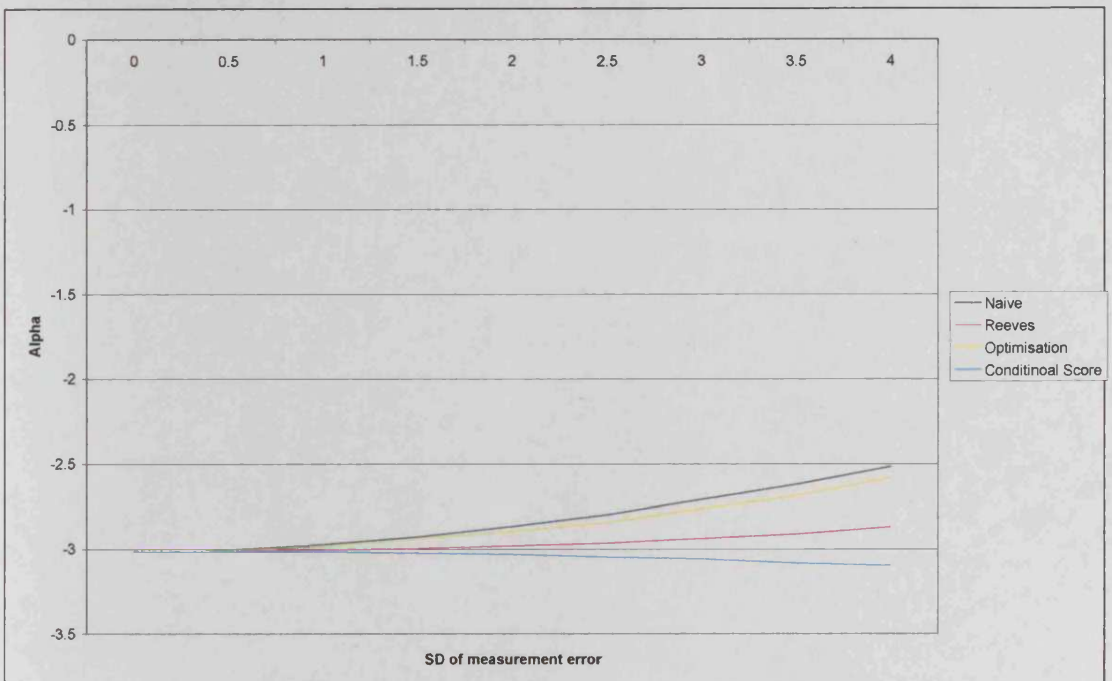


Figure 5.17: Case A Sample size 1000 $\hat{\alpha}_t$ method comparison

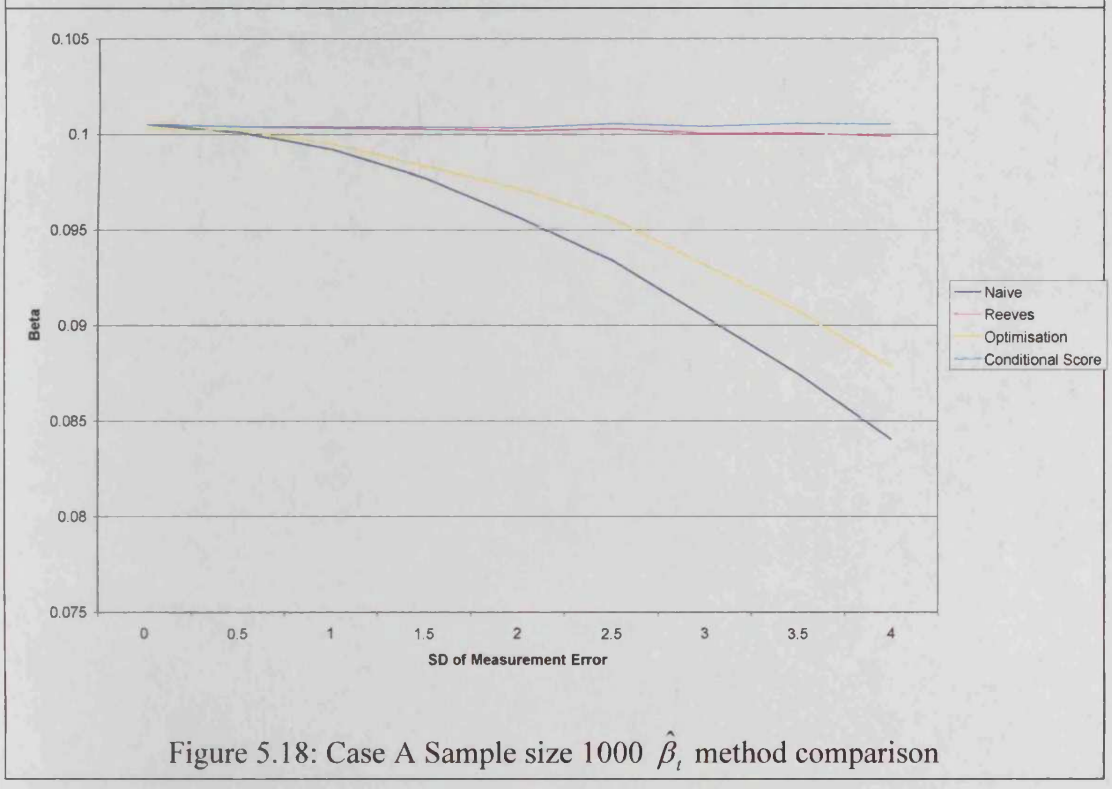


Figure 5.18: Case A Sample size 1000 $\hat{\beta}_t$ method comparison

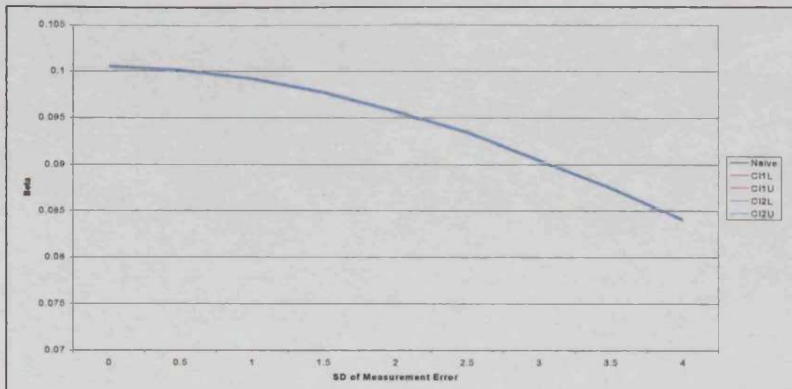


Figure 5.19: Case A Sample size 1000 $\hat{\beta}_t$ Olr Confidence Interval comparison

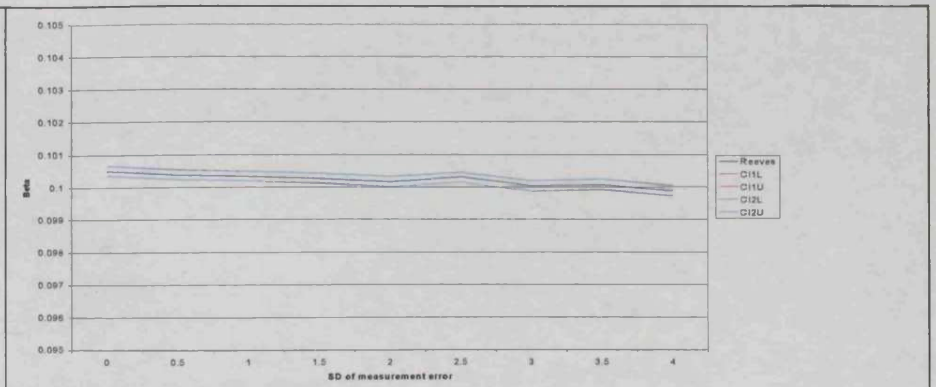


Figure 5.20: Case A Sample size 1000 $\hat{\beta}_t$ Reeves Confidence Interval comparison

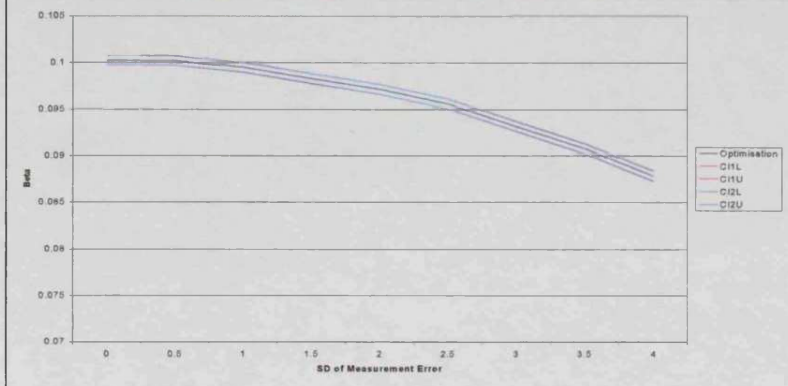


Figure 5.21: Case A Sample size 1000 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

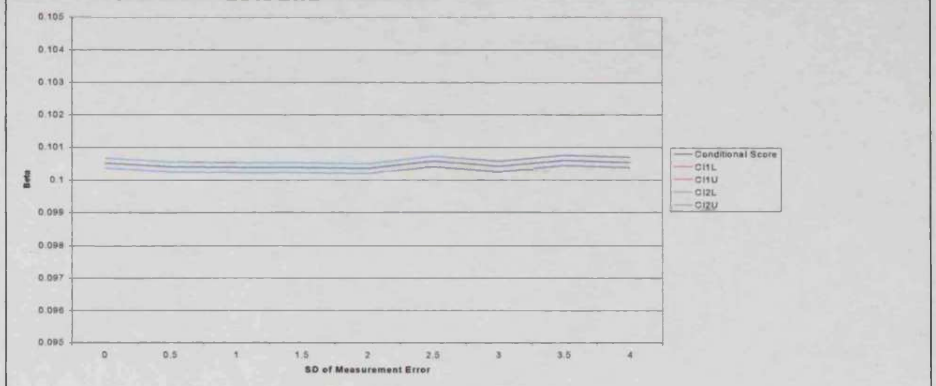


Figure 5.22: Case A Sample size 1000 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.90559	0.45193	3.9056	0.45189	3.89209	0.44996	3.90559	0.45193
0.5	3.88334	0.44641	3.89906	0.44996	3.89096	0.46295	3.89964	0.45017
1	3.83424	0.43329	3.89607	0.44722	3.84763	0.41646	3.89843	0.44791
1.5	3.75793	0.42348	3.89368	0.45448	3.78868	0.42331	3.89896	0.45609
2	3.65548	0.40007	3.8882	0.45248	3.73184	0.43892	3.89739	0.4553
2.5	3.54592	0.38543	3.89624	0.4655	3.65336	0.41197	3.91065	0.47042
3	3.40603	0.36616	3.88405	0.47812	3.53373	0.39273	3.90411	0.48505
3.5	3.27059	0.34292	3.88764	0.48771	3.42279	0.39022	3.9149	0.49733
4	3.12242	0.32018	3.87885	0.50037	3.28996	0.37704	3.91339	0.51352

Table 5.15:Case A Sample Size 1000 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	95.1	94.94	95.1	94.94	95.6	95.2	95.1	94.94
0.5	94.77	94.86	94.77	94.95	94.3	94.2	94.89	94.85
1	94.27	94.41	94.62	94.7	96.2	95.6	94.65	94.98
1.5	93.86	93.8	94.8	94.94	94.4	94.8	94.48	94.87
2	91.28	91.05	94.7	95.02	93.1	93.8	94.88	95.01
2.5	85.78	84.93	94.09	94.75	90.9	90.9	94.58	94.6
3	76.13	74.82	93.12	94.32	85.6	86.8	93.95	94.19
3.5	62.22	59.63	92.15	94.2	78.5	80.4	93.45	93.48
4	45.14	41.98	90.37	94.05	80	80	93	93.62

Table 5.16:Case A Sample Size 1000 Coverage summary results

Conclusion for Case A

In conclusion for case A, as the sample size was increased so the bias in the expected values of $\hat{\beta}_t$ reduced and the associated standard errors decreased in size for all four methods.

For all the sample sizes, the ordinary logistic regression method estimates of the model parameters are severely affected by increased values for σ_v. As a result, the mean of the odds ratio estimates and associated standard deviation reflect the effect of changes in the measurement error standard deviation, which if this were not taken into account could have an impact on any study results.

When considering which correction method to use when there is measurement error in the explanatory variable, both the Reeves and Conditional Score methods' expected values of the model parameters were positively biased for all sample sizes. When the measurement error standard deviation is small the optimisation method also proved to produce expected values with only a small positive bias but its computational demands make it less attractive for routine use.

Case B $\alpha_t = -4.5$ and $\beta_t = 0.1$

For this case the probability of an event, with $Y = 1$, is reduced, from 0.5 on average to 0.23, and so we might expect there to be less information for estimating β_t. This might lead to greater bias and larger standard errors. However, the value of β_t has not changed.

N=100

The results for this sample size are displayed in Figure 5.23 to Figure 5.28 and Table 5.17 to Table 5.20.

Table 5.18 shows the expected values for $\hat{\beta}_t$ for each of the methods and the two standard errors as before.

In terms of bias, the decrease in the prevalence of the disease has generally increased the bias in the mean of $\hat{\beta}_t$ for each of the methods. As was observed for Case A, the same pattern of positive bias for small σ_v followed by a negative bias as σ_v was increased was observed for the ordinary logistic regression and optimization methods. For the Reeves and conditional score methods, the expected values are slightly more positively biased than previously observed, and this bias increased with σ_v. Using Figure 5.24 to compare the methods, the least biased results are produced by the Reeves and conditional score methods when the measurement error standard deviation is large.

The same patterns of change of the standard errors as σ_v was increased, are observed for case B as they were observed for case A. However, the reduction in the prevalence of disease is more reflected in the size of the standard errors of the estimates than the associated bias. For all four methods, the standard errors have increased compared to Case A. The patterns of the confidence intervals shown in Figure 5.25 to Figure 5.28 are also as previously as seen.

The mean of the odds ratio estimates displayed in Table 5.19 have also increased in line with the increased bias in the expected values of $\hat{\beta}_t$. In this case, when there is no measurement error, all four methods over-estimate the odds ratio by approximately 0.73. For the Reeves method, the mean of the odds ratio estimates increase as σ_v increases, so that when $\sigma_v = 4$, the mean odds ratio estimate is almost 1.15 above the true value. The associated standard deviation is also very large in comparison with the mean of the odds ratio estimate. For the ordinary logistic regression method, the odds ratio is only 0.27 less than the true value and the associated standard error is much smaller. For the optimization and conditional score methods, neither method would work in a simulation environment when the measurement error standard deviation was large. Although the negative bias in the ordinary logistic regression expected values of $\hat{\beta}_t$ was greater than the positive bias in the Reeves method, the bias for estimating the odds ratio was actually less for the ordinary logistic regression method.

When considering the coverage terms, the ordinary logistic regression method produced an approximately 90% coverage when $\sigma_v = 4$ with the Reeves method producing approximately 94% coverage, both are in-line with the results from case A.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-4.6938	1.13232	-4.69375	1.13225	-4.7412	1.14074	-4.69375	1.13232
0.5	-4.6642	1.12693	-4.67228	1.12802	-4.6297	1.12613	-4.67677	1.13129
1	-4.667	1.12551	-4.69901	1.13119	-4.7016	1.14159	-4.71769	1.14264
1.5	-4.5945	1.11104	-4.66441	1.12372	-4.6627	1.14807	-4.70643	1.14894
2	-4.5159	1.09536	-4.63519	1.11785	-4.5606	1.14306	-4.71081	1.16228
2.5	-4.4589	1.08266	-4.63818	1.11771	-4.5057	1.15773	-4.76259	1.18929
3	-4.3641	1.06403	-4.60789	1.11417	-4.4214	1.16633	-4.79282	1.21827
3.5	-4.2756	1.04504	-4.58885	1.11289			-4.852	1.25971
4	-4.1497	1.01943	-4.52798	1.10594				

Table 5.17: Case B Sample Size 100 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10489	0.03204	0.03315	0.10489	0.03204	0.03315	0.10672	0.03232	0.03557	0.10489	0.03204	0.03315
0.5	0.10389	0.03188	0.03339	0.10423	0.03197	0.03353	0.1028	0.03193	0.03221	0.10423	0.03198	0.03354
1	0.10393	0.03181	0.03316	0.10528	0.03218	0.03376	0.10493	0.03226	0.03352	0.10529	0.03219	0.03378
1.5	0.10183	0.03139	0.0324	0.10479	0.03223	0.03371	0.10367	0.03254	0.03305	0.10483	0.03225	0.03375
2	0.09943	0.03093	0.03206	0.10458	0.0324	0.03439	0.10075	0.03247	0.03333	0.10464	0.03244	0.03446
2.5	0.09769	0.03053	0.03181	0.10564	0.03282	0.0355	0.09872	0.0329	0.03331	0.10576	0.03292	0.0357
3	0.0947	0.02997	0.03134	0.10583	0.03323	0.03655	0.09673	0.03328	0.03386	0.10603	0.0334	0.03683
3.5	0.09192	0.02939	0.03049	0.10675	0.03379	0.03757				0.10707	0.03413	0.0381
4	0.08806	0.0286	0.02969	0.10663	0.03422	0.03889						

Table 5.18:Case B Sample Size 100 $\hat{\beta}_t$ summary results

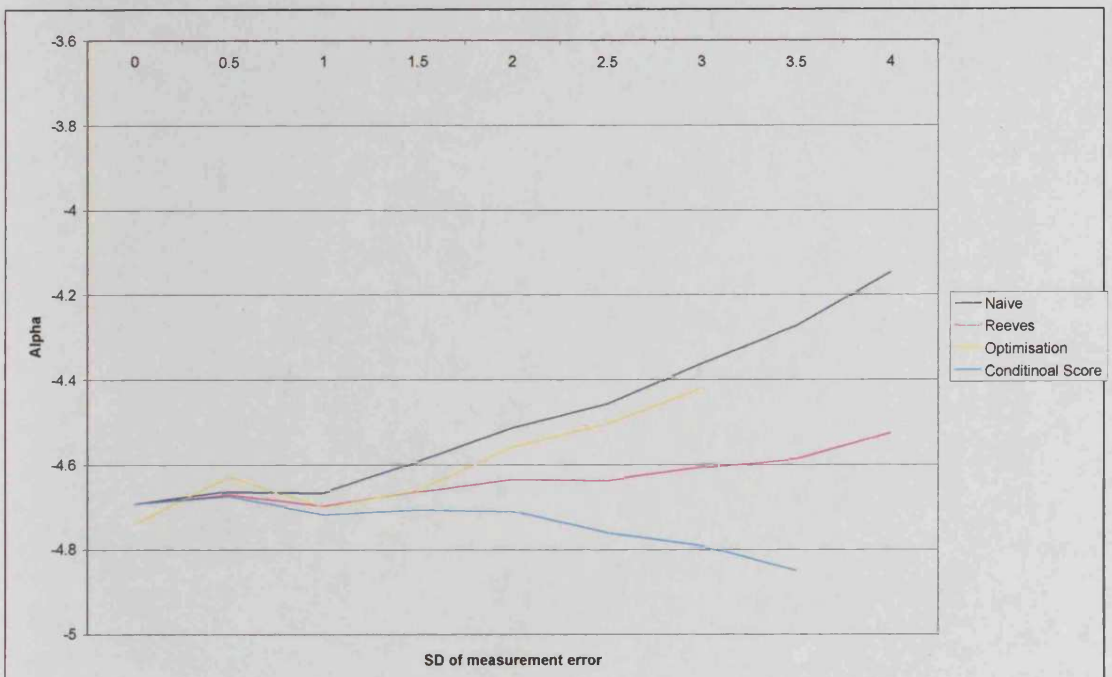


Figure 5.23: Case B Sample size 100 $\hat{\alpha}_t$ method comparison

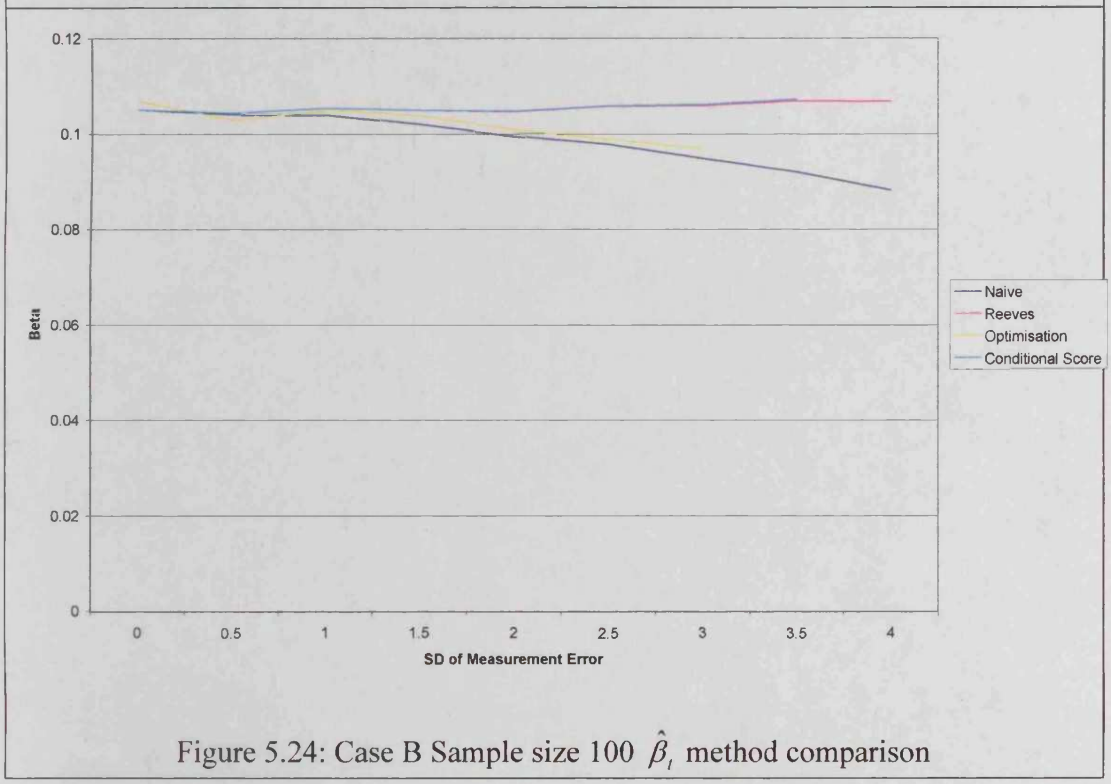


Figure 5.24: Case B Sample size 100 $\hat{\beta}_t$ method comparison

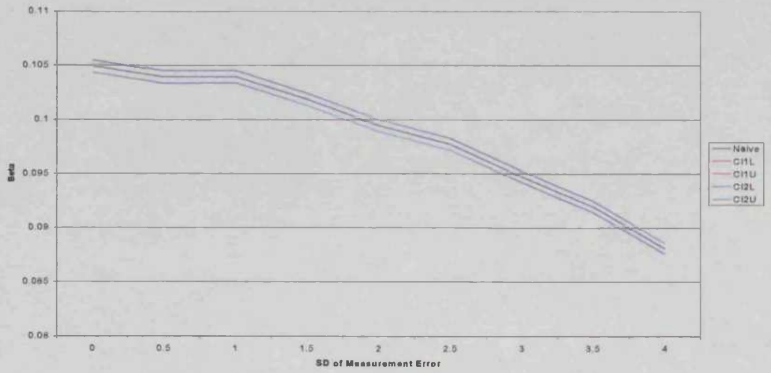


Figure 5.25: Case B Sample size 100 $\hat{\beta}_t$ Olr Confidence Interval comparison

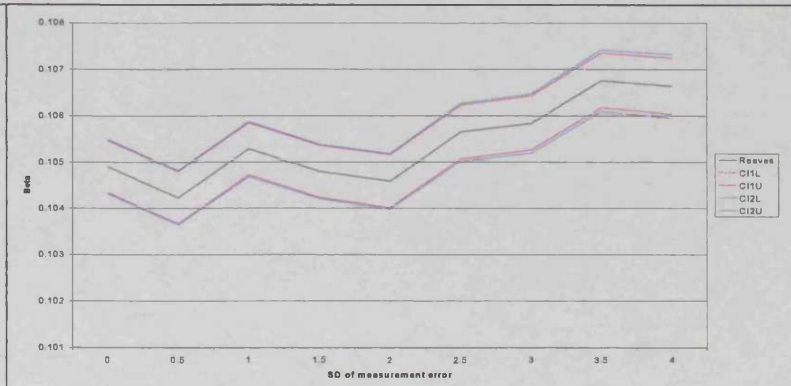


Figure 5.26: Case B Sample size 100 $\hat{\beta}_t$ Reeves Confidence Interval comparison

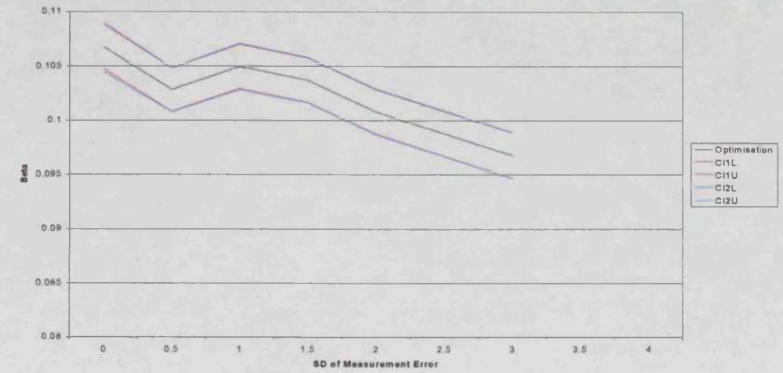


Figure 5.27: Case B Sample size 100 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

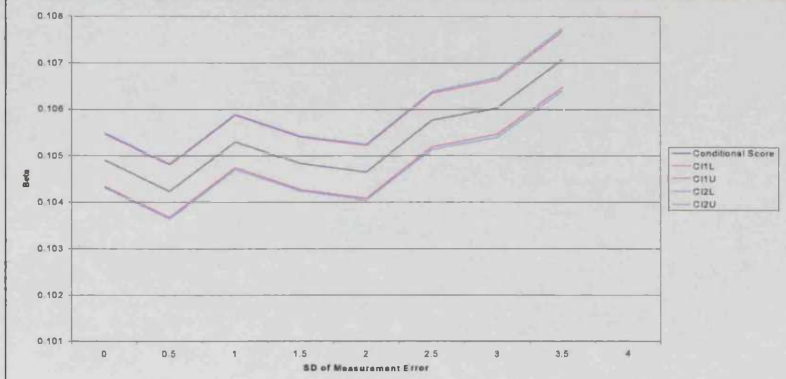


Figure 5.28: Case B Sample size 100 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	4.58553	2.50482	4.58552	2.50484	4.7873	2.92826	4.58553	2.50482
0.5	4.5366	2.56789	4.56231	2.60343	4.44282	2.49335	4.56258	2.60368
1	4.53687	2.64284	4.64189	2.81494	4.60226	2.54384	4.64336	2.80736
1.5	4.3797	2.35847	4.603	2.66705	4.54673	3.04868	4.60736	2.68606
2	4.23432	2.28556	4.62098	2.86069	4.3323	2.24532	4.62874	2.89508
2.5	4.13368	2.42124	4.75622	4.77646	4.26082	2.91459	4.79225	6.62609
3	3.95022	2.01794	4.78428	3.28147	4.14385	2.42597	4.81583	3.50823
3.5	3.78588	1.87932	4.9032	3.82599			4.95826	4.01171
4	3.57515	1.70631	4.97943	5.4307				

Table 5.19:Case B Sample Size 100 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	95.44	95.53	95.43	95.53	94.2	93.8	95.43	95.53
0.5	95.33	95.53	95.32	95.48	95.7	95.7	95.35	95.33
1	95.53	95.62	95.57	95.6	95.3	95.2	95.58	95.47
1.5	95.57	95.57	95.77	95.73	96.4	96.4	95.55	95.76
2	95.48	95.36	95.57	95.63	95.5	95.4	95.73	95.55
2.5	93.98	94.23	94.69	94.95	95.6	95.3	95.33	95.32
3	93.38	93.61	94.68	94.71	95	95	95.43	95.08
3.5	91.82	92	94.26	94.44			95.8	95.33
4	90.42	90.17	93.91	94.2				

Table 5.20:Case B Sample Size 100 Coverage summary results

N=500

The results for this sample size are displayed in Figure 5.29 to Figure 5.34 and Table 5.21 to Table 5.24.

As was observed with case A, the increased sample size reduces the bias associated with estimating the model parameters to a negligible level for the Reeves and Conditional Score methods. For the ordinary logistic regression method, the increased sample size has little effect on the size of the bias associated with estimating α_t and β_t. These trends are reflected in the mean values for the Odds Ratio with the increased sample size reducing the associated standard errors. For the coverage terms, the ordinary logistic regression method gave values as low as 75% for β_t whereas the Conditional Score method reduces only to about 94%.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-4.5318	0.4828	-4.53175	0.48291	-4.54678	0.48428	-4.53175	0.4828
0.5	-4.5365	0.48298	-4.54415	0.4837	-4.5606	0.48615	-4.54771	0.48466
1	-4.4978	0.47971	-4.52763	0.48228	-4.5195	0.4855	-4.54192	0.48613
1.5	-4.4582	0.47566	-4.52376	0.48129	-4.4814	0.48715	-4.55652	0.48993
2	-4.3906	0.46997	-4.50274	0.47978	-4.4073	0.48834	-4.56202	0.49505
2.5	-4.3155	0.46307	-4.48287	0.47811	-4.3621	0.49214	-4.5782	0.50208
3	-4.2251	0.45487	-4.45299	0.47611	-4.2963	0.49645	-4.59295	0.5104
3.5	-4.1329	0.44644	-4.42454	0.47478	-4.2098	0.49906	-4.62077	0.52169
4	-4.0234	0.43652	-4.37853	0.47267	-4.11052	0.50045	-4.64185	0.53423

Table 5.21: Case B Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10082	0.01365	0.01378	0.10082	0.01365	0.01378	0.10118	0.01369	0.01384	0.10082	0.01365	0.01378
0.5	0.10087	0.01365	0.01357	0.10117	0.01369	0.01363	0.1015	0.01374	0.0136	0.10117	0.01369	0.01363
1	0.09972	0.01355	0.01369	0.10091	0.01371	0.01391	0.10052	0.01374	0.0144	0.10092	0.0137	0.01391
1.5	0.09853	0.01343	0.0136	0.10117	0.01378	0.01408	0.09903	0.01378	0.01392	0.10118	0.01375	0.01408
2	0.09639	0.01326	0.01334	0.101	0.01386	0.01418	0.09719	0.01386	0.0136	0.10102	0.01382	0.01419
2.5	0.09413	0.01305	0.01327	0.10117	0.01398	0.01458	0.09563	0.01399	0.01331	0.1012	0.01392	0.0146
3	0.09129	0.0128	0.01291	0.10112	0.01412	0.01475	0.09377	0.01415	0.0139	0.10115	0.01403	0.01478
3.5	0.08839	0.01254	0.01253	0.10135	0.01431	0.01497	0.09114	0.01426	0.01389	0.10141	0.01419	0.01501
4	0.08511	0.01224	0.01228	0.10142	0.0145	0.01544	0.08847	0.01434	0.01324	0.10152	0.01437	0.01551

Table 5.22:Case B Sample Size 500 $\hat{\beta}_t$ summary results

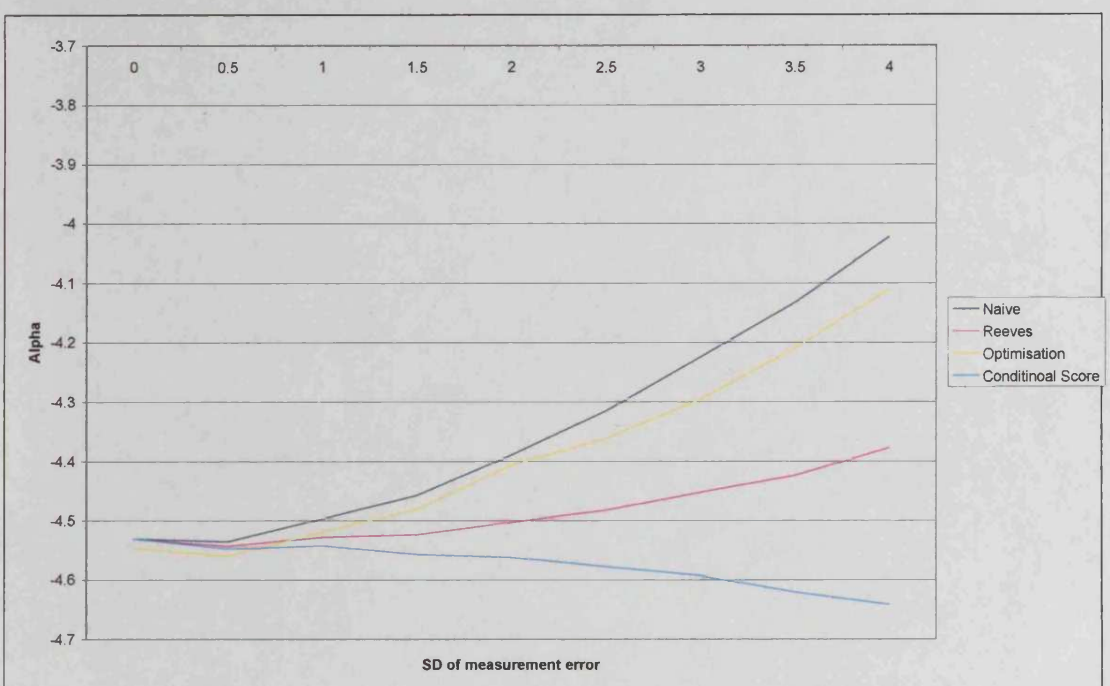


Figure 5.29: Case B Sample size 500 $\hat{\alpha}_t$ method comparison

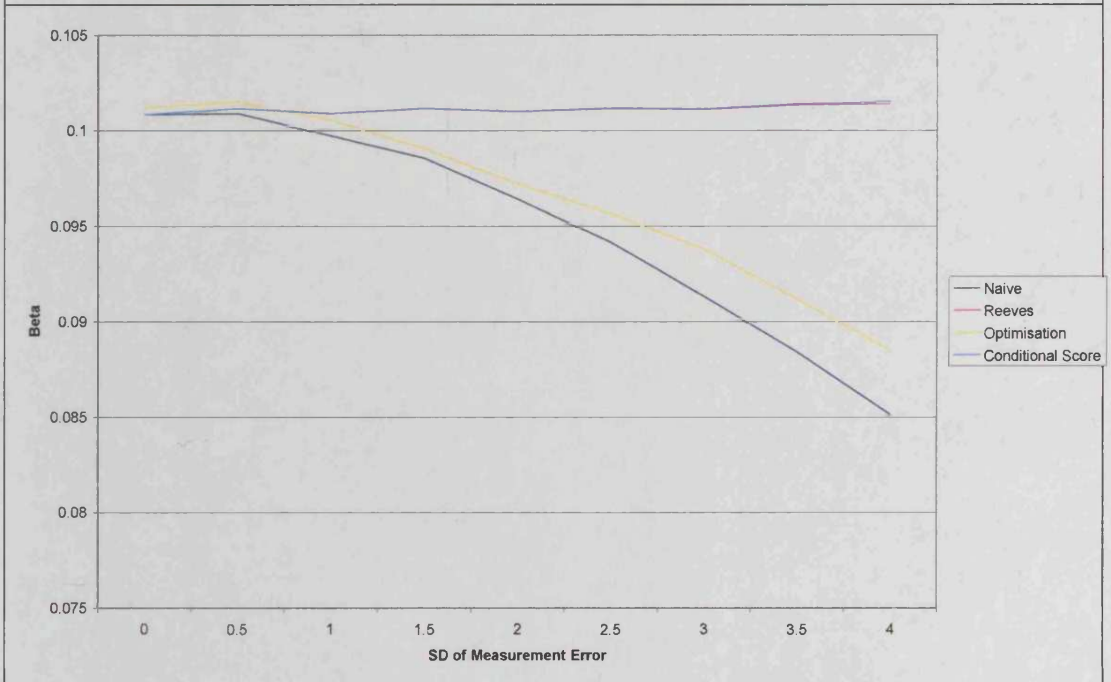


Figure 5.30: Case B Sample size 500 $\hat{\beta}_t$ method comparison

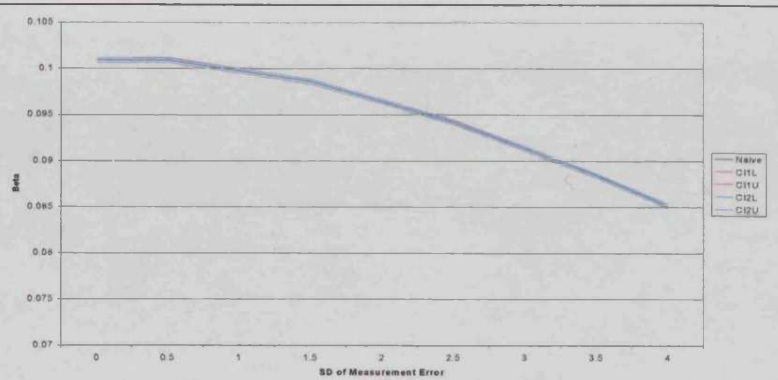


Figure 5.31: Case B Sample size 500 $\hat{\beta}_t$ Olr Confidence Interval comparison

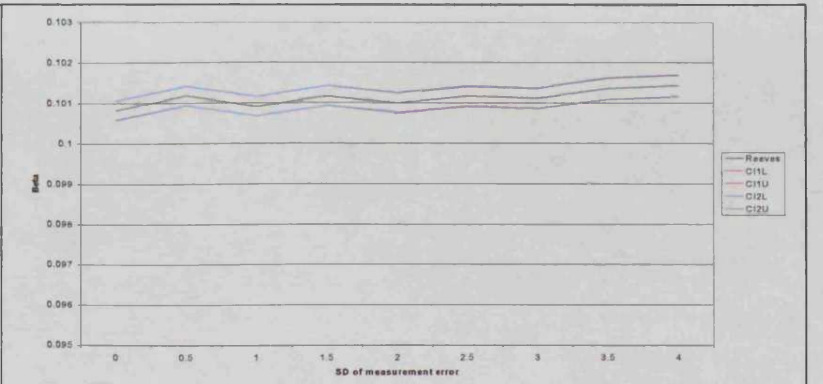


Figure 5.32: Case B Sample size 500 $\hat{\beta}_t$ Reeves Confidence Interval comparison

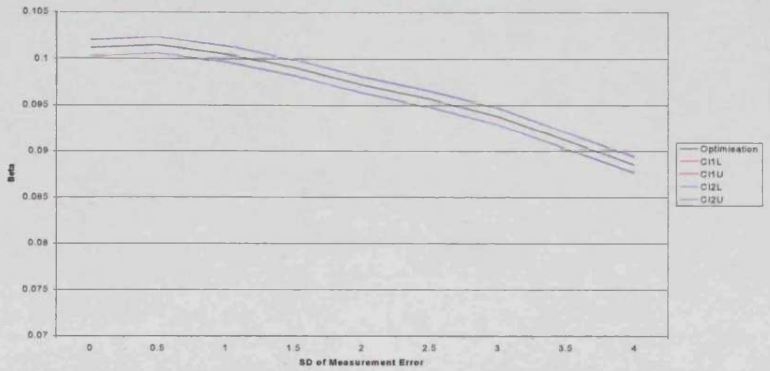


Figure 5.33: Case B Sample size 500 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

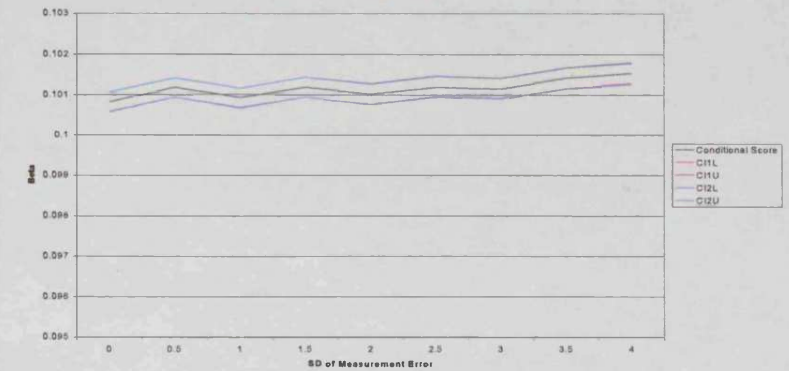


Figure 5.34: Case B Sample size 500 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.965	0.75983	3.96501	0.75979	3.98515	0.76992	3.965	0.75983
0.5	3.96545	0.7456	3.98215	0.75188	3.99982	0.75472	3.98224	0.75192
1	3.90597	0.74396	3.97143	0.76912	3.95586	0.79509	3.97179	0.7694
1.5	3.84268	0.72849	3.98746	0.78496	3.87274	0.76308	3.98813	0.78539
2	3.73134	0.69206	3.97919	0.78818	3.77394	0.71414	3.98038	0.78899
2.5	3.61866	0.66823	3.9925	0.81701	3.69238	0.67661	3.99475	0.81929
3	3.47949	0.6215	3.99141	0.82263	3.60698	0.70465	3.99384	0.82497
3.5	3.34302	0.57939	4.00661	0.8408	3.48037	0.66784	4.01037	0.84467
4	3.19638	0.54367	4.0159	0.87625	3.35183	0.61416	4.02238	0.88315

Table 5.23:Case B Sample Size 500 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	94.9	95.1	94.91	95.1	95.2	95.2	94.91	95.1
0.5	95.08	95.26	95.11	95.25	95	95.3	94.92	95.19
1	94.99	95.04	95.05	95.2	93.8	94.3	95.02	94.88
1.5	94.38	94.42	94.75	94.85	94.6	94.8	95.04	94.86
2	93.58	93.3	94.79	94.98	95.2	94.8	94.95	94.85
2.5	91.69	91.53	94.54	94.59	94.7	94.6	94.65	94.26
3	88.39	87.45	93.92	94.15	91.2	92.1	94.44	94.05
3.5	83.82	82.16	92.94	93.45	89.1	89.7	94.49	93.99
4	77.3	74.52	92.37	93.75	86.3	87	93.71	93.61

Table 5.24:Case B Sample Size 500 Coverage summary results

N=1000

The results for this sample size are displayed in Figure 5.35 to Figure 5.40 and Table 5.25 to Table 5.28.

The further increased sample size had the same effect on the expected values of the model parameters and associated standard errors by each of the methods that were observed for case A. The same trends and patterns are also apparent.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-4.5218	0.33998	-4.52177	0.34006	-4.52778	0.34058	-4.52177	0.33998
0.5	-4.5106	0.33924	-4.51813	0.33975	-4.5106	0.33992	-4.52158	0.34041
1	-4.4839	0.33756	-4.51354	0.33938	-4.4963	0.34105	-4.52743	0.34203
1.5	-4.4323	0.33439	-4.49708	0.33835	-4.4456	0.34225	-4.52871	0.34427
2	-4.376	0.33069	-4.48726	0.33758	-4.416	0.34458	-4.54499	0.34809
2.5	-4.3009	0.32577	-4.46707	0.33635	-4.3399	0.34561	-4.55941	0.35277
3	-4.2095	0.32005	-4.43576	0.33497	-4.2723	0.34819	-4.57196	0.35852
3.5	-4.1111	0.31378	-4.4	0.33362	-4.1824	0.35009	-4.59006	0.36565
4	-4.0015	0.30665	-4.35285	0.33194	-4.09485	0.35194	-4.6066	0.37374

Table 5.25: Case B Sample Size 1000 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10053	0.00961	0.00967	0.10053	0.00961	0.00967	0.10065	0.00962	0.0099	0.10053	0.00961	0.00967
0.5	0.1002	0.00959	0.00969	0.1005	0.00961	0.00972	0.10015	0.00961	0.00939	0.1005	0.00961	0.00972
1	0.09938	0.00954	0.00964	0.10055	0.00965	0.00978	0.09994	0.00965	0.00939	0.10056	0.00964	0.00979
1.5	0.0978	0.00944	0.00947	0.1004	0.00968	0.0098	0.09833	0.0097	0.00964	0.10041	0.00966	0.0098
2	0.09607	0.00933	0.00924	0.10062	0.00975	0.00981	0.09744	0.00977	0.00942	0.10064	0.00972	0.00982
2.5	0.09376	0.00918	0.00919	0.1007	0.00983	0.01008	0.09518	0.00982	0.00926	0.10073	0.00978	0.01009
3	0.09094	0.00901	0.00908	0.10063	0.00993	0.01035	0.09322	0.00992	0.01015	0.10067	0.00985	0.01037
3.5	0.0879	0.00882	0.00882	0.10064	0.01005	0.01051	0.09055	0.01	0.00938	0.1007	0.00995	0.01054
4	0.0846	0.0086	0.00871	0.10062	0.01017	0.0109	0.08813	0.01009	0.00978	0.10069	0.01006	0.01094

Table 5.26:Case B Sample Size 1000 $\hat{\beta}_t$ summary results

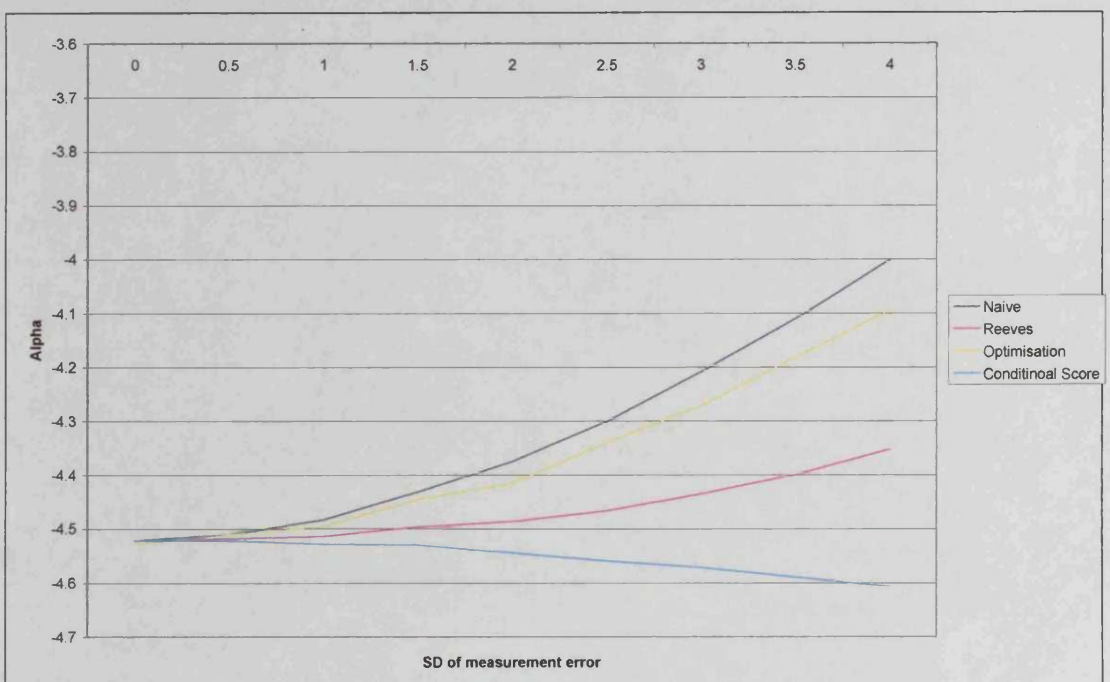


Figure 5.35: Case B Sample size 1000 $\hat{\alpha}_i$ method comparison

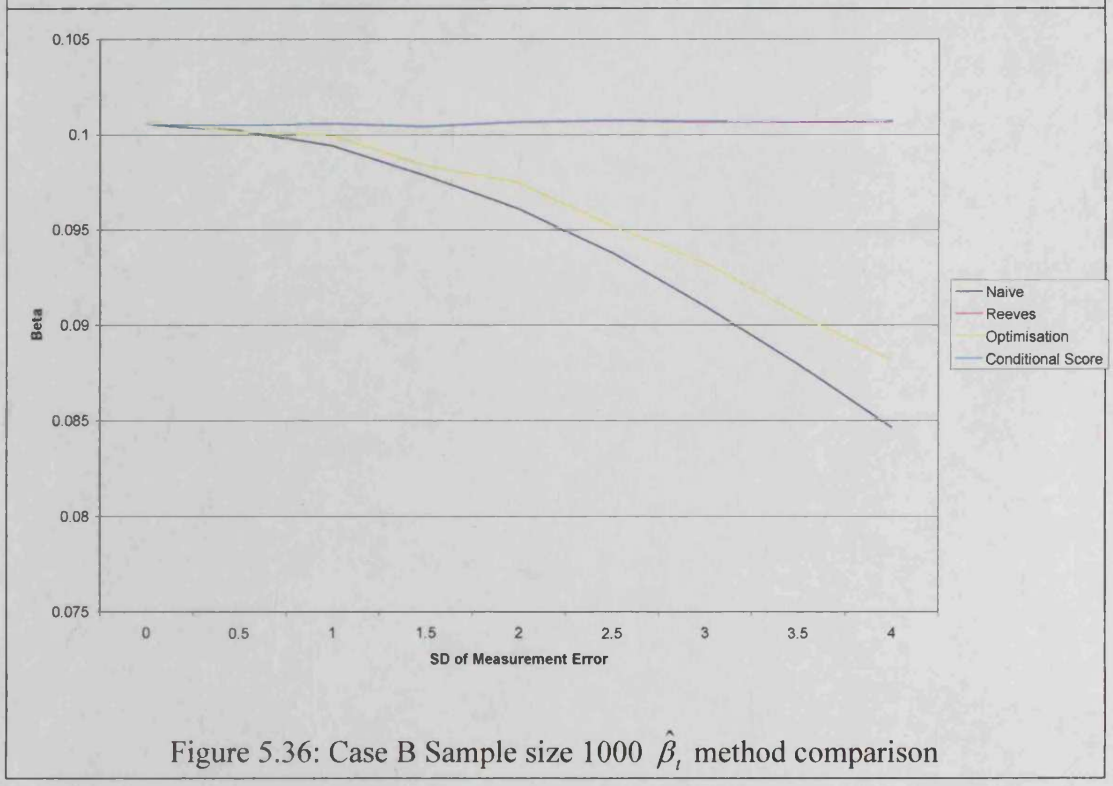


Figure 5.36: Case B Sample size 1000 $\hat{\beta}_i$ method comparison

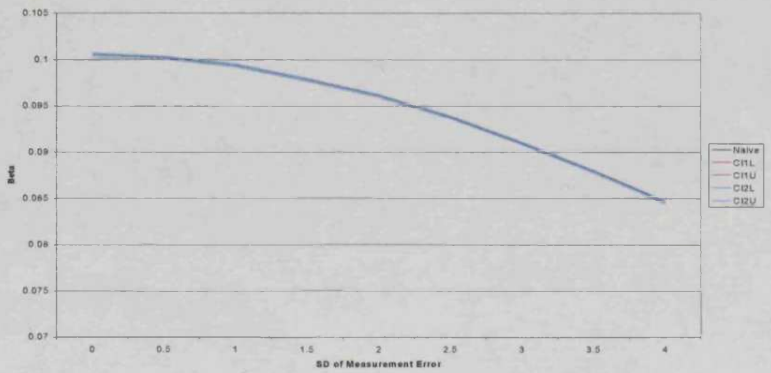


Figure 5.37: Case B Sample size 1000 $\hat{\beta}_t$ Olr Confidence Interval comparison

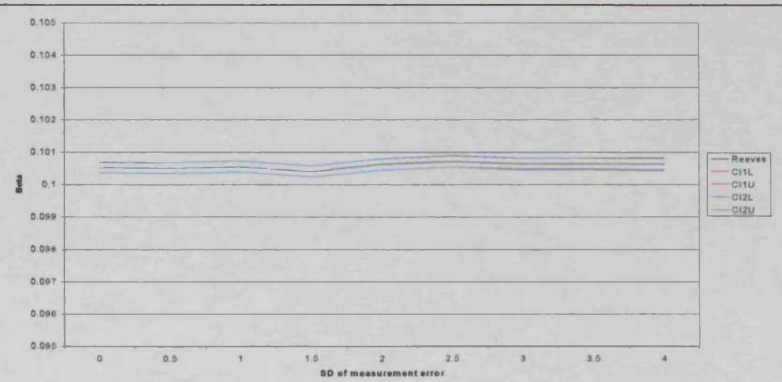


Figure 5.38: Case B Sample size 1000 $\hat{\beta}_t$ Reeves Confidence Interval comparison

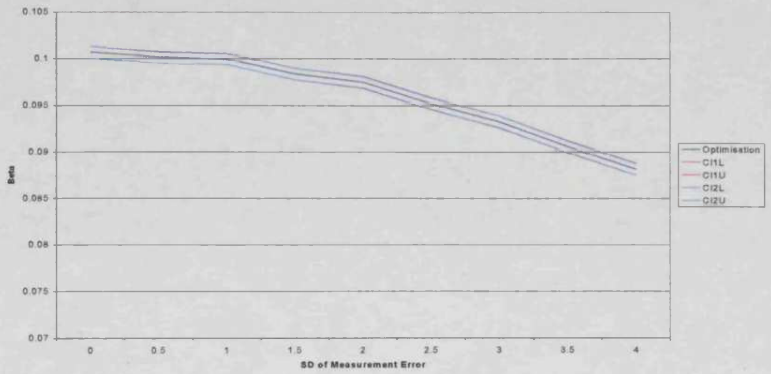


Figure 5.39: Case B Sample size 1000 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

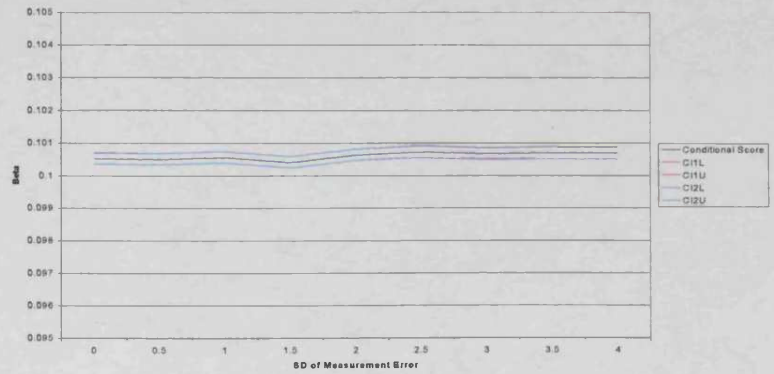


Figure 5.40: Case B Sample size 1000 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.91436	0.51935	3.91436	0.51936	3.92233	0.53067	3.91436	0.51935
0.5	3.8973	0.51742	3.91317	0.52161	3.89219	0.49455	3.91326	0.52164
1	3.85375	0.50729	3.91645	0.52382	3.88161	0.50028	3.91673	0.52388
1.5	3.77152	0.48766	3.90863	0.52353	3.80006	0.50133	3.90922	0.52364
2	3.68297	0.46414	3.91994	0.52568	3.75301	0.48274	3.92112	0.52621
2.5	3.57009	0.44894	3.9266	0.54355	3.63954	0.46449	3.9281	0.54445
3	3.43605	0.42647	3.92491	0.55829	3.55015	0.50123	3.9271	0.55946
3.5	3.29636	0.39743	3.92647	0.56787	3.41975	0.43927	3.92971	0.56999
4	3.15233	0.37523	3.92831	0.59039	3.31236	0.44928	3.93255	0.59316

Table 5.27:Case B Sample Size 1000 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	94.98	94.93	94.98	94.93	94.8	94.4	94.98	94.93
0.5	95.02	94.9	94.96	95.01	96.3	96.1	95	95.12
1	94.61	94.52	94.69	94.74	96.5	95.7	94.87	94.68
1.5	94.08	94.02	94.86	94.89	94.7	95	94.58	94.57
2	92.47	92.27	94.88	95.08	94.9	94.6	95.04	94.99
2.5	88.98	88.31	94.2	94.38	92.5	93.4	94.61	94.43
3	82.52	80.64	93.48	94.32	87.1	88	94.25	93.74
3.5	73.05	69.31	92.51	93.98	84.8	84.4	93.99	93.63
4	60.55	54.54	90.82	93.8	75.5	75.5	93.65	93.15

Table 5.28:Case B Sample Size 1000 Coverage summary results

Conclusion for Case B

In conclusion, the effect of reducing the prevalence of the disease size while leaving β_t unchanged was a slight increase in both the bias and the standard errors in the estimates of the parameters. As the sample size was increased, so the bias in the mean of $\hat{\beta}_t$ and the associated standard errors decreased.

This case also shows that if any method is used to estimate the Odds Ratio when an explanatory variable could be subject to measurement error, then misleading results could be concluded for small sample sizes. For a sample size of 500 and above, both the Reeves and Conditional Score methods can correct reasonably well for the measurement error when estimating the model parameters, even when there is a large measurement error standard deviation. For further precision in these estimates a large sample size is recommended.

Case C $\alpha_t = -6$ and $\beta_t = 0.1$

Here the prevalence of the disease has been reduced much further, to about 7%, but the regression coefficient of the explanatory variable has remained the same.

N=100

The results for this sample size are displayed in Figure 5.41 to Figure 5.46 and Table 5.29 to Table 5.32.

In the two previous cases, for all four methods when the sample size was 100 and $\sigma_v = 0$, the expected values of $\hat{\beta}_t$ compared with the true value of β_t, had a slight positive bias. In this case, when n=100, and $\sigma_v = 0$, Table 5.30 shows that all the expected values of $\hat{\beta}_t$ have a considerable negative bias in comparison.

As σ_v increased, for the ordinary logistic regression method, the expected values of $\hat{\beta}_t$ become more negatively biased. For the Reeves and conditional score methods, the negative bias reduces as σ_v is increased. The optimization method followed the same pattern as the ordinary logistic regression method but with less bias.

Figure 5.42 shows a graphical comparison between the methods where it can be seen that the Reeves method produced the least biased results. Figure 5.43 to Figure 5.46 show the confidence intervals scaled to each method displaying the same pattern previously observed in the other cases. Again showing that the bias associated with the expected values of $\hat{\beta}_t$ are real and are not due to chance.

When comparing the standard errors the same patterns of change for the methods observed for cases A and B were also observed for case C. For the Reeves and conditional score methods, the differences between the two standard errors is greater that previously observed with the empirical standard error being the larger of the two. For the ordinary logistic regression method and the optimization method, however, the likelihood standard error is larger in comparison to the empirical standard error.

Comparing the ordinary logistic regression and Reeves methods' mean of the odds ratio estimates in Table 5.31, when $\sigma_v = 4$ the ordinary logistic regression method under-estimates the mean of the odds ratio estimates by 0.51 whereas the Reeves method over-estimates by 1.01. The associated standard error of the odds ratio for the Reeves method is also very large, showing there is great variation in the estimates. The standard errors are larger than for Cases A and B with the same sample size, due to the very small number of events likely to be observed in this case.

In terms of the coverage, when $\sigma_v = 4$, the ordinary logistic regression method provides an approximate 89% coverage in comparison to the Reeves method with 93%.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-5.6546	1.74792	-5.65464	1.7423	-5.65464	1.74808	-5.65464	1.74808
0.5	-5.6443	1.74552	-5.65128	1.74043	-5.6699	1.73776	-5.65504	1.75166
1	-5.6148	1.73881	-5.64221	1.73787	-5.6311	1.73476	-5.65734	1.76136
1.5	-5.5915	1.72453	-5.65207	1.73102	-5.6851	1.76833	-5.68673	1.77405
2	-5.513	1.70967	-5.61575	1.72331	-5.5374	1.77593	-5.67754	1.79692
2.5	-5.4587	1.68585	-5.61328	1.71151	-5.5494	1.78377	-5.71598	1.82389
3	-5.3778	1.6651	-5.58794	1.70365	-5.514	1.82628	-5.7415	1.86606
3.5	-5.2912	1.63934	-5.56071	1.69571			-5.77933	1.91767
4	-5.2274	1.61358	-5.5587	1.69044				

Table 5.29: Case C Sample Size 100 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.08982	0.04622	0.04542	0.08982	0.04646	0.04542	0.08982	0.04622	0.04542	0.08982	0.04622	0.04542
0.5	0.08957	0.04616	0.04563	0.08986	0.04651	0.04581	0.09083	0.04615	0.044	0.08984	0.04629	0.04579
1	0.08858	0.04601	0.04601	0.08972	0.0467	0.04674	0.09002	0.04617	0.0439	0.08966	0.04647	0.04667
1.5	0.08805	0.04557	0.04451	0.09061	0.0468	0.04612	0.0906	0.04702	0.04334	0.09048	0.04657	0.04597
2	0.0854	0.04517	0.0444	0.08983	0.04711	0.0473	0.08618	0.04765	0.04319	0.08958	0.04692	0.04699
2.5	0.08409	0.04446	0.04391	0.09095	0.04737	0.04848	0.08722	0.04785	0.04307	0.09059	0.04722	0.04804
3	0.08148	0.04385	0.04363	0.09113	0.04787	0.0503	0.08531	0.04903	0.04369	0.09061	0.04784	0.04963
3.5	0.07885	0.04315	0.04297	0.09169	0.04854	0.05201				0.09102	0.04865	0.05113
4	0.07689	0.04228	0.04238	0.09346	0.04924	0.05453						

Table 5.30:Case C Sample Size 100 $\hat{\beta}_t$ summary results

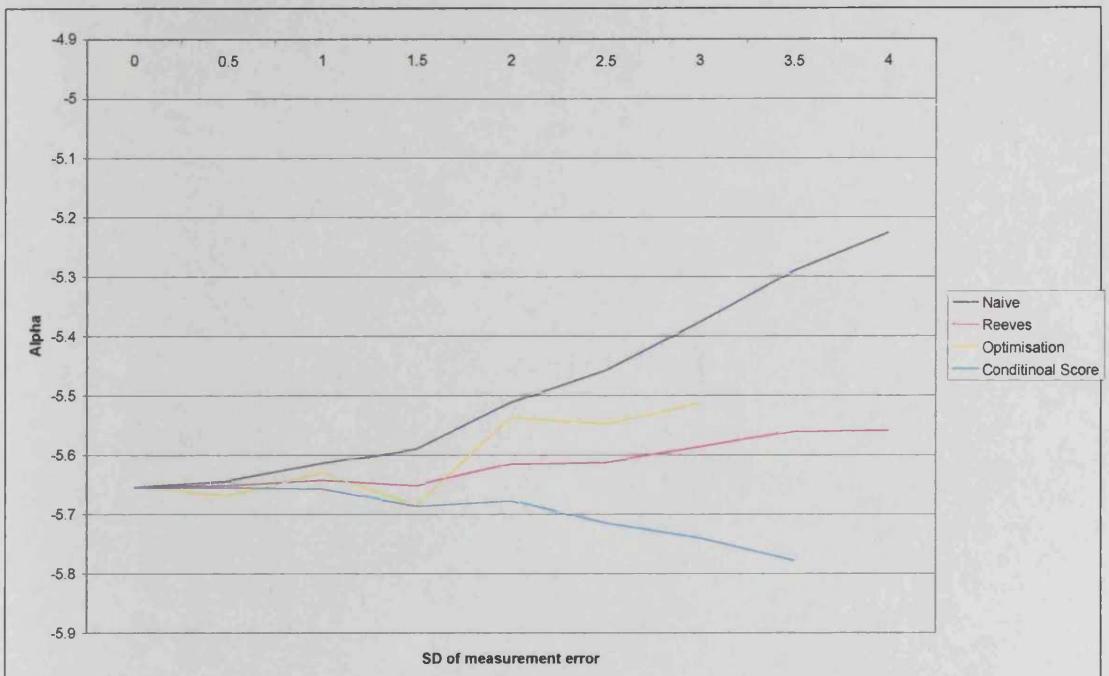


Figure 5.41: Case C Sample size 100 $\hat{\alpha}_t$ method comparison

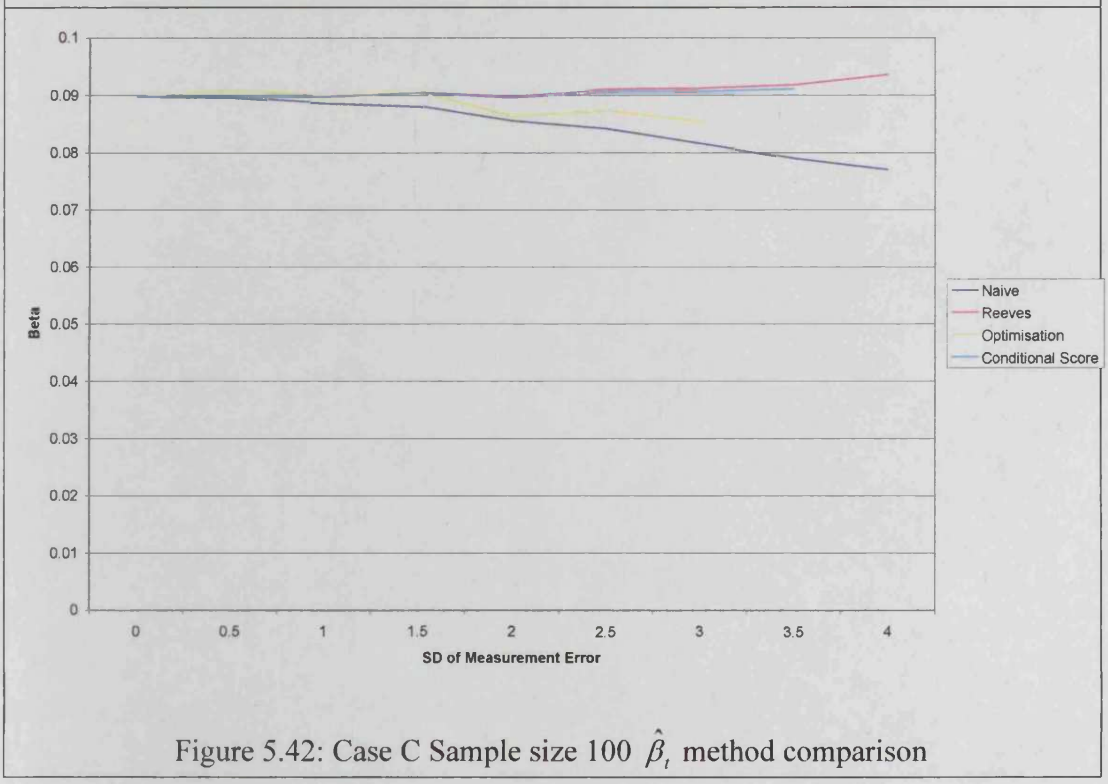


Figure 5.42: Case C Sample size 100 $\hat{\beta}_t$ method comparison

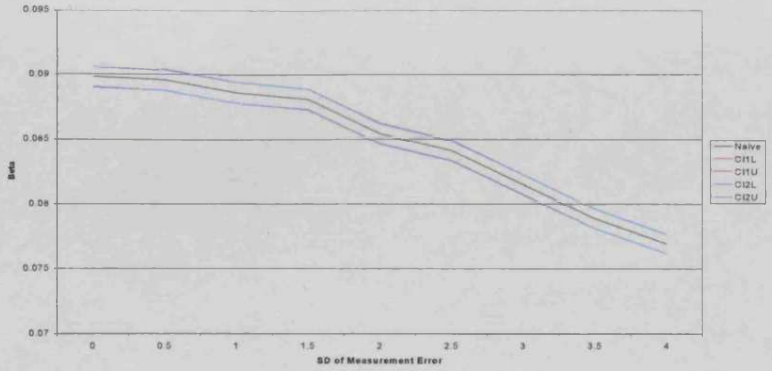


Figure 5.43: Case C Sample size 100 $\hat{\beta}_t$ Olr Confidence Interval comparison

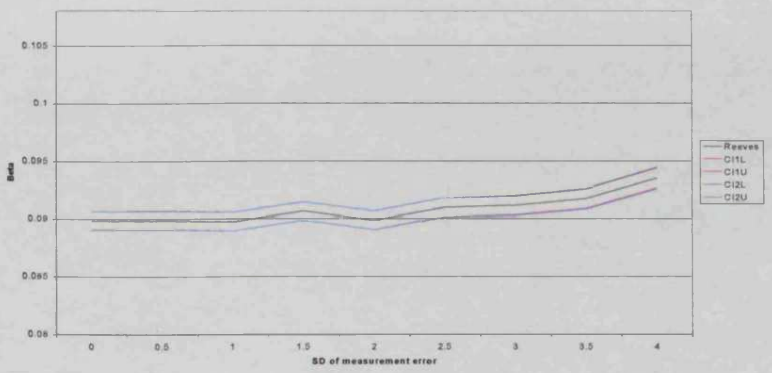


Figure 5.44: Case C Sample size 100 $\hat{\beta}_t$ Reeves Confidence Interval comparison

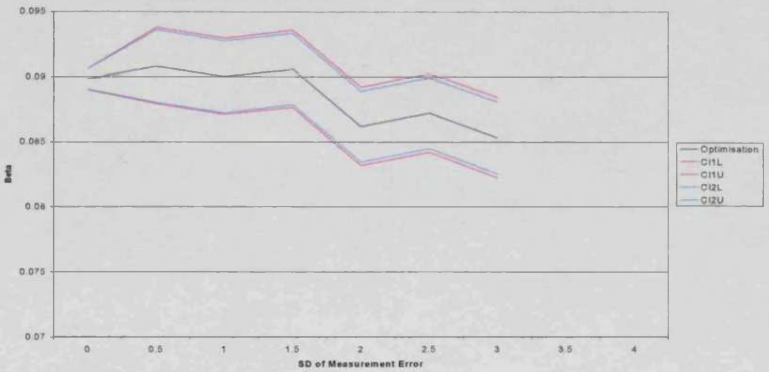


Figure 5.45: Case C Sample size 100 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

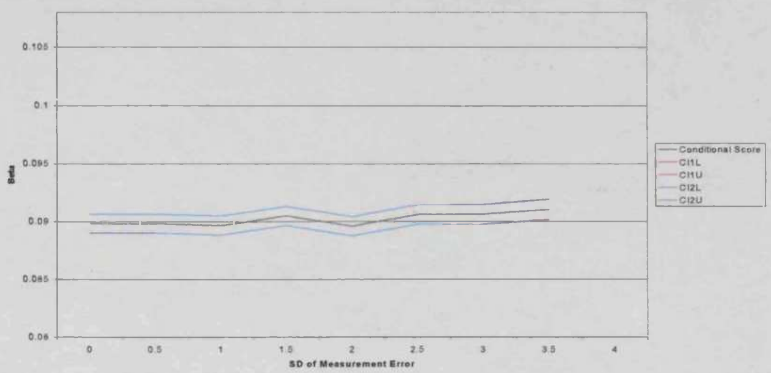


Figure 5.46: Case C Sample size 100 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	4.06409	2.92211	4.0641	2.92209	4.06409	2.92211	4.06409	2.92211
0.5	4.06725	3.05652	4.0907	3.10227	4.09087	2.90455	4.08896	3.09502
1	4.01639	2.98729	4.10917	3.16953	4.02397	2.66584	4.10211	3.1465
1.5	3.93272	2.71094	4.13392	3.0582	4.06547	2.96412	4.1188	3.0155
2	3.8029	2.72302	4.1596	3.46283	3.82928	2.73861	4.12679	3.308
2.5	3.72207	2.64509	4.28502	3.892	3.89054	3.03716	4.24017	3.79543
3	3.59007	2.71605	4.49286	3.83031	3.80432	2.7467	4.33724	5.16517
3.5	3.44076	2.35285	4.54292	5.70959			4.42067	4.45617
4	3.33693	2.24076	4.86819	7.33446				

Table 5.31:Case C Sample Size 100 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	94.61	94.11	94.04	94.72	94.32	93.58	94.04	94.72
0.5	94.37	93.62	93.67	94.37	94	94.8	94.48	93.55
1	94.17	93.23	93.77	94.4	94	94.8	94.23	93.39
1.5	94.05	93.43	93.68	94.62	94.7	95.3	94.66	93.98
2	93.45	92.64	93.27	94.04	94	95.1	94.39	93.6
2.5	92.91	92.38	93.14	94.32	92.9	95.5	94.15	93.45
3	91.99	91.28	92.72	94.01	93.6	95.6	94.22	93.22
3.5	90.94	90.38	92.31	93.68			93.91	93.18
4	89.12	88.99	91.5	93.02				

Table 5.32:Case C Sample Size 100 Coverage summary results

N=500

The results for this sample size are displayed in Figure 5.47 to Figure 5.52 and Table 5.33 to Table 5.36.

For the increased sample size of 500, the bias is now positive for all the methods when $\sigma_v = 0$ and all the patterns and trends for the mean estimates of the model parameters are as seen in previous cases for this sample size. The main point to observe is that the Conditional Score method is producing marginally better results than the Reeves method as σ_v was increased. This coupled with the fact that there is less variation in the mean of the Odds Ratio and the coverage terms almost being at the 95% level, this method is now slightly better than the Reeves method for estimating the model parameters when there is measurement error in the explanatory variable.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-6.0822	0.79034	-6.08223	0.79083	-6.07286	0.78904	-6.08223	0.79034
0.5	-6.0753	0.78996	-6.08294	0.79109	-6.0668	0.79244	-6.08661	0.7928
1	-6.0365	0.78541	-6.06649	0.78832	-6.0295	0.79581	-6.08096	0.79509
1.5	-6.0004	0.78025	-6.06639	0.78615	-5.981	0.79499	-6.09948	0.80131
2	-5.9285	0.77066	-6.04146	0.78084	-5.9435	0.80181	-6.10155	0.80745
2.5	-5.8605	0.76102	-6.02964	0.77679	-5.9071	0.80948	-6.12616	0.81809
3	-5.7752	0.7504	-6.0061	0.77288	-5.8476	0.8172	-6.14971	0.83204
3.5	-5.6655	0.73659	-5.96031	0.76723	-5.7369	0.82098	-6.16054	0.84654
4	-5.5639	0.72245	-5.92426	0.76229	-5.69008	0.83551	-6.19404	0.86548

Table 5.33: Case C Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10157	0.02046	0.02099	0.10157	0.02047	0.02099	0.10131	0.02042	0.02075	0.10157	0.02046	0.02099
0.5	0.10131	0.02044	0.02059	0.10162	0.0205	0.02067	0.10112	0.02054	0.02052	0.1016	0.0205	0.02066
1	0.10024	0.02033	0.02056	0.10145	0.02052	0.02089	0.09974	0.02064	0.02101	0.10138	0.02051	0.02085
1.5	0.0991	0.02017	0.0205	0.1018	0.0206	0.02125	0.09864	0.02067	0.01999	0.10164	0.02058	0.02117
2	0.0971	0.01992	0.02026	0.10181	0.02066	0.02156	0.09736	0.02087	0.02005	0.10152	0.02063	0.02143
2.5	0.09512	0.01964	0.01965	0.10233	0.02078	0.02164	0.09636	0.02112	0.01922	0.10187	0.02074	0.02141
3	0.09252	0.01934	0.01957	0.10264	0.02096	0.02244	0.09449	0.02138	0.02031	0.102	0.02089	0.02213
3.5	0.08933	0.01897	0.01917	0.10264	0.02115	0.02301	0.09147	0.02159	0.01968	0.1018	0.02104	0.02259
4	0.08635	0.01857	0.01891	0.10321	0.02137	0.02393	0.09002	0.02204	0.02087	0.10213	0.02125	0.02342

Table 5.34:Case C Sample Size 500 $\hat{\beta}_t$ summary results

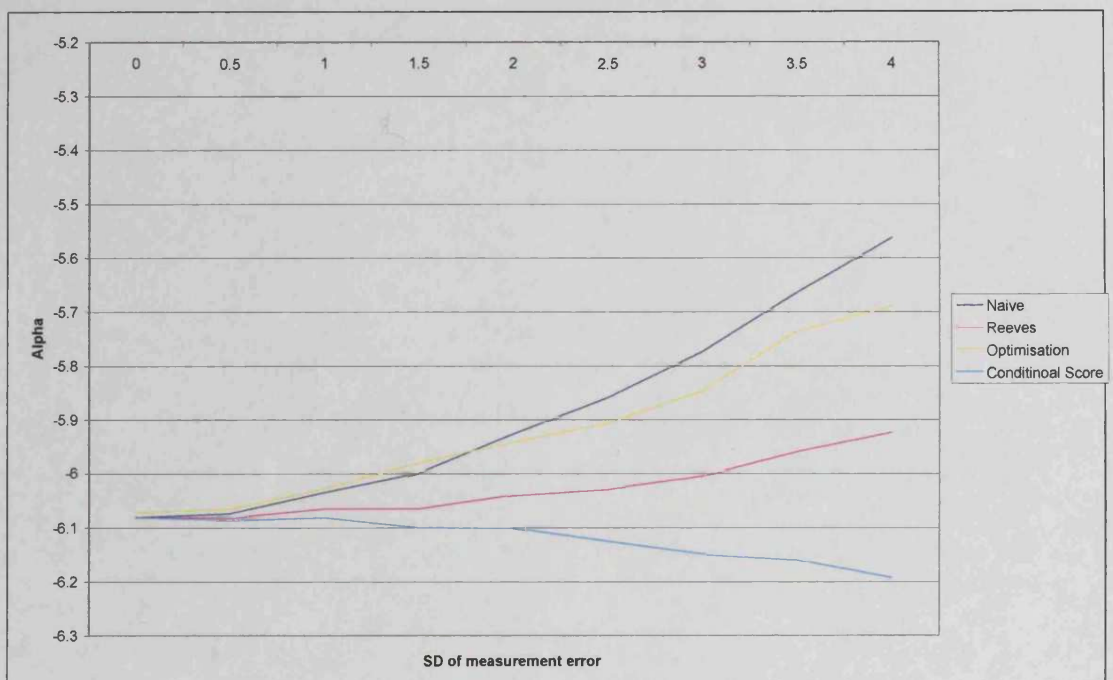


Figure 5.47: Case C Sample size 500 $\hat{\alpha}_t$ method comparison

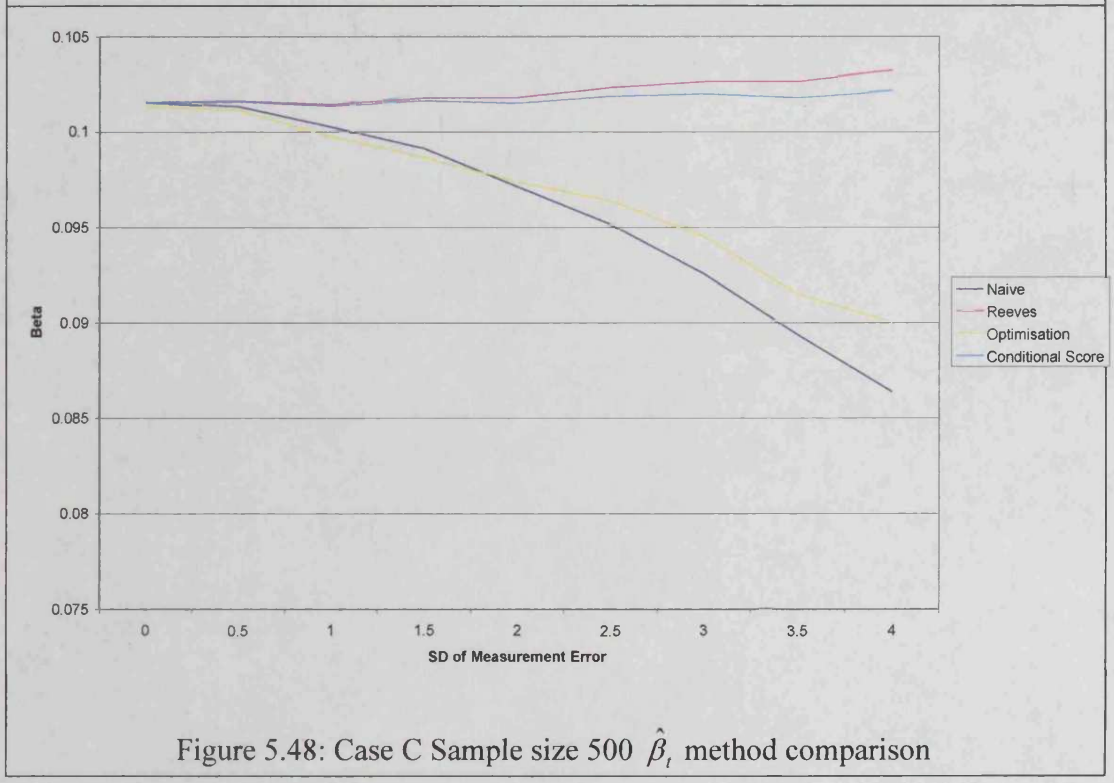


Figure 5.48: Case C Sample size 500 $\hat{\beta}_t$ method comparison

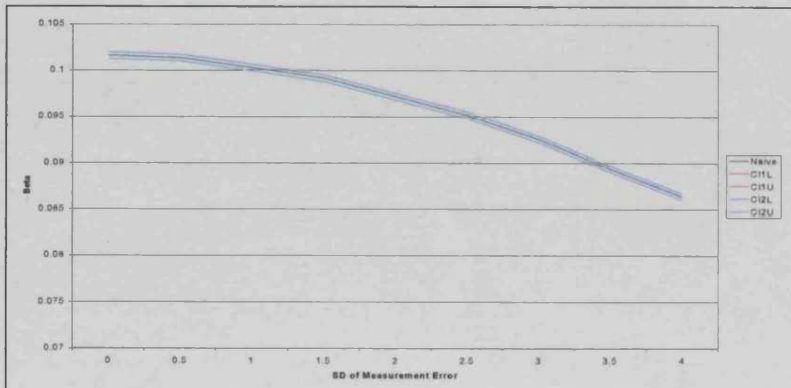


Figure 5.49: Case C Sample size 500 $\hat{\beta}_1$ Olr Confidence Interval comparison

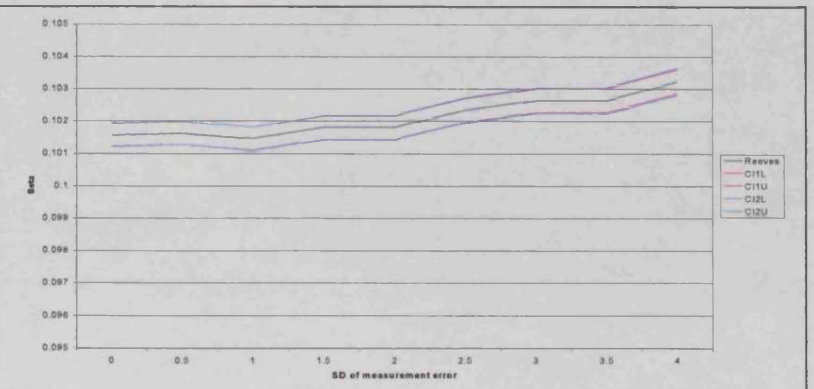


Figure 5.50: Case C Sample size 500 $\hat{\beta}_1$ Reeves Confidence Interval comparison

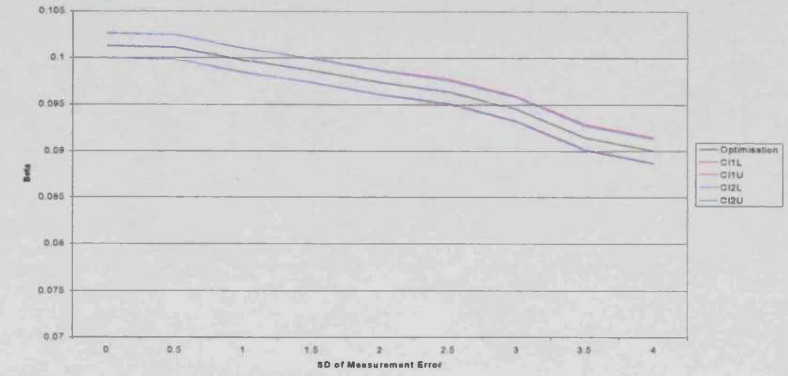


Figure 5.51: Case C Sample size 500 $\hat{\beta}_1$ Optimisation Confidence Interval comparison

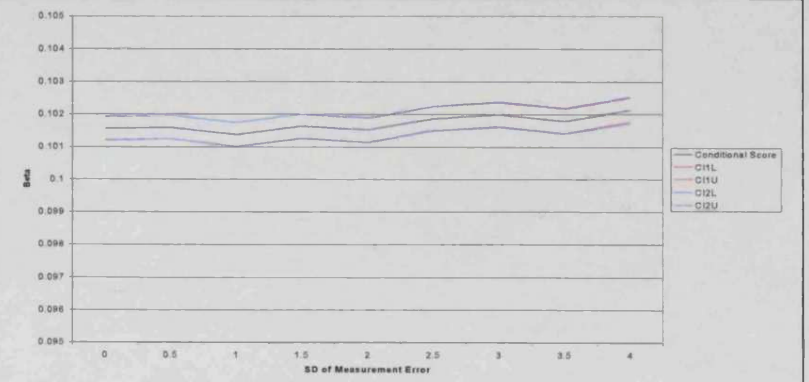


Figure 5.52: Case C Sample size 500 $\hat{\beta}_1$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	4.10153	1.24075	4.10153	1.24076	4.08171	1.2014	4.10153	1.24075
0.5	4.08107	1.2131	4.0993	1.22441	4.06617	1.16219	4.09822	1.22342
1	4.02087	1.18085	4.09253	1.22449	4.00084	1.19755	4.0879	1.22054
1.5	3.96027	1.17551	4.11973	1.27655	3.92572	1.11114	4.10894	1.2671
2	3.84947	1.10452	4.12311	1.27173	3.85873	1.08396	4.10469	1.2576
2.5	3.73977	1.04481	4.15459	1.30078	3.79685	1.03428	4.12514	1.27488
3	3.61029	1.01137	4.18744	1.38477	3.72118	1.12545	4.14579	1.34758
3.5	3.45312	0.94688	4.19867	1.43543	3.56007	0.98929	4.14392	1.38499
4	3.31335	0.88751	4.2483	1.51297	3.50818	1.05689	4.17751	1.45573

Table 5.35:Case C Sample Size 500 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	95.33	95.48	95.34	95.49	94	94.6	95.34	95.49
0.5	95.12	95.28	95.12	95.28	94.4	94.7	95.13	95.05
1	94.95	95.11	94.95	94.99	94.8	95.5	95.12	95.15
1.5	94.63	94.79	94.78	94.84	96	95.8	95.12	95.16
2	94.31	94.55	94.72	94.91	95.1	95.4	94.95	94.68
2.5	92.88	92.95	93.85	93.85	96.5	96.7	95.24	94.99
3	91.7	91.58	93.78	93.92	94.5	95.4	94.99	94.52
3.5	90.24	90.12	93.71	93.38	94.5	94.9	94.68	94.15
4	87.61	86.52	92.64	92.74	92.6	92.8	94.32	93.78

Table 5.36:Case C Sample Size 500 Coverage summary results

N=1000

The results for this sample size are displayed in Figure 5.53 to Figure 5.58 and Table 5.37 to Table 5.40.

In the previous two cases when the sample size was increased to 1000 this slightly reduced the bias in the mean of the estimates. In this case the increased sample size has caused more bias in the mean of the estimates from the ordinary logistic regression method. When comparing the Reeves and conditional score methods, the conditional score method produced slightly better estimates with more precision than the Reeves method. This can mainly be seen in the mean of the odds ratio estimates and their associated standard deviation.

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-6.0441	0.552	-6.0441	0.55239	-6.02794	0.55193	-6.0441	0.552
0.5	-6.0305	0.55069	-6.03803	0.5515	-6.0297	0.55311	-6.04144	0.55262
1	-6.0012	0.5479	-6.03089	0.5501	-6.0002	0.55299	-6.0445	0.55455
1.5	-5.9587	0.54475	-6.02385	0.54913	-5.9675	0.55747	-6.05497	0.55912
2	-5.8967	0.53874	-6.00849	0.54619	-5.9479	0.56193	-6.06483	0.56373
2.5	-5.8152	0.5313	-5.98208	0.54279	-5.8459	0.5652	-6.07207	0.56976
3	-5.7332	0.52371	-5.96128	0.54007	-5.7925	0.56988	-6.09574	0.57864
3.5	-5.6326	0.51425	-5.92449	0.53641	-5.7276	0.57674	-6.11213	0.58817
4	-5.527	0.50442	-5.88345	0.53318	-5.5985	0.5774	-6.13593	0.60029

Table 5.37: Case C Sample Size 1000 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Optimisation			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10084	0.01429	0.01445	0.10084	0.01429	0.01445	0.10047	0.0143	0.01397	0.10084	0.01429	0.01445
0.5	0.10051	0.01425	0.01429	0.10081	0.01429	0.01434	0.1005	0.01433	0.01381	0.10079	0.01429	0.01434
1	0.09968	0.01418	0.01433	0.10086	0.01432	0.01455	0.09978	0.01434	0.01397	0.10079	0.01431	0.01453
1.5	0.09834	0.01409	0.01438	0.10098	0.01439	0.01488	0.09861	0.01448	0.01399	0.10082	0.01437	0.01483
2	0.09654	0.01392	0.01396	0.10114	0.01444	0.01483	0.09794	0.01461	0.01446	0.10085	0.0144	0.01473
2.5	0.09418	0.01372	0.01376	0.1012	0.01452	0.01511	0.09483	0.01477	0.0143	0.10074	0.01445	0.01495
3	0.09175	0.0135	0.01358	0.10161	0.01463	0.01551	0.0934	0.01493	0.01463	0.10097	0.01454	0.01531
3.5	0.08882	0.01324	0.01329	0.10182	0.01476	0.01587	0.09126	0.01516	0.01402	0.10097	0.01463	0.01558
4	0.08576	0.01297	0.01321	0.10219	0.01493	0.01659	0.08802	0.01527	0.01399	0.10111	0.01476	0.0162

Table 5.38:Case C Sample Size 1000 $\hat{\beta}_t$ summary results

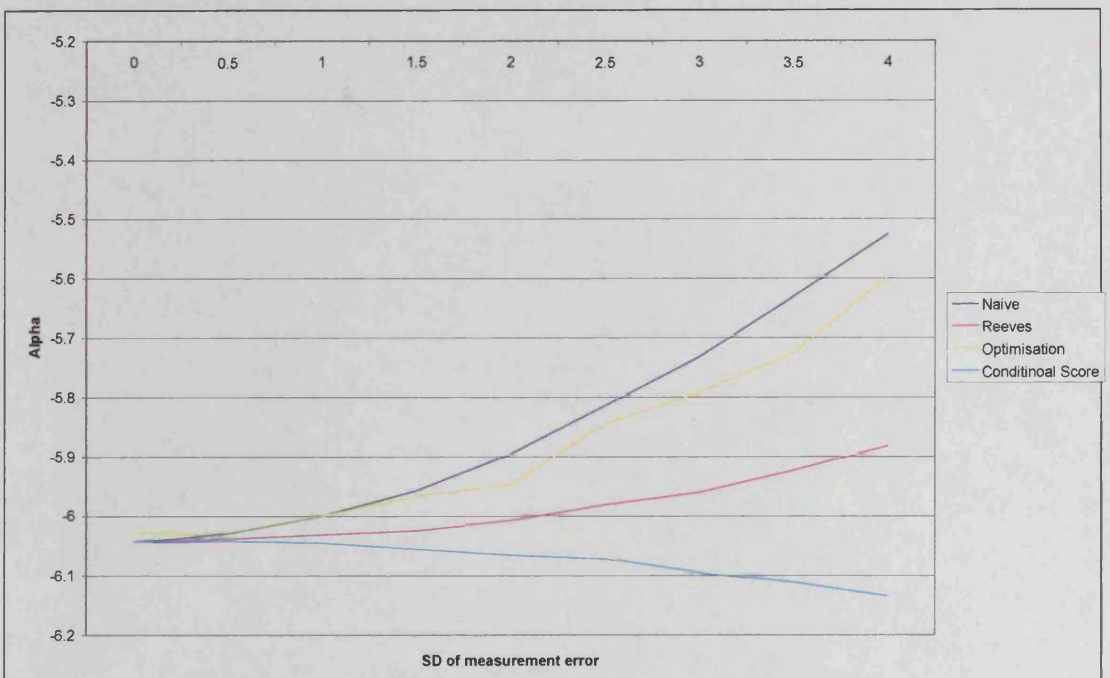


Figure 5.53: Case C Sample size 1000 $\hat{\alpha}_t$ method comparison

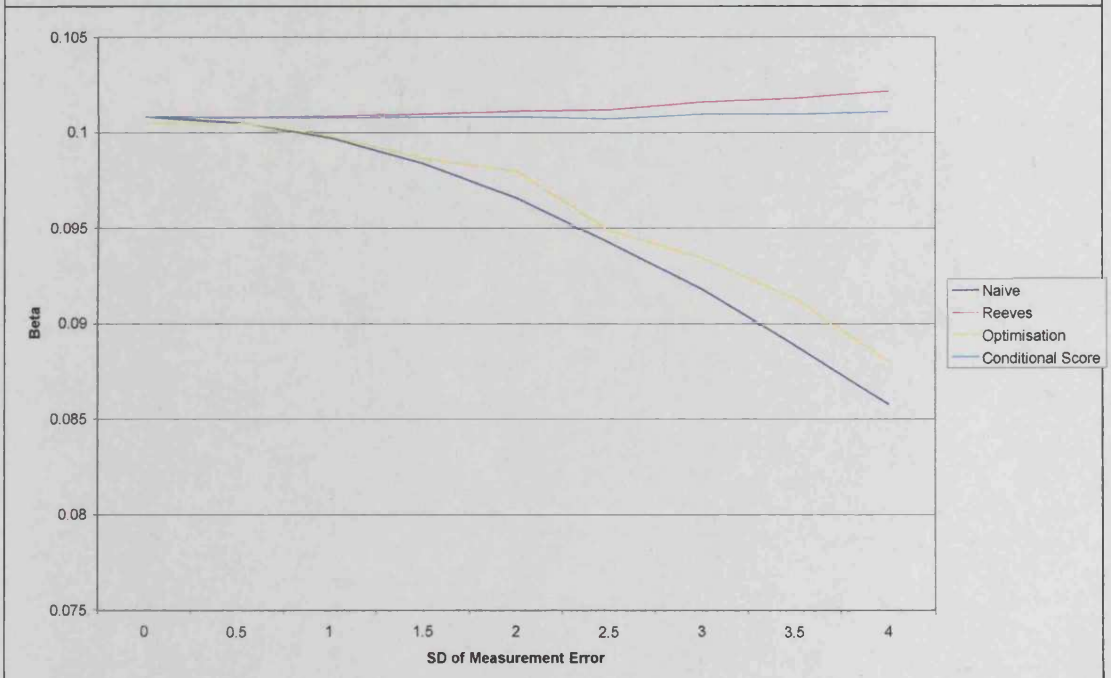


Figure 5.54: Case C Sample size 1000 $\hat{\beta}_t$ method comparison

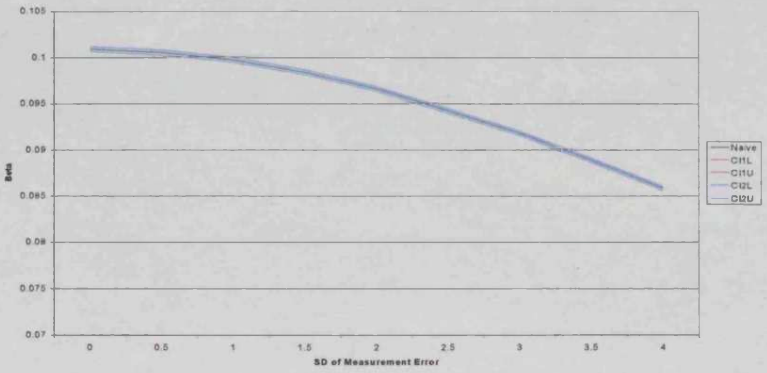


Figure 5.55: Case C Sample size 1000 $\hat{\beta}_t$ Olr Confidence Interval comparison

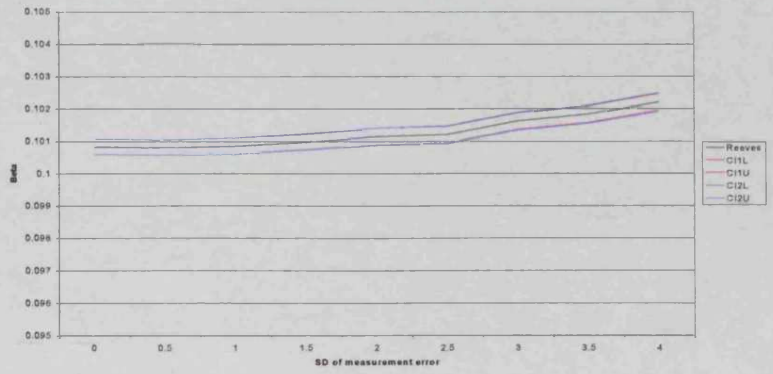


Figure 5.56: Case C Sample size 1000 $\hat{\beta}_t$ Reeves Confidence Interval comparison

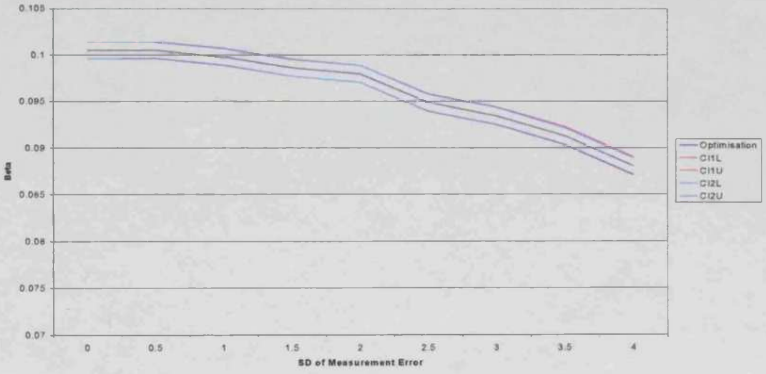


Figure 5.57: Case C Sample size 1000 $\hat{\beta}_t$ Optimisation Confidence Interval comparison

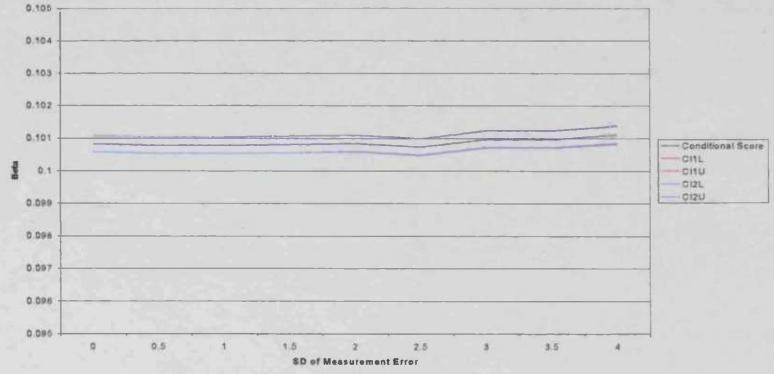


Figure 5.58: Case C Sample size 1000 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.97302	0.79655	3.97302	0.79656	3.94781	0.76179	3.97302	0.79655
0.5	3.95374	0.78361	3.9703	0.79028	3.94755	0.74463	3.96929	0.78962
1	3.90998	0.77688	3.97546	0.80313	3.91107	0.75038	3.97126	0.80094
1.5	3.84056	0.76253	3.98511	0.82097	3.84994	0.73863	3.97563	0.81572
2	3.74461	0.7276	3.99384	0.82891	3.82141	0.7756	3.97688	0.81936
2.5	3.62497	0.68966	3.99944	0.84185	3.66188	0.72533	3.97314	0.82691
3	3.50678	0.65983	4.02649	0.87529	3.59592	0.73927	3.99003	0.85551
3.5	3.36841	0.6185	4.04229	0.8996	3.48703	0.67359	3.99277	0.87056
4	3.23139	0.58926	4.07161	0.95198	3.3385	0.66002	4.0077	0.91276

Table 5.39:Case C Sample Size 1000 OR summary results

	Olr		Reeves		Optimisation		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t	α_t	β_t
0	94.85	94.88	94.86	94.91	95.8	96.3	94.86	94.91
0.5	94.95	95.19	94.92	95.19	95.5	96	95.12	95.32
1	94.88	95.08	94.97	95.11	96.6	95.9	95.02	94.92
1.5	94.63	94.49	94.77	94.85	96.3	95.6	94.74	94.7
2	94.01	94.02	94.68	94.8	94.1	94.7	95.13	94.89
2.5	92.51	92.24	94.22	94.28	94	94	94.86	94.28
3	90.15	89.68	93.67	93.76	91.9	92.7	94.52	94.22
3.5	86.21	84.73	93.12	93.38	91.8	90.9	94.21	93.79
4	81.23	78.22	92.19	92.59	89.1	89.4	93.88	93.21

Table 5.40:Case C Sample Size 1000 Coverage summary results

Conclusion for Case C

The reduction in the prevalence of the disease had a marked effect on the bias, especially for a sample size of 100. In this case, though the ordinary logistic regression method produced the most negatively biased estimates of the regression parameter, this method produced the least biased estimates of the true odds ratio value. The coverage was substantially worse using this method than for those which corrected for measurement error.

As before the effect of the measurement error cannot be ignored. In this case the conditional score method proved to provide the best mean estimates of the model parameters when the sample size was greater than 500, though the difference compared to the Reeves method is marginal.

Case D $\alpha_t = 0$ and $\beta_t = -0.1$

The prevalence of the disease is the same as seen for case C. The difference in this case is that the relationship between the disease and the explanatory variable is negative.

In some cases the optimisation method would not converge or estimated unrealistic values for the model parameters. Therefore, the simulation could not provide comparable expected values of the model parameters and hence the method has been excluded from this simulation study.

N=100

The results for this sample size are displayed in Figure 5.59 to Figure 5.63 and Table 5.41 to Table 5.44.

As we observed for case C the reduction in the prevalence of the disease and the small sample size has had an effect on the bias of estimates of β_t for all methods when there is no measurement error, Table 5.42. All three methods expected values of $\hat{\beta}_t$ are positively biased, that is the mean estimates are closer to 0, and as a result they also have negative bias when estimating α_t. As σ_v is increased, the ordinary logistic regression method expected values of $\hat{\beta}_t$ are increasingly positively biased whereas for the Reeves and conditional score methods, the bias reduces as σ_v is increased, following the same general pattern as was observed for all the previous cases. The associated standard errors of these estimates are in-line with those seen for case C when n=100.

Figure 5.61 to Figure 5.63 show that the pattern of the confidence intervals is similar to that of the expected values of $\hat{\beta}_t$. Their width confirms that, as in the previous cases, the bias is real and not due to sampling variation.

In this case the true value of the odds ratio is 0.26. The positive bias associated with all the methods when estimating β_t led to positive bias in the mean of the odds ratio estimates, Table 5.43. For the ordinary logistic regression method, when $\sigma_v = 4$, the mean of the odds ratio estimates had increased to 0.416 whereas the Reeves method it was 0.363. In this case, the associated odds ratio standard errors for the two methods were approximately the same. In comparison to the other three cases when n=100, the Reeves method produced the least biased mean of the odds ratio estimates.

The coverage term patterns, Table 5.44, were the same as previously seen in the other three cases.

	Olr		Reeves		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-0.2591	1.14836	-0.25907	1.1707	-0.25907	1.14839
0.5	-0.2741	1.14931	-0.26711	1.17248	-0.2684	1.1506
1	-0.2968	1.14637	-0.26956	1.17398	-0.27434	1.15067
1.5	-0.3121	1.13606	-0.25193	1.17438	-0.26224	1.1454
2	-0.3825	1.13245	-0.28002	1.18776	-0.29584	1.14872
2.5	-0.4259	1.11257	-0.27283	1.1818	-0.2939	1.13675
3	-0.4757	1.0997	-0.26492	1.19153	-0.29001	1.13345
3.5	-0.562	1.0847	-0.29323	1.19929	-0.31916	1.12944
4	-0.6183	1.06513	-0.28752	1.2028		

Table 5.41: Case D Sample Size 100 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	-0.09	0.04609	0.04516	-0.09	0.04637	0.04516	-0.09	0.04609	0.04516
0.5	-0.0896	0.04614	0.04556	-0.0899	0.04653	0.04574	-0.0899	0.04626	0.04572
1	-0.0884	0.04587	0.04496	-0.0895	0.0467	0.04568	-0.0895	0.04632	0.04561
1.5	-0.0882	0.04561	0.04479	-0.0908	0.04724	0.04643	-0.0906	0.04662	0.04624
2	-0.0857	0.04531	0.04516	-0.0902	0.04805	0.04813	-0.0899	0.04709	0.04781
2.5	-0.0838	0.04451	0.04333	-0.0906	0.04854	0.04783	-0.0903	0.04725	0.04736
3	-0.0822	0.04398	0.04376	-0.092	0.04976	0.05052	-0.0915	0.048	0.04984
3.5	-0.0792	0.04328	0.04344	-0.0921	0.05098	0.05267	-0.0914	0.04881	0.05173
4	-0.077	0.04245	0.04242	-0.09364	0.05236	0.05473			

Table 5.42:Case D Sample Size 100 $\hat{\beta}_t$ summary results

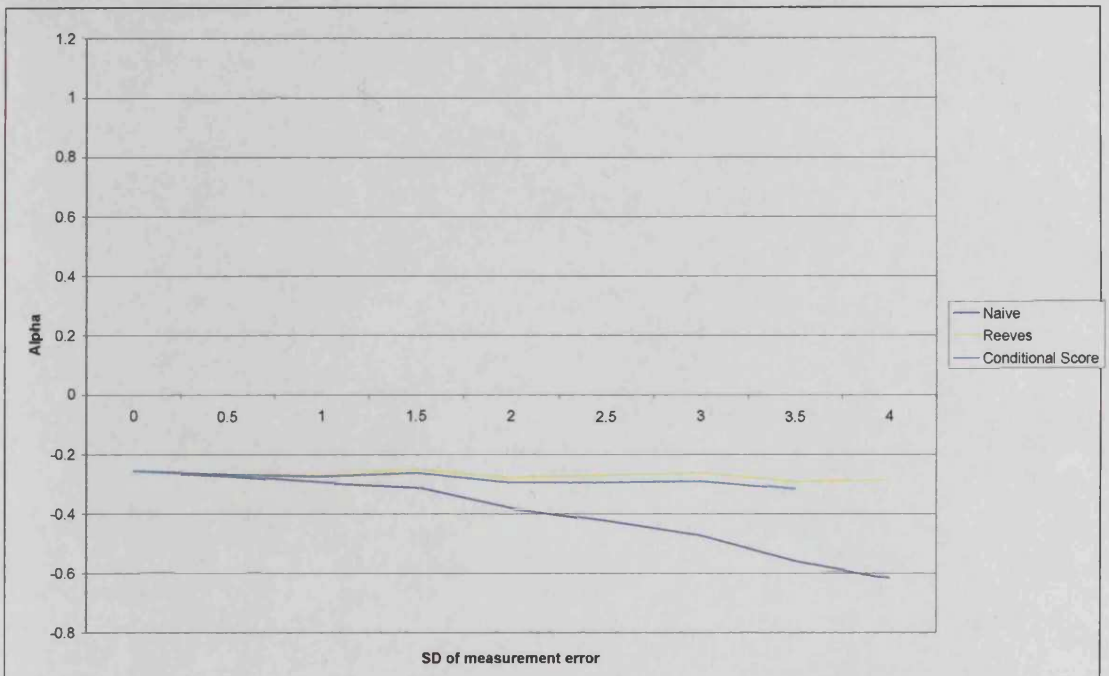


Figure 5.59: Case D Sample size 100 $\hat{\alpha}_t$ method comparison

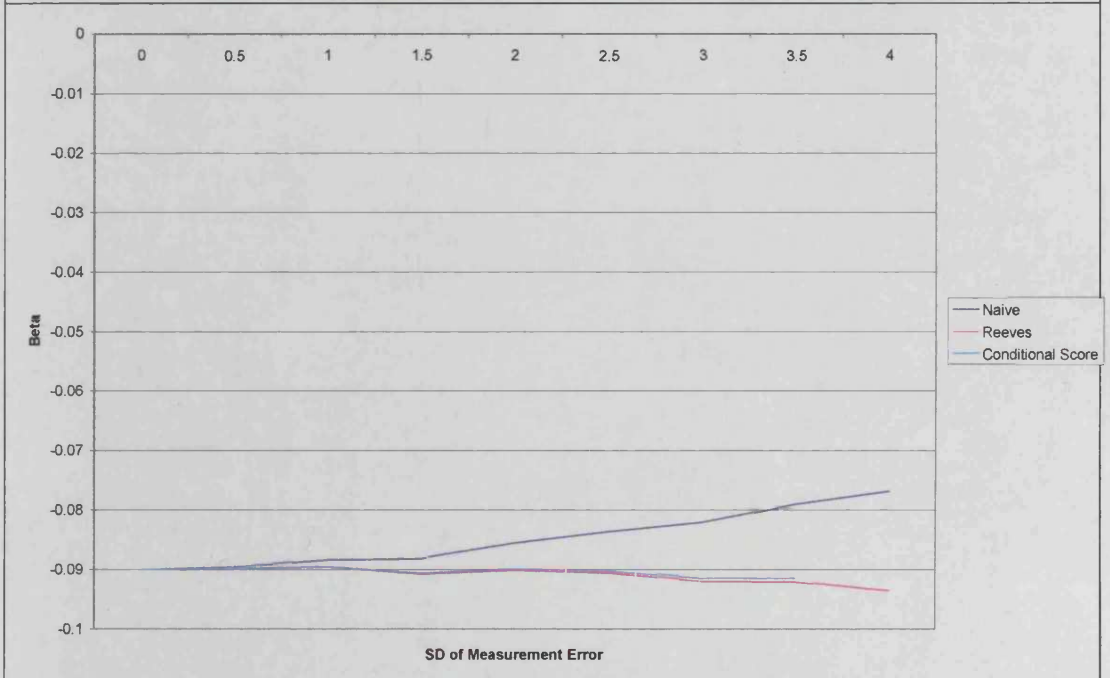


Figure 5.60: Case D Sample size 100 $\hat{\beta}_t$ method comparison

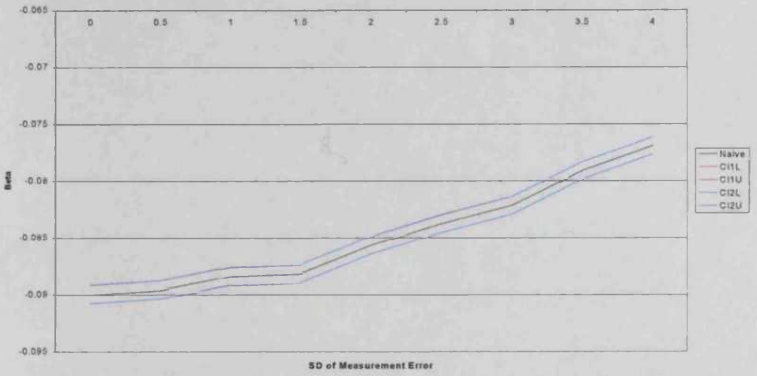


Figure 5.61: Case D Sample size 100 $\hat{\beta}_t$ Olr Confidence Interval comparison

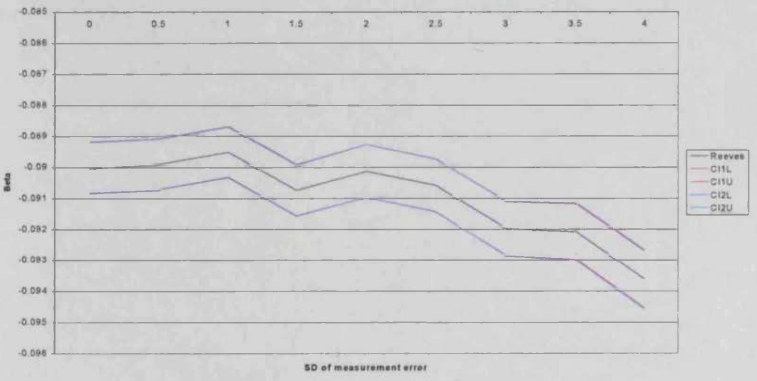


Figure 5.62: Case D Sample size 100 $\hat{\beta}_t$ Reeves Confidence Interval comparison

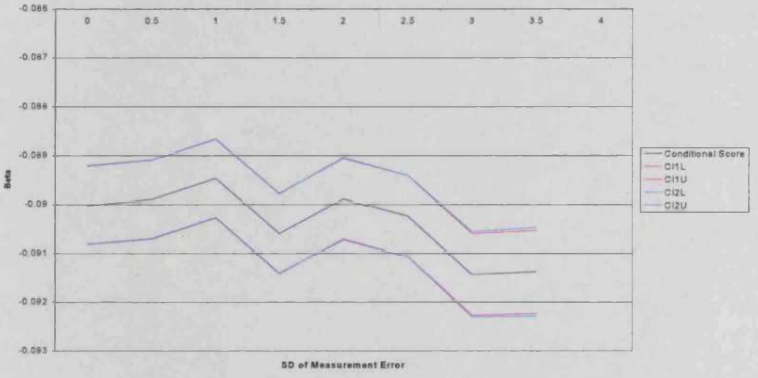


Figure 5.63: Case D Sample size 100 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	0.35782	0.24235	0.35782	0.24236	0.35782	0.24235
0.5	0.36018	0.24283	0.35927	0.24315	0.35931	0.24313
1	0.36458	0.24366	0.36097	0.24495	0.36111	0.24481
1.5	0.36533	0.24882	0.3573	0.25288	0.35762	0.25245
2	0.37857	0.25153	0.36453	0.25692	0.36506	0.25655
2.5	0.38305	0.24727	0.36122	0.25648	0.36201	0.2563
3	0.39184	0.248	0.36101	0.25971	0.36212	0.25924
3.5	0.40778	0.25769	0.36709	0.27547	0.3684	0.27427
4	0.41579	0.25096	0.36319	0.27208		

Table 5.43:Case D Sample Size 100 OR summary results

	Olr		Reeves		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	93.55	93.46	95.64	94.27	93.86	93.75
0.5	93.88	93.72	95.97	94.48	93.82	93.6
1	93.25	93.37	95.38	94.35	94.05	93.65
1.5	93.68	93.33	95.68	94.56	93.9	93.78
2	93.43	92.88	95.88	94.81	93.08	93.29
2.5	93.35	92	95.55	94.4	93.41	93.74
3	93.02	91.6	95.64	94.94	92.78	93.01
3.5	92.82	90.72	95.65	94.77	92.15	92.79
4	91.61	88.88	95.05	94.45		

Table 5.44:Case D Sample Size 100 Coverage summary results

N=500

The results for this sample size are displayed in Figure 5.64 to Figure 5.68 and Table 5.45 to Table 5.48.

As it has been seen for all previous cases the increase in sample size from 100 to 500 reduced the bias in the mean estimates of the model parameters for all the methods.

In this case, the Reeves and Conditional Score methods are almost unbiased in estimating α_t whereas the ordinary logistic regression method estimates a positive value for the mean of $\hat{\alpha}_t$ when $\sigma_v = 0$ and a negative value when $\sigma_v = 4$. When estimating β_t, the ordinary logistic regression method estimates are in line with previous cases. For the conditional score method, the expected values of $\hat{\beta}_t$ are slightly less biased than the Reeves method. On examination of the mean of the Odds Ratios estimates there is very little difference between the two methods in practice and the precision associated with these values are also approximately the same. In this case, the coverage for the Reeves method is slightly higher than that observed for the conditional score method.

	Olr		Reeves		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	0.00849	0.49471	0.00849	0.49493	0.00849	0.4947
0.5	-0.0004	0.49472	0.00725	0.49577	0.00563	0.49523
1	-0.025	0.49232	0.00493	0.49589	-0.00127	0.49379
1.5	-0.0578	0.48886	0.00803	0.49648	-0.00505	0.49182
2	-0.0997	0.48435	0.01358	0.49742	-0.00789	0.48923
2.5	-0.1574	0.47757	0.01152	0.4972	-0.01826	0.48467
3	-0.2308	0.47083	-0.00063	0.49791	-0.03755	0.48023
3.5	-0.3016	0.46335	-0.00618	0.49862	-0.04758	0.47518
4	-0.3852	0.4547	-0.02562	0.4984	-0.06865	0.46871

Table 5.45: Case D Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	-0.1014	0.02043	0.02071	-0.1014	0.02044	0.02071	-0.1014	0.02043	0.02071
0.5	-0.1012	0.02043	0.02063	-0.1015	0.02051	0.02071	-0.1014	0.02049	0.0207
1	-0.1001	0.02031	0.02081	-0.1013	0.02058	0.02114	-0.1012	0.02049	0.0211
1.5	-0.0989	0.02015	0.02033	-0.1016	0.02076	0.02106	-0.1014	0.02056	0.02098
2	-0.0973	0.01996	0.02024	-0.1021	0.02101	0.02155	-0.1018	0.02067	0.0214
2.5	-0.095	0.01964	0.02004	-0.1022	0.02125	0.02206	-0.1018	0.02074	0.02183
3	-0.0923	0.01932	0.01949	-0.1023	0.02159	0.02233	-0.1017	0.02087	0.02201
3.5	-0.0896	0.01898	0.01922	-0.1029	0.02201	0.02307	-0.1021	0.02107	0.02267
4	-0.0862	0.01856	0.01894	-0.10302	0.02241	0.02398	-0.10195	0.02123	0.02346

Table 5.46:Case D Sample Size 500 $\hat{\beta}_t$ summary results

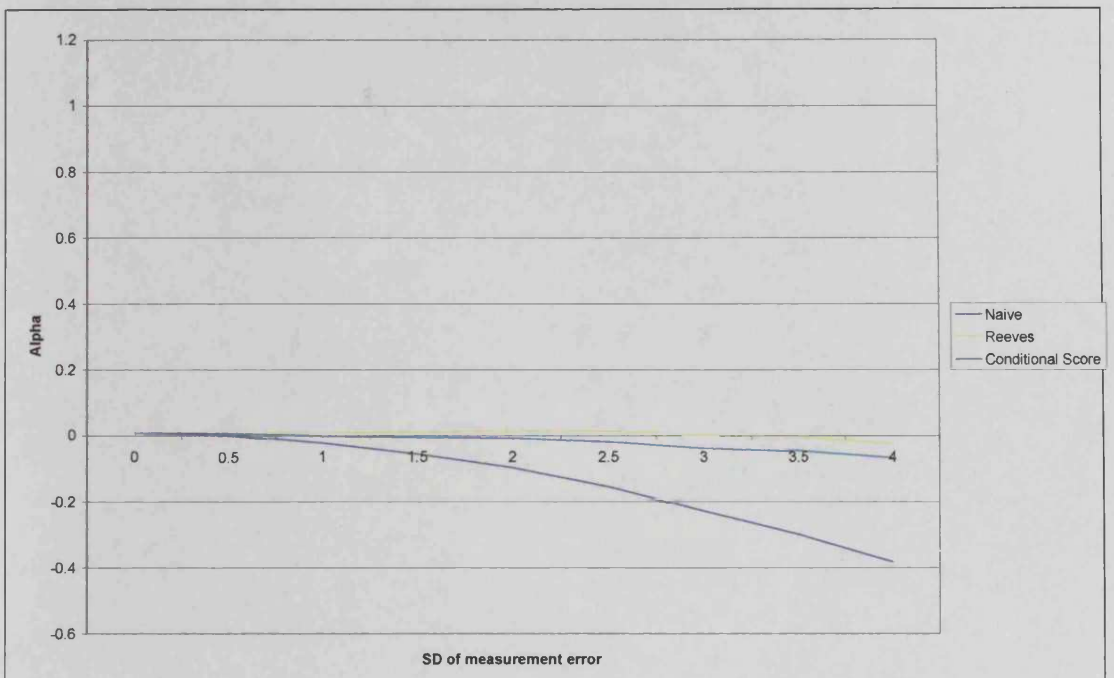


Figure 5.64: Case D Sample size 500 $\hat{\alpha}_t$ method comparison

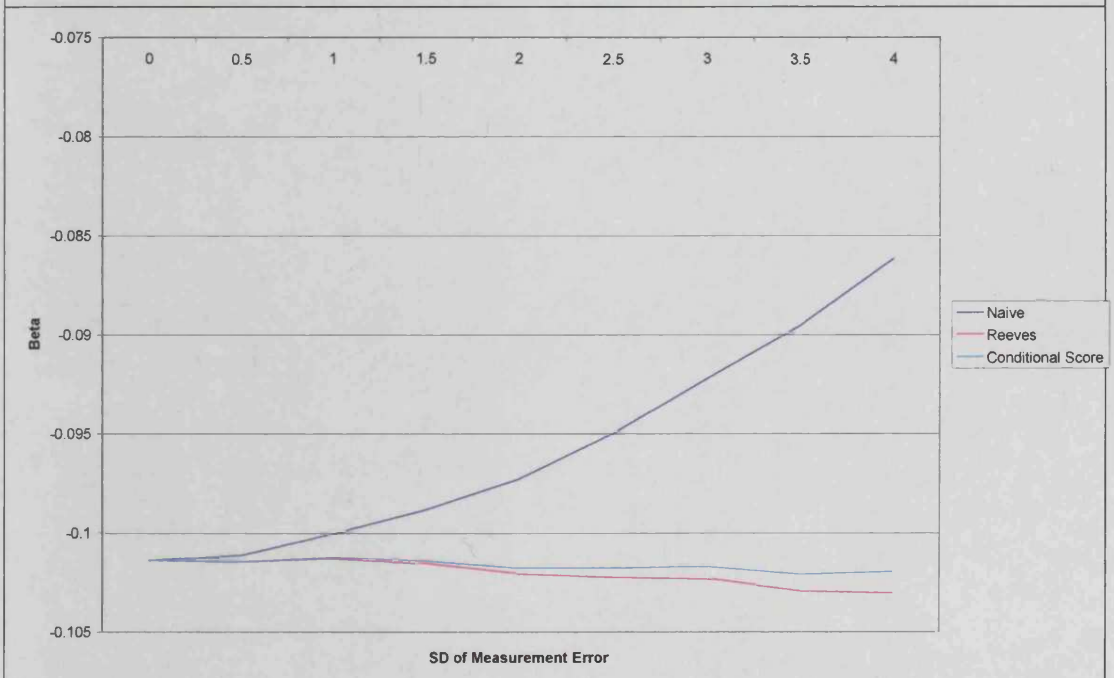


Figure 5.65: Case D Sample size 500 $\hat{\beta}_t$ method comparison

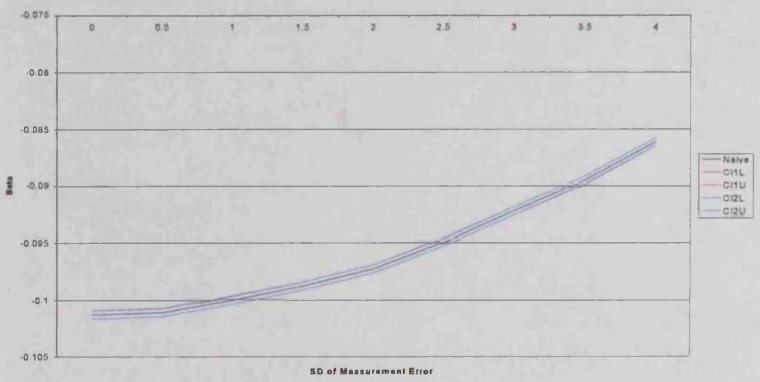


Figure 5.66: Case D Sample size 50 $\hat{\beta}_t$ Olr Confidence Interval comparison

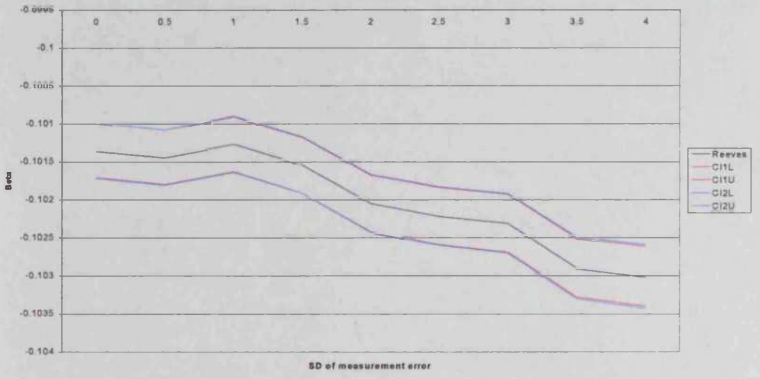


Figure 5.67: Case D Sample size 500 $\hat{\beta}_t$ Reeves Confidence Interval comparison

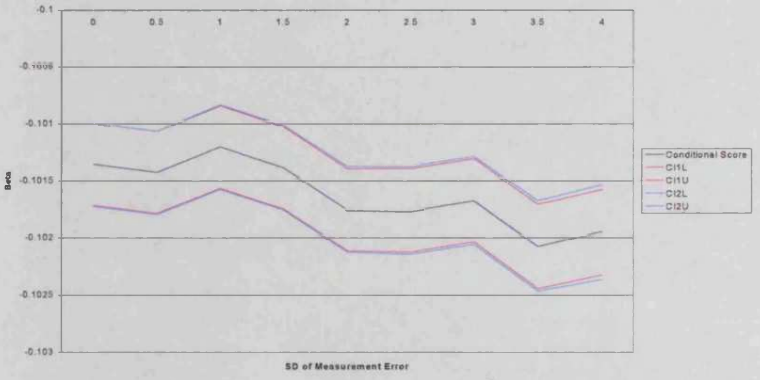


Figure 5.68: Case D Sample size 500 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	0.2647	0.07326	0.2647	0.07326	0.2647	0.07326
0.5	0.26541	0.07327	0.2644	0.07327	0.26445	0.07325
1	0.26943	0.07462	0.26538	0.07461	0.26562	0.07455
1.5	0.27342	0.07442	0.26437	0.07441	0.26489	0.07429
2	0.27902	0.07568	0.26304	0.0757	0.26397	0.07549
2.5	0.28767	0.07734	0.26296	0.07744	0.26437	0.07711
3	0.29808	0.07806	0.2629	0.07833	0.26486	0.07783
3.5	0.30879	0.07963	0.26154	0.08014	0.26408	0.07961
4	0.32275	0.08135	0.262	0.0822	0.26524	0.08151

Table 5.47:Case D Sample Size 500 OR summary results

	Olr		Reeves		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	95.35	95.4	95.36	95.4	95.56	95.5
0.5	95.4	95.26	95.32	95.29	95.23	95.14
1	95.48	95.42	95.33	95.54	94.98	95.03
1.5	95.1	94.72	94.91	95.04	95.14	95.12
2	95.14	94.44	94.87	95.01	94.78	94.96
2.5	94.38	93.88	95	95.23	94.45	94.35
3	93.11	92.22	94.58	94.85	94.55	94.75
3.5	90.93	89.97	94.4	94.95	93.8	94.12
4	87.62	86.68	94.03	94.45	93.12	93.65

Table 5.48:Case D Sample Size 500 Coverage summary results

N=1000

The results for this sample size are displayed in Figure 5.69 to Figure 5.73 and Table 5.49 to Table 5.52.

As with cases A and B, the increase in the sample size to 1000, has reduced the bias and standard errors associated with the mean of the estimates of the model parameters by all four methods.

In this case, the conditional score method produced the least biased results though there was a marginal difference between this method and the Reeves method especially when considering the standard errors associated with these mean estimates.

	Olr		Reeves		Conditional Score	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	0.00401	0.34613	0.00401	0.34629	0.00401	0.34613
0.5	-0.0035	0.34588	0.00408	0.34663	0.00244	0.34626
1	-0.0224	0.34442	0.00739	0.34688	0.00109	0.34541
1.5	-0.063	0.34224	0.00214	0.34748	-0.01104	0.34421
2	-0.1061	0.33851	0.00566	0.34745	-0.01583	0.34175
2.5	-0.1652	0.33402	0.00152	0.34744	-0.02842	0.33867
3	-0.2289	0.32932	-0.00085	0.34787	-0.03825	0.33544
3.5	-0.3027	0.32371	-0.01083	0.34776	-0.05361	0.33126
4	-0.3819	0.31809	-0.02516	0.34797	-0.0703	0.32699

Table 5.49: Case D Sample Size 1000 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Conditional Score		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	-0.1007	0.01428	0.01433	-0.1007	0.01429	0.01433	-0.1007	0.01428	0.01433
0.5	-0.1005	0.01427	0.01425	-0.1008	0.01432	0.0143	-0.1007	0.01431	0.0143
1	-0.0997	0.0142	0.01428	-0.1009	0.01439	0.0145	-0.1008	0.01433	0.01448
1.5	-0.0982	0.01409	0.01427	-0.1008	0.01451	0.01477	-0.1007	0.01437	0.01472
2	-0.0964	0.01392	0.01408	-0.101	0.01464	0.01496	-0.1007	0.0144	0.01486
2.5	-0.0941	0.01372	0.0137	-0.1011	0.01482	0.01505	-0.1007	0.01445	0.01489
3	-0.0917	0.0135	0.01368	-0.1016	0.01506	0.01561	-0.1009	0.01454	0.01539
3.5	-0.0888	0.01324	0.0133	-0.1018	0.0153	0.01589	-0.101	0.01463	0.01559
4	-0.0859	0.01298	0.01311	-0.10232	0.01562	0.01644	-0.10122	0.01477	0.01606

Table 5.50:Case D Sample Size 1000 $\hat{\beta}_t$ summary results

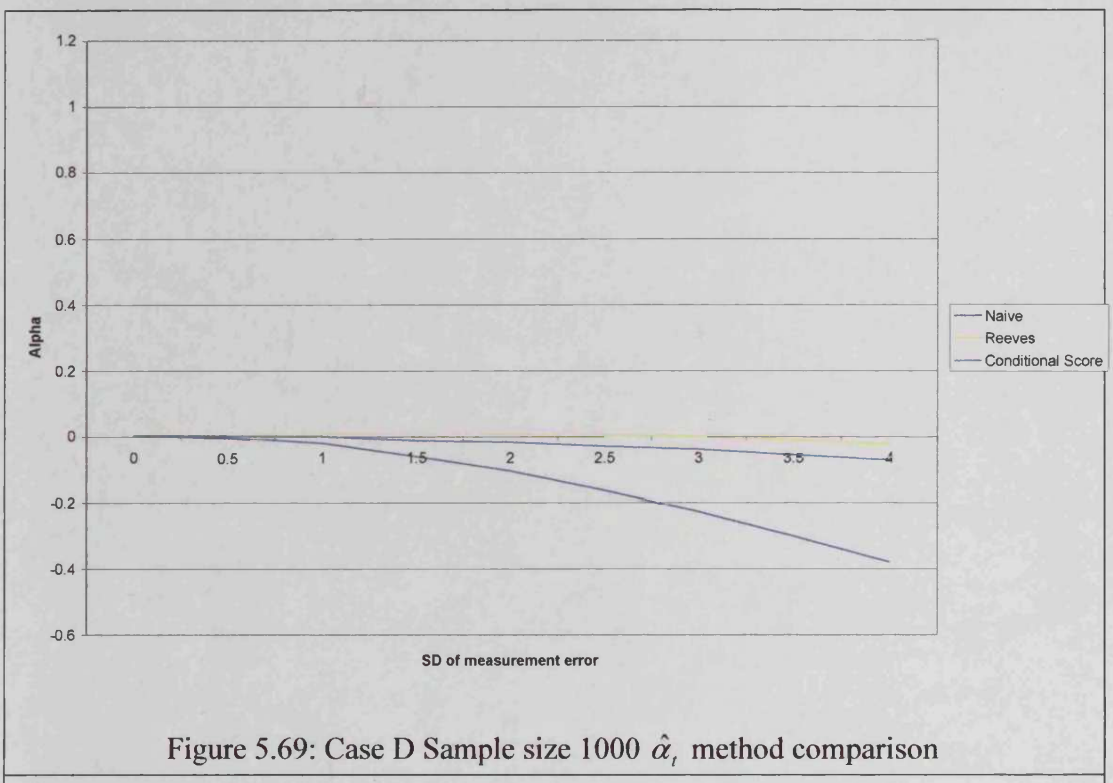


Figure 5.69: Case D Sample size 1000 $\hat{\alpha}_t$ method comparison

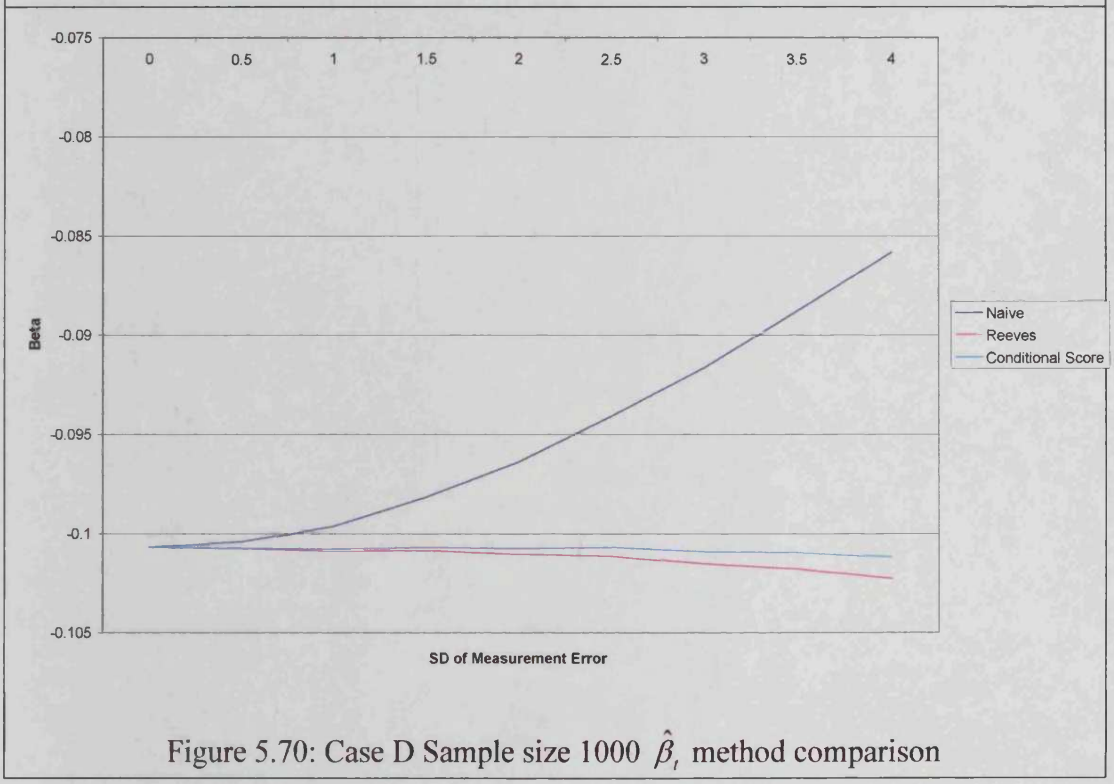


Figure 5.70: Case D Sample size 1000 $\hat{\beta}_t$ method comparison

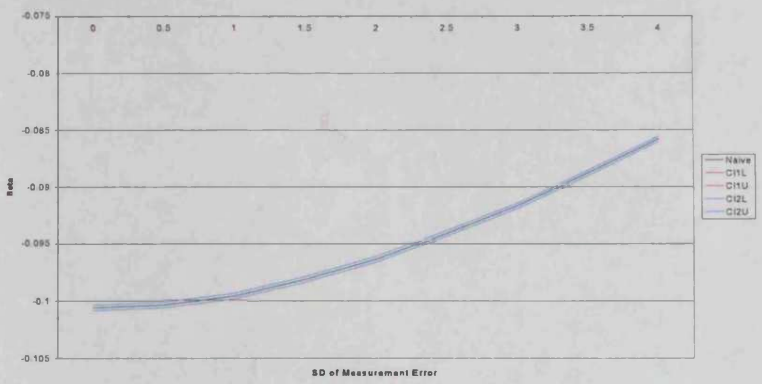


Figure 5.71: Case D Sample size 1000 $\hat{\beta}_t$ Olr Confidence Interval comparison

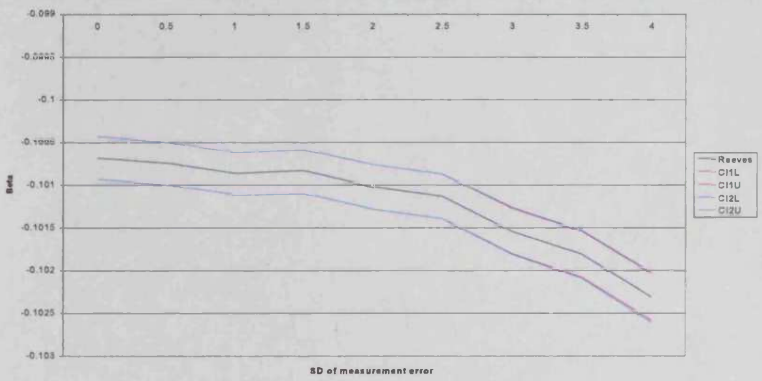


Figure 5.72: Case D Sample size 1000 $\hat{\beta}_t$ Reeves Confidence Interval comparison

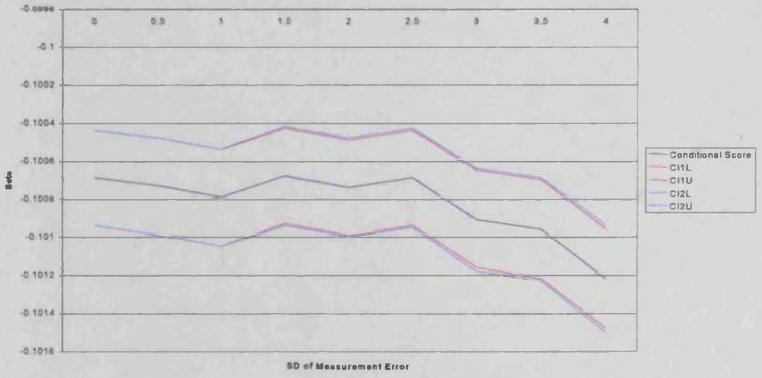


Figure 5.73: Case D Sample size 1000 $\hat{\beta}_t$ Conditional Score Confidence Interval comparison

	Olr		Reeves		Conditional Score	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	0.2619	0.0504	0.26191	0.0504	0.2619	0.0504
0.5	0.26269	0.0503	0.26167	0.05029	0.26172	0.05029
1	0.26546	0.05099	0.26139	0.05097	0.26164	0.05093
1.5	0.27075	0.05174	0.26164	0.0517	0.26219	0.05162
2	0.27718	0.05215	0.26111	0.05211	0.26207	0.05196
2.5	0.28568	0.05253	0.2608	0.05252	0.26228	0.05228
3	0.2952	0.05424	0.25977	0.05429	0.26186	0.05395
3.5	0.30658	0.05459	0.259	0.05479	0.2618	0.05438
4	0.31888	0.05614	0.25768	0.05652	0.26123	0.05597

Table 5.51:Case D Sample Size 1000 OR summary results

	Olr		Reeves		Conditional Score	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	94.9	95.03	94.91	95.05	94.91	95.05
0.5	95.12	94.93	95.04	94.94	95.35	95.28
1	94.57	94.77	94.39	94.92	95.02	95
1.5	95.1	94.65	94.99	95.15	94.81	94.87
2	94.18	94.12	95.07	95.17	94.64	94.58
2.5	92.66	91.92	94.45	94.46	94.12	94.55
3	89.85	89.42	94.42	94.85	93.68	94.13
3.5	85.25	84.83	93.95	94.5	93.45	93.8
4	78.02	78.19	93.9	94.57	92.46	93.46

Table 5.52:Case D Sample Size 1000 Coverage summary results

Conclusion for Case D

The prevalence of the disease for this case is the same as case C. The effect of the small prevalence of disease and a small sample size, is that all methods have a significant bias associated with the expected values of $\hat{\beta}_t$. This level of bias is consistent with that observed for case C.

When the sample size is small, for both cases C and D, the trend of the mean values of $\hat{\beta}_t$ for the ordinary logistic regression method, resulted in an increasing bias as σ_v is increased. As the size of the bias is larger at $\sigma_v = 0$ than observed for cases A and B, when the measurement error standard deviation is large, the bias is also large. In comparison, for the methods of Reeves and conditional score, the bias reduces as σ_v is increased. Therefore, for case D, the Reeves and conditional score methods have less bias in the expected values of $\hat{\beta}_t$ than the ordinary logistic regression method.

Comparing the standard errors and coverage terms between case C and D, the results are comparable.

When the sample size is increased to 500, the levels of bias and standard errors associated with the estimation of the model parameters are comparable with the previous cases.

In conclusion, when the prevalence of the disease is small and the sample size is small, as with all other cases, the ordinary logistic regression method produced very biased results and should not be used to estimate the relationship between the

variables in such cases. The correction methods produced less biased estimates but again they will not estimate the true relationship. If there is a medium sample size, there is little difference between the estimates of the model parameters produced by the Reeves and conditional score methods and therefore, as in all previous cases, the Reeves method is recommended.

5.4.2 Conclusion of Simulation Study

This simulation study has shown that even for no measurement error in the explanatory variable, the method of ordinary logistic regression gives estimates of β_t biased away from zero. As the measurement error standard deviation is increased, this bias becomes increasingly negative. In comparison, for the Reeves and conditional score methods', the expected values of $\hat{\beta}_t$ are only slightly positively biased.

When comparing the methods' likelihood and empirical standard errors, the ordinary logistic regression method produced the smallest values. For this method, the likelihood standard error does not take into account the measurement error associated with the explanatory variable. As a result, the standard errors are too small compared to those produced by the Reeves and conditional score methods as well as the empirical standard errors that take into account the measurement error. Therefore, the small standard errors produced by the ordinary logistic regression method do not reflect the reality and give spurious precision to the estimates of $\hat{\beta}_t$. For the Reeves and conditional score methods, the standard errors were approximately the same in all cases with the empirical standard errors slightly larger than the likelihood, though the difference is marginal.

From examination of the associated confidence intervals for each of the methods expected values of $\hat{\beta}_t$, the width of the confidence intervals were the same for all values of σ_v. In all cases, it can be concluded that the levels of bias observed are real and not due to chance.

In terms of the coverage associated with the estimation of the model parameters, for the ordinary logistic regression, the large bias associated with the expected values of $\hat{\beta}_t$ and the small standard errors, meant that the coverage in some cases when the measurement error standard deviation was large could be as low as 50%. In comparison, the Reeves method coverage for the model parameters was only reduced to approximately 90% when the measurement error standard deviation was large and the prevalence of the disease was small. The conditional score method produced the best coverage even when the measurement error standard deviation was large and the prevalence of disease was small with a low of approximately 93%.

For the expected values of the odds ratio estimates the nature of the non-linear transform could be seen for each of the methods. The result is that a negative bias will be reduced in terms of the estimate of the odds ratio whereas a positive bias will be increased. For the case of a small sample size, the negative bias associated with the ordinary logistic regression estimates of the model parameters is reduced providing reasonable estimates of the odds ratio in comparison to any correction method.

Overall, the Reeves and conditional score methods produced the least biased expected values of $\hat{\beta}_t$. For the various cases considered within this study, the methods were comparable with respect to bias and associated standard errors. As we discussed in chapter 4, the Reeves method is simpler to implement than the conditional score method. So when all the model assumptions can be met, the Reeves method would be the best to implement. The effect of when these model assumptions cannot be met on the expected values of $\hat{\beta}_t$ will be considered in chapter 6.

5.4.3 Berkson measurement error model

So far, we have only studied the methods associated with the classical measurement error model. However, as was explained throughout chapter 4, there are certain applications for which the Berkson error model is appropriate. Hence, the following investigation was conducted for Case A, to compare the methods of ordinary logistic regression, the Reeves method for the Berkson measurement error model and the Rosner approximate method explained in section 4.2.2. A further simulation for case C for a sample size of 500 is also considered. As explained in chapter 4, the Rosner method assumes that the prevalence of the disease is small and hence case C is examined, therefore, providing a fair comparison between the methods.

The data sets were generated as before except for the measurement error model.

N=100

The results for this sample size are displayed in Figure 5.74 to Figure 5.78 and Table 5.53 to Table 5.56.

For this sample size, Table 5.54, the Rosner method produces only marginally less biased estimates than the ordinary logistic regression method for the expected values of both $\hat{\alpha}_t$ and $\hat{\beta}_t$. As was seen for the classical measurement error model, the Reeves method produced the least biased expected values of $\hat{\alpha}_t$ and $\hat{\beta}_t$ for the Berkson measurement error model.

When considering the confidence interval patterns, Figure 5.76 to Figure 5.78, the scales of the axes have been changed for each method to show the detail. The same pattern and trends that were observed for the classical measurement error model are observed for the ordinary logistic regression and Reeves method. The Rosner method confidence intervals for the expected values of $\hat{\beta}_t$, also follow the same pattern as the expected values of $\hat{\beta}_t$, as σ_v increased. The width of the intervals shows that the bias levels observed for the Rosner method are real and not just due to sampling variation.

With the mean of the odds ratio estimates, the Rosner method mirrors the trend of the ordinary logistic regression method estimates. As σ_v increases, the associated standard errors are quite large for the Rosner method compared to the ordinary logistic regression method, though smaller than that for the Reeves method.

The coverage, Table 5.56, for the Rosner method is better than that of the ordinary logistic regression case, but inferior to that of Reeves. Comparing this table with Table 5.8, which is the same regression model but with the different measurement error model, the results for the Reeves method are broadly comparable.

	Olr		Reeves		Rosner	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.1348	0.8551	-3.13477	0.85514	-3.13477	0.85514
0.5	-3.132	0.8547	-3.14016	0.8563	-3.13468	0.85623
1	-3.1053	0.84924	-3.1372	0.85537	-3.11558	0.85512
1.5	-3.0455	0.84	-3.1151	0.8536	-3.06765	0.85311
2	-2.9764	0.82888	-3.09566	0.85264	-3.01336	0.85195
2.5	-2.9177	0.81749	-3.09612	0.85395	-2.97265	0.85291
3	-2.8123	0.80008	-3.0538	0.85102	-2.88352	0.85022
3.5	-2.7183	0.78313	-3.02645	0.85063	-2.80763	0.85018
4	-2.5993	0.76334	-2.97132	0.84824	-2.70215	0.84919

Table 5.53: Case A Sample Size 100 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Rosner		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10442	0.0275	0.02844	0.10442	0.0275	0.02844	0.10442	0.0275	0.02844
0.5	0.10439	0.02749	0.02857	0.10472	0.02759	0.0287	0.10448	0.02754	0.02865
1	0.10345	0.0273	0.02808	0.10477	0.02768	0.02858	0.10379	0.02751	0.02837
1.5	0.10154	0.027	0.02821	0.10445	0.02783	0.02933	0.10228	0.02746	0.02883
2	0.0992	0.02661	0.02753	0.10426	0.02808	0.02948	0.10042	0.02741	0.02856
2.5	0.09724	0.02621	0.02705	0.10501	0.02848	0.03008	0.09906	0.02745	0.02859
3	0.09371	0.02563	0.02613	0.10452	0.02882	0.03035	0.09607	0.02738	0.02813
3.5	0.09059	0.02504	0.02595	0.10487	0.02931	0.03183	0.09355	0.02739	0.02856
4	0.08663	0.02437	0.02478	0.10446	0.02979	0.03207	0.09003	0.02737	0.02779

Table 5.54:Case A Sample Size 100 $\hat{\beta}_t$ summary results

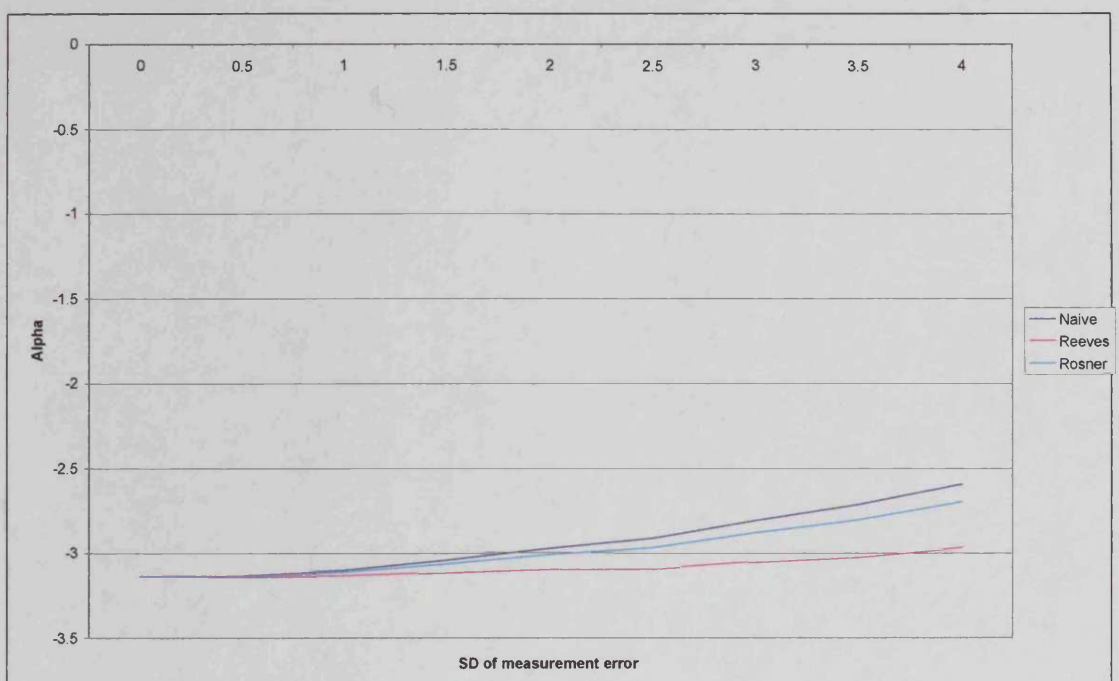


Figure 5.74: Case A Sample size 100 $\hat{\alpha}_t$ method comparison

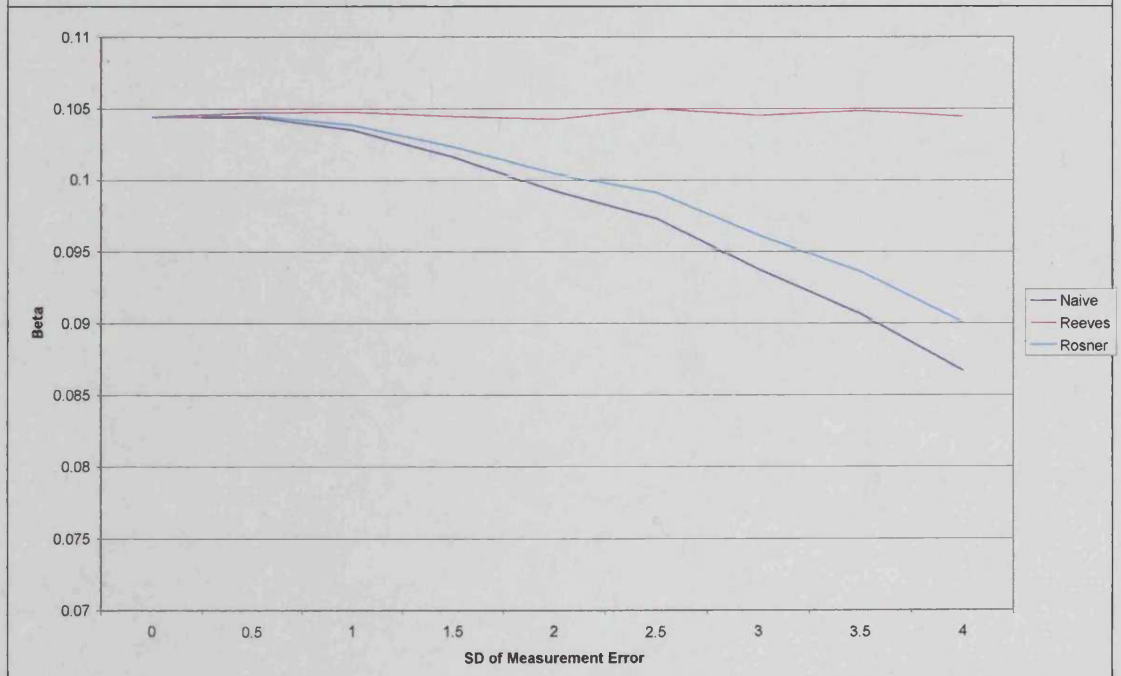


Figure 5.75: Case A Sample size 100 $\hat{\beta}_t$ method comparison

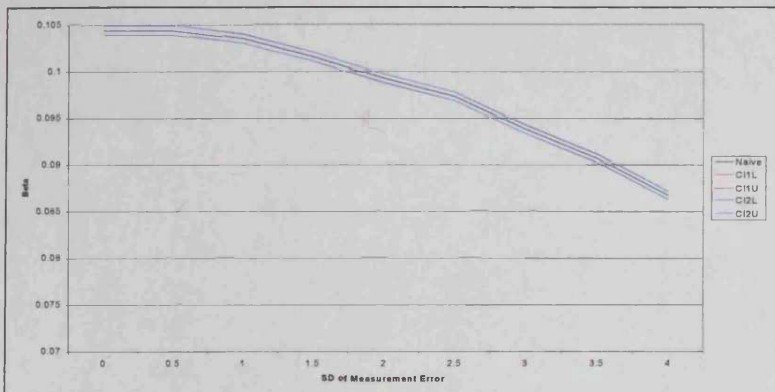


Figure 5.76: Case A Sample size 100 $\hat{\beta}_1$ Olr Confidence Interval comparison

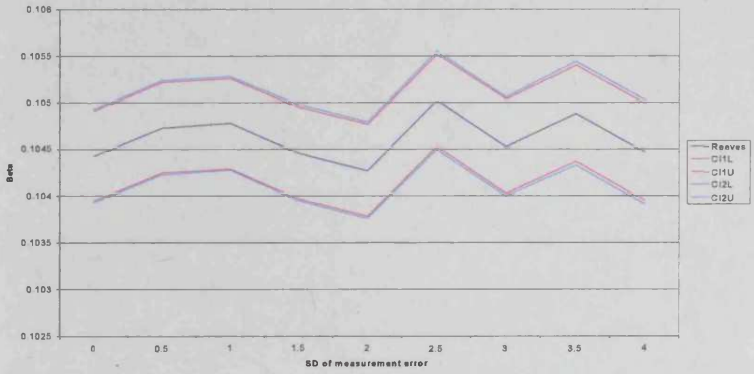


Figure 5.77: Case A Sample size 100 $\hat{\beta}_1$ Reeves Confidence Interval comparison

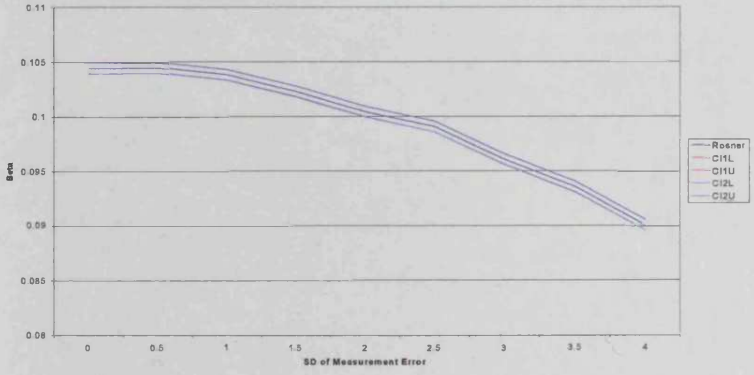


Figure 5.78: Case A Sample size 100 $\hat{\beta}_1$ Rosner Confidence Interval comparison

	Olr		Reeves		Rosner	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	4.4284	2.04893	4.42838	2.04896	4.4284	2.04893
0.5	4.42786	2.00569	4.45118	2.02926	4.43523	2.01834
1	4.36462	2.02499	4.45696	2.13123	4.39359	2.0865
1.5	4.25419	1.92922	4.456	2.14773	4.31456	2.03997
2	4.10406	1.79738	4.44862	2.18626	4.2012	1.99158
2.5	3.98593	1.69414	4.5117	2.26783	4.12642	1.95332
3	3.78256	1.53757	4.49192	2.29374	3.9547	1.84403
3.5	3.62394	1.48228	4.56806	2.71541	3.8379	1.95807
4	3.41386	1.30342	4.55103	2.77576	3.6435	1.82939

Table 5.55:Case A Sample Size 100 OR summary results

	Olr		Reeves		Rosner	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	95.51	95.48	95.51	95.5	95.51	95.5
0.5	95.52	95.57	95.49	95.58	95.5	95.59
1	95.71	95.62	95.79	95.65	95.75	95.65
1.5	95.22	95.05	95.34	95.16	95.15	95.08
2	94.35	94.38	94.92	94.98	94.68	94.65
2.5	94.32	94.23	95.25	95.4	94.81	94.78
3	93.18	93.02	95.15	95.37	94.38	94.27
3.5	91.65	91.39	95.1	95.22	93.81	93.6
4	89.18	88.62	94.18	94.4	92.54	92.4

Table 5.56:Case A Sample Size 100 Coverage summary results

N=500

The results for this sample size are displayed in Figure 5.79 to Figure 5.83 and Table 5.57 to Table 5.60.

The Rosner and Reeves methods' expected values of $\hat{\beta}_t$ follow the same pattern as n=100 but also with less bias. The pattern and trend of the associated standard errors for the ordinary logistic regression and the Reeves method are the same for the Berkson measurement error model as were observed for the classical measurement error model. The Rosner method's standard errors follow the same pattern as the ordinary logistic regression standard errors, that is, the likelihood based standard error is smaller than the empirical standard error and both decreased as σ_v increased.

In terms of the coverage, the ordinary logistic regression coverage is as low as approximately 68%. For the Rosner method, the coverage is approximately 80% for $\sigma_v = 4$. These two coverage terms display the effect of the associated biases and decreasing size of the standard errors. For the Reeves method, the coverage stays at approximately 93%.

	Olr		Reeves		Rosner	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.0242	0.36963	-3.02418	0.36965	-3.02418	0.36965
0.5	-3.0201	0.36922	-3.02771	0.36986	-3.02212	0.36982
1	-2.9898	0.36684	-3.01962	0.36932	-2.99757	0.36919
1.5	-2.9457	0.36348	-3.011	0.36896	-2.96247	0.3687
2	-2.8824	0.35878	-2.99403	0.36832	-2.91038	0.36795
2.5	-2.8074	0.35317	-2.97373	0.36774	-2.84806	0.36733
3	-2.7239	0.34672	-2.95051	0.36715	-2.7777	0.36683
3.5	-2.6357	0.33973	-2.92556	0.36673	-2.70255	0.36667
4	-2.5272	0.33164	-2.87869	0.36562	-2.60469	0.36628

Table 5.57: Case A Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Rosner		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10086	0.01188	0.01199	0.10086	0.01188	0.01199	0.10086	0.01188	0.01199
0.5	0.10063	0.01186	0.01206	0.10093	0.01189	0.01211	0.10069	0.01188	0.01209
1	0.09959	0.01178	0.01192	0.10078	0.01193	0.01211	0.09985	0.01186	0.01201
1.5	0.09819	0.01166	0.01159	0.10082	0.012	0.012	0.09875	0.01185	0.01179
2	0.09605	0.0115	0.01151	0.10062	0.01209	0.01223	0.09698	0.01182	0.01185
2.5	0.09357	0.01131	0.01137	0.10052	0.01221	0.01248	0.09492	0.0118	0.01187
3	0.09075	0.01109	0.0112	0.10048	0.01236	0.01279	0.09254	0.01179	0.01186
3.5	0.08784	0.01085	0.01097	0.10066	0.01255	0.01311	0.09006	0.01179	0.01181
4	0.08424	0.01057	0.01058	0.10029	0.01274	0.01328	0.08681	0.01178	0.01156

Table 5.58:Case A Sample Size 500 $\hat{\beta}_t$ summary results

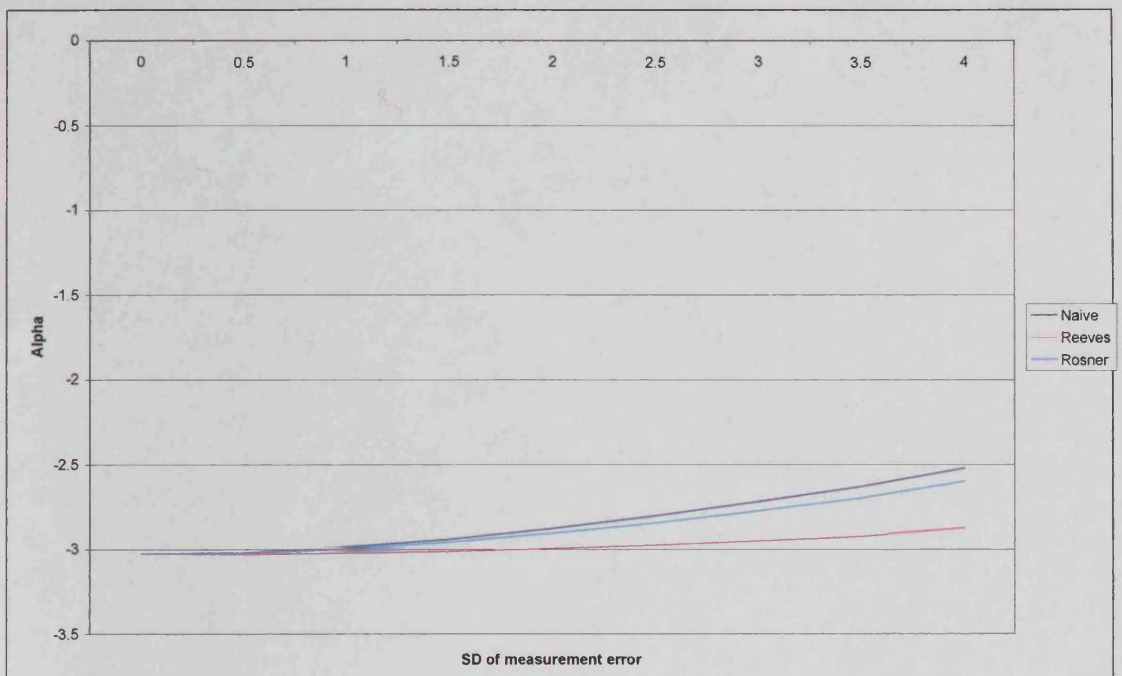


Figure 5.79: Case A Sample size 500 $\hat{\alpha}_i$ method comparison

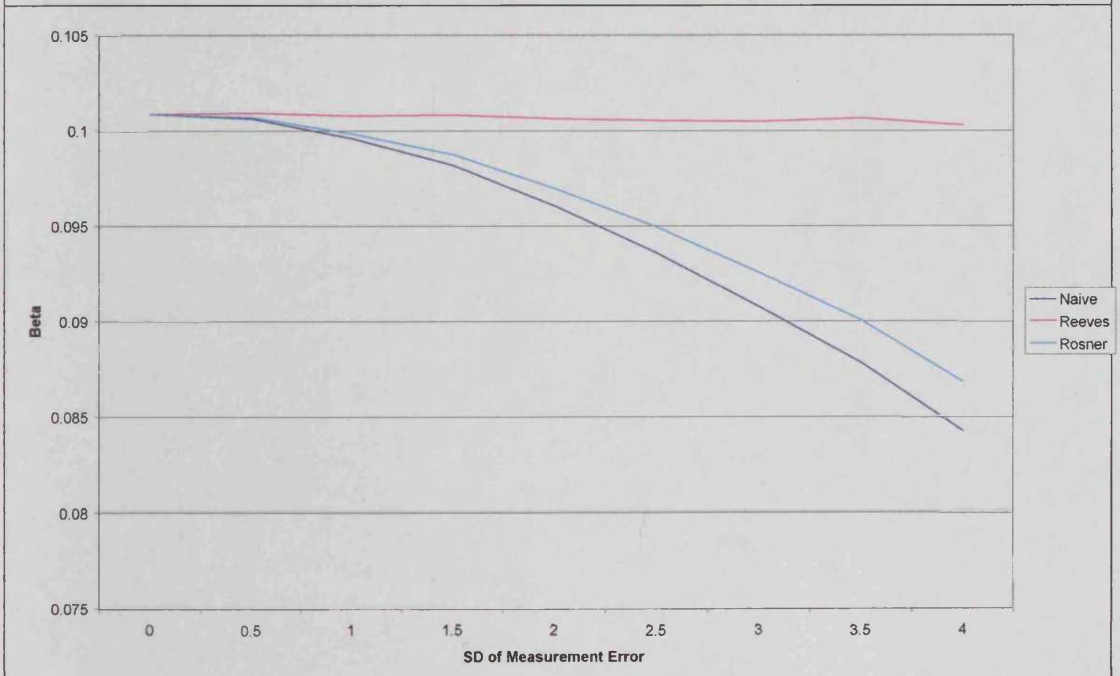


Figure 5.80: Case A Sample size 500 $\hat{\beta}_i$ method comparison

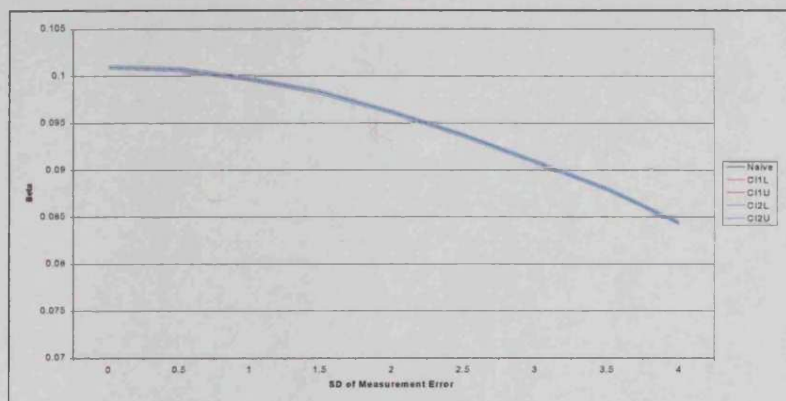


Figure 5.81: Case A Sample size 500 $\hat{\beta}_t$ Olr Confidence Interval comparison

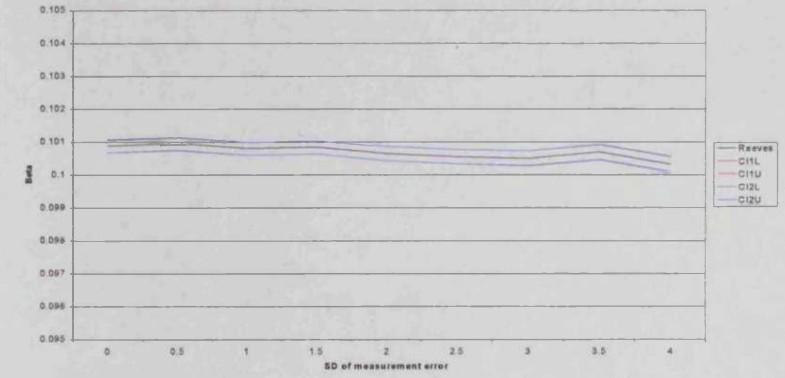


Figure 5.82: Case A Sample size 500 $\hat{\beta}_t$ Reeves Confidence Interval comparison

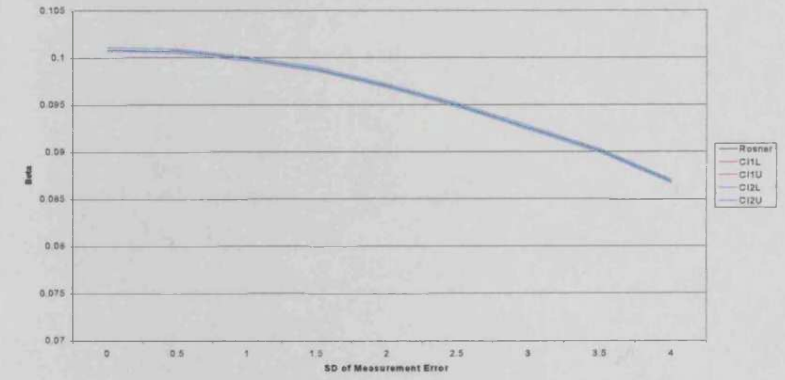


Figure 5.83: Case A Sample size 500 $\hat{\beta}_t$ Rosner Confidence Interval comparison

	Olr		Reeves		Rosner	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.95014	0.65389	3.95014	0.65389	3.95014	0.65389
0.5	3.93862	0.65758	3.955	0.66309	3.94237	0.65963
1	3.88279	0.6384	3.94705	0.65967	3.89712	0.64615
1.5	3.80734	0.60822	3.94833	0.65412	3.83774	0.62442
2	3.6985	0.58908	3.94011	0.66921	3.74803	0.61609
2.5	3.57539	0.5607	3.93699	0.68167	3.64516	0.59888
3	3.44106	0.53149	3.93753	0.7006	3.53011	0.58089
3.5	3.3069	0.49992	3.95021	0.72177	3.41346	0.55962
4	3.14775	0.45904	3.93245	0.7309	3.26534	0.52508

Table 5.59:Case A Sample Size 500 OR summary results

	Olr		Reeves		Rosner	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	95.47	95.48	95.47	95.48	95.47	95.48
0.5	95.22	95.06	95.19	95.12	95.22	95.08
1	94.74	94.62	94.91	94.78	94.75	94.73
1.5	94.53	94.42	94.9	95	94.69	94.65
2	93.19	92.97	94.77	94.81	93.66	93.81
2.5	90.08	89.9	94.35	94.67	92.03	91.9
3	85.76	85.13	94.19	94.52	89.55	89.38
3.5	78.26	76.78	93.57	94.41	85.42	84.82
4	68.75	66.42	92.45	94.08	79.33	78.47

Table 5.60:Case A Sample Size 500 Coverage summary results

N=1000

The results for this sample size are displayed in Figure 5.84 to Figure 5.88 and Table 5.61 to Table 5.64.

Again, Table 5.62 shows that for each of the methods the bias in the expected values of $\hat{\beta}_t$ and associated standard errors has been reduced with the increased sample size though the pattern and trend are as before.

As before, the summary estimate of the standard errors is smaller for the method of ordinary logistic regression than the other two methods. In previous cases the method by Reeves produced the largest summary estimate for the standard errors as the measurement error was increased. In this case, for each of the sample sizes the method by Rosner produced the largest estimate for the summary estimate of the standard errors. For this sample size, the difference between the method of ordinary logistic regression and the method by Reeves could be up to as much as 10%.

The coverage for the Rosner method can go as low as approximately 63%, though this is better than the ordinary logistic regression with a low of approximately 44%, the Reeves method still produced the best coverage with a low of 90%.

	Olr		Reeves		Rosner	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.0164	0.26048	-3.01642	0.26049	-3.01642	0.26049
0.5	-3.003	0.25981	-3.01048	0.26025	-3.00488	0.26023
1	-2.9758	0.25829	-3.00534	0.26002	-2.98328	0.25993
1.5	-2.9328	0.25598	-2.99753	0.25981	-2.94898	0.25962
2	-2.8711	0.25269	-2.98182	0.25936	-2.89814	0.25909
2.5	-2.7944	0.24869	-2.95944	0.25887	-2.83364	0.25858
3	-2.7125	0.24418	-2.93725	0.25843	-2.76443	0.2582
3.5	-2.6223	0.23925	-2.90971	0.25808	-2.68662	0.25807
4	-2.5171	0.23362	-2.86577	0.25732	-2.5918	0.2578

Table 5.61: Case A Sample Size 1000 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Rosner		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10053	0.00837	0.0085	0.10053	0.00837	0.0085	0.10053	0.00837	0.0085
0.5	0.10007	0.00834	0.00837	0.10036	0.00837	0.0084	0.10013	0.00836	0.00838
1	0.09915	0.00829	0.00833	0.10032	0.0084	0.00846	0.09939	0.00835	0.00839
1.5	0.09775	0.00821	0.00819	0.10035	0.00845	0.00848	0.09829	0.00834	0.00833
2	0.09569	0.0081	0.00809	0.1002	0.00851	0.00859	0.09659	0.00832	0.00832
2.5	0.09315	0.00796	0.00793	0.10002	0.00859	0.00869	0.09445	0.00831	0.00826
3	0.0904	0.00781	0.00779	0.09999	0.0087	0.00888	0.09212	0.0083	0.00824
3.5	0.08742	0.00764	0.00765	0.10005	0.00882	0.00911	0.08955	0.0083	0.00821
4	0.08392	0.00744	0.00747	0.09975	0.00896	0.00934	0.08639	0.00829	0.00814

Table 5.62:Case A Sample Size 1000 $\hat{\beta}_t$ summary results

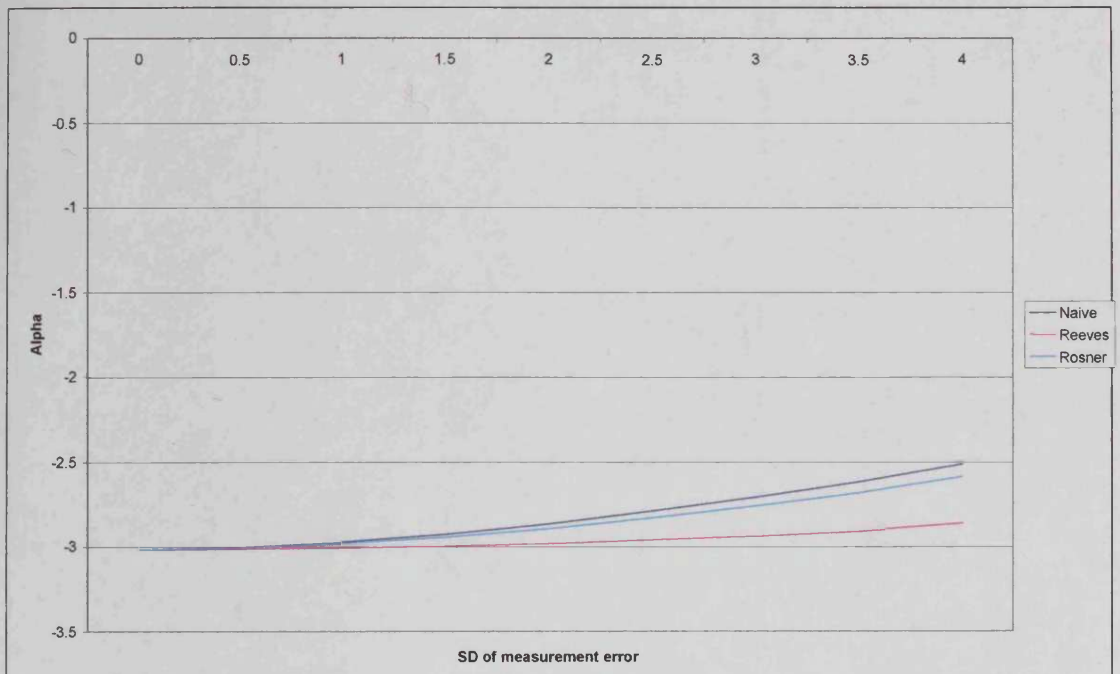


Figure 5.84: Case A Sample size 1000 $\hat{\alpha}_t$ method comparison

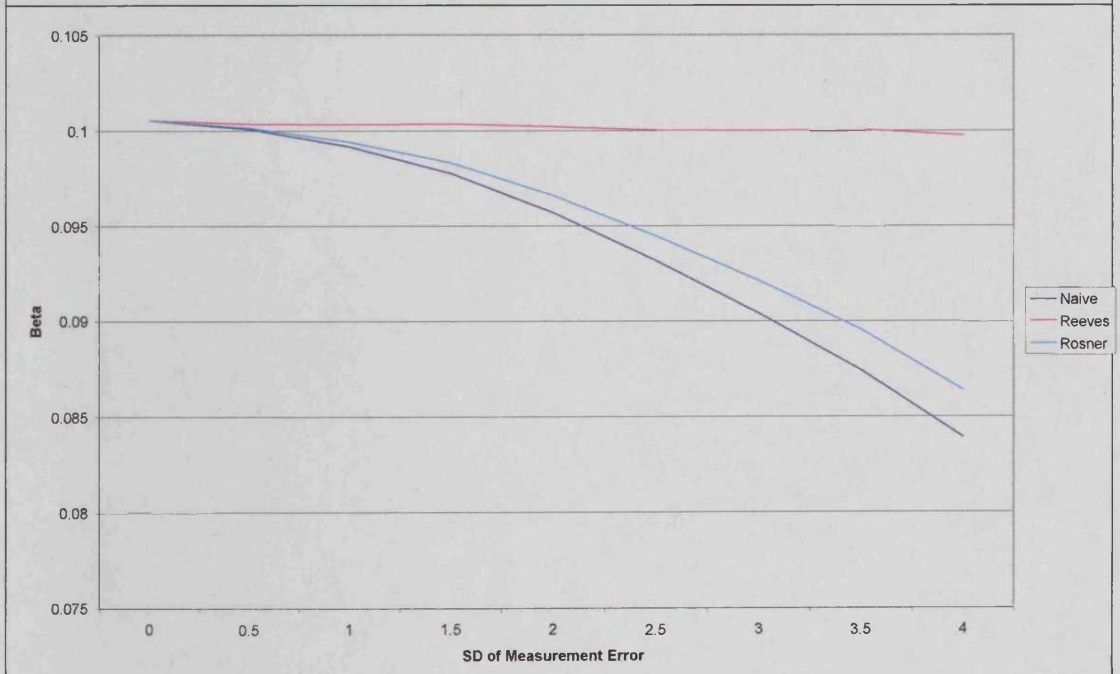


Figure 5.85: Case A Sample size 1000 $\hat{\beta}_t$ method comparison

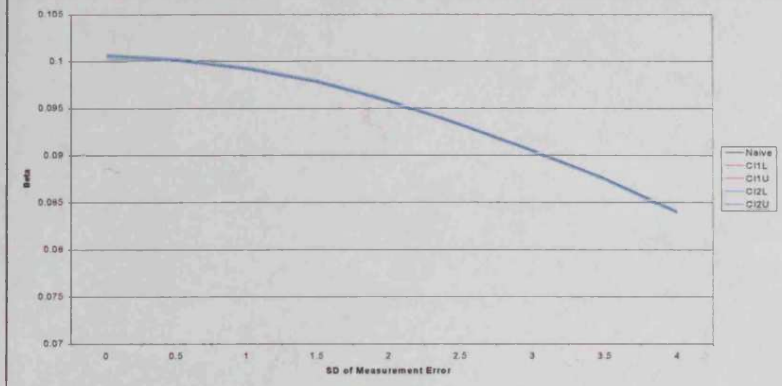


Figure 5.86: Case A Sample size 1000 $\hat{\beta}_t$ Olr Confidence Interval comparison

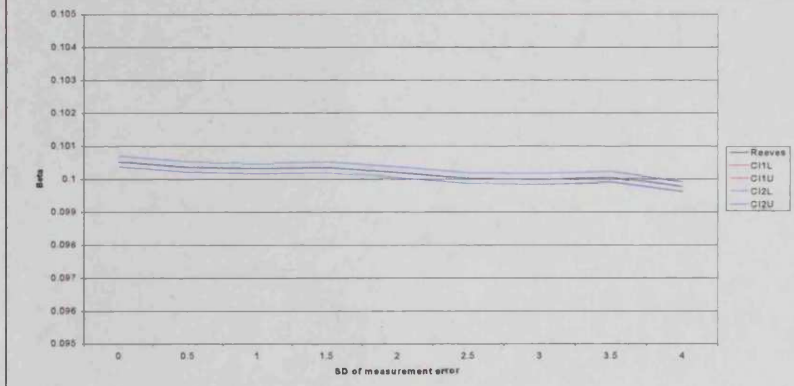


Figure 5.87: Case A Sample size 1000 $\hat{\beta}_t$ Reeves Confidence Interval comparison

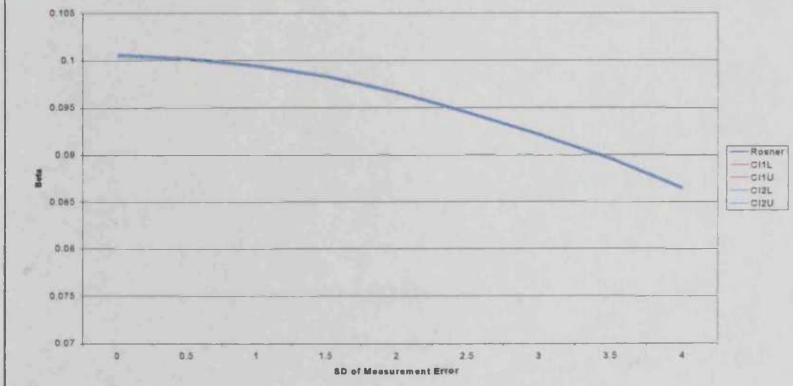


Figure 5.88: Case A Sample size 1000 $\hat{\beta}_t$ Rosner Confidence Interval comparison

	Olr		Reeves		Rosner	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	3.90694	0.45357	3.90694	0.45356	3.90694	0.45357
0.5	3.88191	0.44427	3.89761	0.44783	3.88535	0.44556
1	3.83359	0.43535	3.89539	0.44939	3.84687	0.44028
1.5	3.76142	0.42031	3.89726	0.45095	3.78963	0.43077
2	3.65745	0.40285	3.89024	0.45548	3.70347	0.41992
2.5	3.53353	0.38169	3.88123	0.46099	3.59807	0.40581
3	3.40414	0.36142	3.88114	0.47128	3.48656	0.39249
3.5	3.26924	0.34034	3.8857	0.48398	3.36746	0.37763
4	3.11779	0.3173	3.87153	0.49579	3.22681	0.3593

Table 5.63:Case A Sample Size 1000 OR summary results

	Olr		Reeves		Rosner	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	94.93	94.98	94.94	94.98	94.94	94.98
0.5	95.11	95.18	95.15	95.08	95.14	95.14
1	94.9	94.78	95.23	95.17	95.11	94.9
1.5	93.84	93.73	95.04	94.95	94.28	94.28
2	90.78	90.43	94.52	94.73	92.4	92.17
2.5	85.51	84.58	94.08	94.48	88.82	88.56
3	76.75	75.02	93.55	94.58	83.79	83.21
3.5	63.41	60.54	92.44	94.38	75.66	74.16
4	46	42.32	90.6	93.92	64.21	61.85

Table 5.64:Case A Sample Size 1000 Coverage summary results

Case C $\alpha_t = -6$ and $\beta_t = 0.1$

For this case the prevalence of the disease is small and therefore fits with the method assumption made by Rosner et al. A sample size of 500 has been chosen to provide a representative case.

N=500

With the smaller prevalence of disease, the expected values of $\hat{\beta}_t$ from the Rosner method, are approximately the same as those produced by the ordinary logistic regression method. This means that the expected values are negatively biased as σ_v is increased. The likelihood and empirical standard errors are slightly larger than the those produced by the ordinary logistic regression method, with the width of the associated confidence intervals showing that the bias is real and not due to chance. For the coverage terms, an approximate 93% low is obtained for the Rosner method compared to the approximate 80% and 93% achieved by the ordinary logistic regression and Reeves method.

	Olr		Reeves		Rosner	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-6.065	0.78854	-6.06501	0.78902	-6.06501	0.78902
0.5	-6.0692	0.78885	-6.07685	0.78991	-6.07063	0.79094
1	-6.0506	0.78729	-6.08069	0.79014	-6.05613	0.79428
1.5	-5.9994	0.78034	-6.0653	0.78624	-6.01153	0.79536
2	-5.934	0.77122	-6.04702	0.78141	-5.95466	0.79723
2.5	-5.8419	0.76018	-6.00993	0.77596	-5.87247	0.79997
3	-5.7683	0.75005	-5.99871	0.7725	-5.81029	0.80646
3.5	-5.6554	0.73557	-5.94918	0.76615	-5.7086	0.8105
4	-5.5564	0.72198	-5.91529	0.76191	-5.62126	0.81768

Table 5.65: Case C Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr			Reeves			Rosner		
σ_v	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_L(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.10112	0.02042	0.02042	0.10112	0.02043	0.02042	0.10112	0.02043	0.02042
0.5	0.10119	0.02041	0.02055	0.1015	0.02047	0.02063	0.1012	0.02048	0.02055
1	0.10052	0.02036	0.02073	0.10173	0.02056	0.02106	0.10054	0.02061	0.02074
1.5	0.09909	0.02018	0.02046	0.10179	0.0206	0.02119	0.09915	0.02071	0.02048
2	0.09723	0.01993	0.02008	0.10194	0.02067	0.02138	0.09732	0.02086	0.02012
2.5	0.09454	0.01963	0.01992	0.10169	0.02077	0.02192	0.09467	0.02106	0.01998
3	0.09233	0.01934	0.01961	0.10242	0.02095	0.02249	0.09252	0.02137	0.01969
3.5	0.08906	0.01895	0.01918	0.10231	0.02112	0.02303	0.08929	0.02167	0.01928
4	0.08612	0.01857	0.01867	0.10289	0.02137	0.02358	0.08639	0.02206	0.01881

Table 5.66:Case C Sample Size 500 $\hat{\beta}_t$ summary results

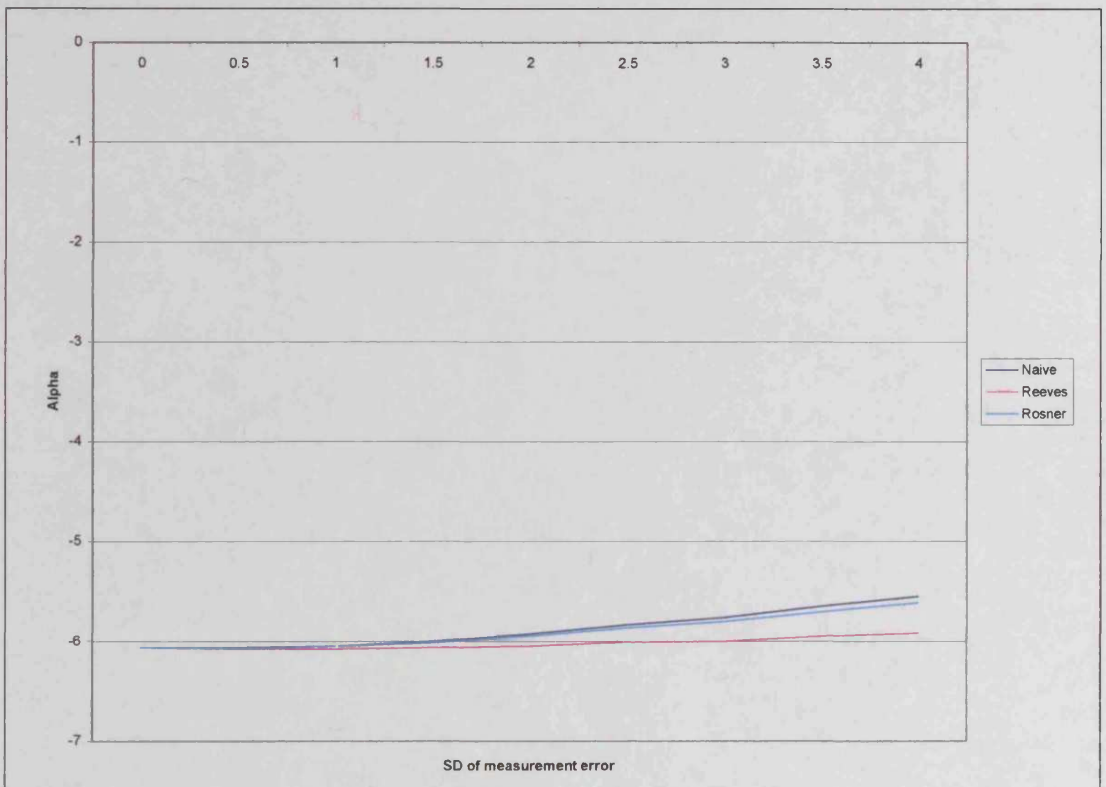


Figure 5.89: Case C Sample size 500 $\hat{\alpha}_t$ method comparison

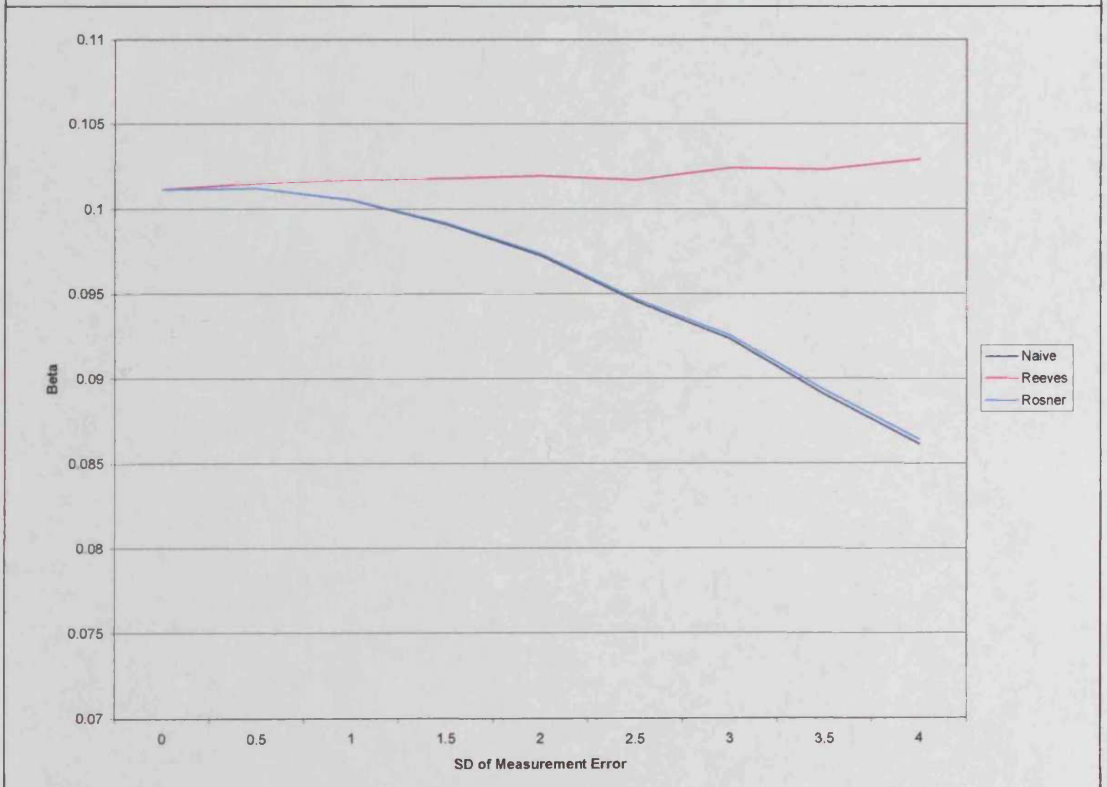


Figure 5.90: Case C Sample size 500 $\hat{\beta}_t$ method comparison

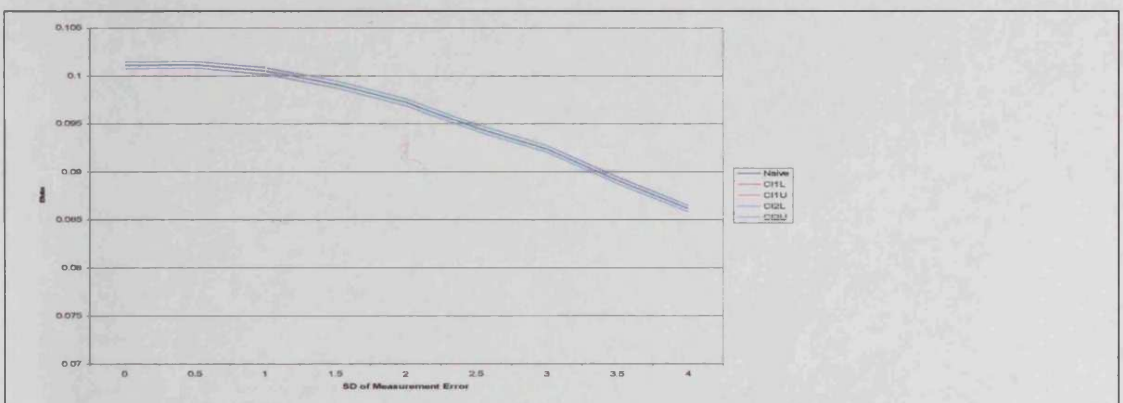


Figure 5.91: Case C Sample size 500 $\hat{\beta}_t$ Olr Confidence Interval comparison

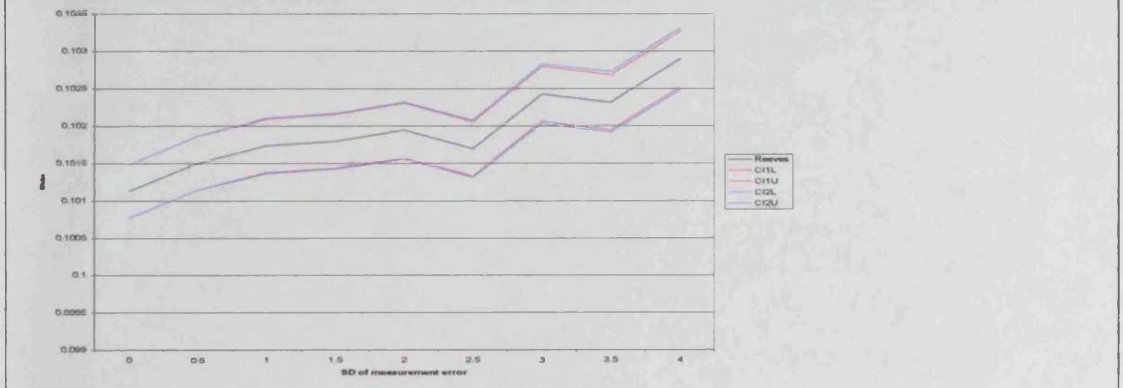


Figure 5.92: Case C Sample size 500 $\hat{\beta}_t$ Reeves Confidence Interval comparison

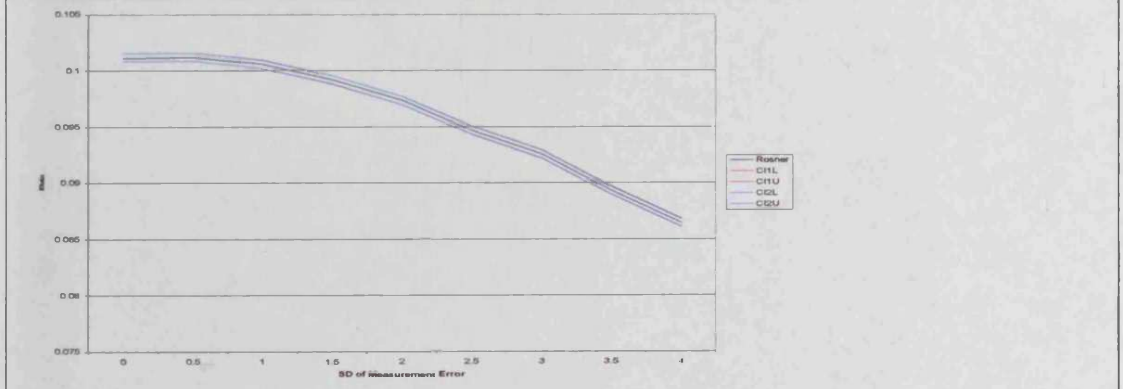


Figure 5.93: Case C Sample size 500 $\hat{\beta}_t$ Rosner Confidence Interval comparison

	Olr		Reeves		Rosner	
σ_v	OR	$SE(OR)$	OR	$SE(OR)$	OR	$SE(OR)$
0	4.06672	1.18458	4.06672	1.18459	4.06672	1.18458
0.5	4.07271	1.19259	4.09086	1.20349	4.07309	1.19288
1	4.03927	1.20503	4.11167	1.25028	4.04079	1.20616
1.5	3.95812	1.16897	4.11716	1.27145	3.96136	1.17139
2	3.85455	1.11467	4.12908	1.28762	3.85992	1.11869
2.5	3.71399	1.04976	4.12344	1.304	3.72159	1.05545
3	3.60173	1.01106	4.17622	1.38332	3.61178	1.01875
3.5	3.44028	0.93597	4.17958	1.41767	3.45228	0.94511
4	3.30025	0.8702	4.22269	1.47675	3.3141	0.88073

Table 5.67:Case C Sample Size 500 OR summary results

	Olr		Reeves		Rosner	
σ_v	α_t	β_t	α_t	β_t	α_t	β_t
0	95.35	95.53	95.35	95.53	95.35	95.53
0.5	95.18	95.34	95.19	95.32	95.26	95.41
1	94.95	95.14	94.99	95.11	95.09	95.35
1.5	94.9	95.11	95.08	95.07	95.21	95.63
2	94.54	94.72	94.96	94.81	95.21	95.72
2.5	93.02	93.15	94.25	94.39	94.5	95.02
3	91.78	91.98	93.98	94.16	93.96	94.67
3.5	89.9	89.72	93.53	93.22	92.95	93.62
4	87.23	85.99	93.1	93.28	91.82	92.58

Table 5.68:Case C Sample Size 500 Coverage summary results

Conclusion for Berkson Measurement Error Model

This simulation study has shown that when comparing the ordinary logistic regression, Reeves and Rosner methods, the best correction method is that of the Reeves method.

For both case A, and case C where the prevalence of the disease is small, which is an assumption made by the Rosner method, the bias associated with the expected values of $\hat{\beta}_t$ as well as the size of the standard errors, were in the same order of magnitude as those observed for the ordinary logistic regression method. Therefore, even when the method assumption of a small prevalence of disease holds, the Rosner method is not a recommended correction method.

5.5 Conclusion

This chapter has given an overview of the results in estimating the logistic regression measurement error model parameters by some of the current methods that are available to correct for measurement error.

The examples of relating blood pressure to coronary heart disease and fasting plasma glucose to retinopathy status, showed that without implementing a correction method it is not obvious what the effect that measurement error in the risk factor can have on the estimates of the relationship between the risk factor and the disease status produced by the ordinary logistic regression method. Further to this, it cannot be concluded from those examples which correction factor method should be used, as the true values for the model parameters are unknown. However, these examples showed that when a risk factor is known to be measured with error and there is an estimate of the distribution of this error, then some form of correction method should be used. This leaves the question of which correction method should be used taking into account the sample size, the prevalence of the disease and the size of the measurement error standard deviation. The simulation study within this chapter aimed to answer some of these questions.

For all the cases that were considered in the simulation study, the expected values of $\hat{\beta}_t$ produced by the ordinary logistic regression method were biased as the measurement error standard deviation increased. In general the ordinary logistic regression method will not correctly estimate the relationship between the explanatory variable and the disease status when there is measurement error associated with the explanatory variable.

The method of optimising the logistic log likelihood proved to only work in certain cases. That is, if the sample size was small or the starting values were far from the true values, then the method would not converge and hence the method produced no estimates for the model parameters. In general, when the optimisation method would work, the estimates were biased, with the bias becoming increasingly negative as σ_v was increased. When the measurement error standard deviation was small, the optimisation method produced the least biased results with the smallest associated standard errors. When the measurement error standard deviation was large then the other two correction methods produced the most viable results.

Throughout the simulation study the methods of Reeves and Conditional Score produced comparable results. The Reeves method also produced the least biased results when the relationship between the observed and true values followed a Berkson measurement error model. The Reeves method is a simple correction procedure that requires estimates of the model parameters from a standard statistical package and full specification of the data distributions. The Conditional Score method is somewhat more involved. A specific program must be written and implemented and cannot be achieved using a standard statistical package. The simulation study has shown there is little difference in terms of bias, size of standard errors and coverage, between these two methods when X is Normally distributed. Therefore, when the model assumptions hold, the Reeves method has proven to be the easier method to implement as well as producing comparable results with other correction methods.

The simulation study exercise has shown that the Reeves and conditional score methods are comparable in the cases that we have investigated, with the best method being that of the Reeves method, as it is simpler to implement than the conditional score method. However, this simulation study was based on X following a normal distribution. How robust these methods are to departures from these model assumptions is investigated in chapter 6.

Chapter 6

6 Simulation Study to investigate the effects of the methods model assumptions in various settings

6.1 Introduction

In chapter 5 methods for estimating the model parameters for the logistic regression measurement error problem were compared in cases where the model assumptions held. In this chapter a simulation exercise is again used to compare the methods when these model assumptions are not true, providing guidance on which method is appropriate to use in which cases.

One of the model assumptions made by the Reeves and optimization methods is that X follows a normal distribution. There could be many cases where this is not the case,

such as when an experiment is actually designed; then the values of X may have a uniform distribution or even be concentrated near the extremes. For example if X follows a Chi-square distribution, what effect does this have on the estimates of the model parameters by the various methods considered.

Throughout chapter 5, it was assumed that the measurement error variance was known. In practice, the measurement error variance is estimated from a subsidiary or validation study and therefore the estimate can be subject to variation. Again, what effect does this have on the methods estimates of the model parameters?

The aim of this chapter is to see the effect on the methods considered in chapter 5. The estimates of the model parameters and associated standard errors will be compared for the two situations explained above, namely:

- The case where X does not follow a Normal distribution.
- The case where the variance of the measurement error is estimated from a validation study.

This simulation study is designed to look at the effect on the estimators and to see whether the breaking of such model assumptions makes a difference on how well the methods can estimate the true model parameters.

6.2 Case where X does not follow a Normal Distribution

6.2.1 Introduction

In many applications it cannot always be assumed that the true explanatory variable follows a Normal distribution. It is often convenient to make such an assumption

about the data involved in practical situations, as many methods require this model assumption in order to estimate the parameters. If the model assumption does not hold, then more complex methods may be required.

When considering which method to use to estimate the model parameters in the logistic regression measurement error problem, the effect that a Non-Normal covariate could have on the results must be taken into account. In this comparison study, all the methods, except for the conditional-score method, assume that X is Normally distributed. To investigate the best method in this situation, a Chi squared distribution with equivalent variation to the previously used Normal distribution will be investigated. The methods of ordinary logistic regression, Reeves and conditional score will be compared. As the Reeves and conditional score methods produced the best results from chapter 5, they are the only correction methods that are considered within this investigation.

6.2.2 Simulation Study

To investigate and to compare the methods from chapter 5 where the assumption that X is normally distributed is not true, a simulation study was conducted according to the following model.

7) Sample size N, was chosen to be 100, 500 and 1000 respectively.
8) The true X values were assumed to follow a scaled chi-squared distribution. Two were used, with 5 and 20 degrees of freedom respectively, and these were scaled linearly to have a mean of 30 and standard deviation of 10 to make them

comparable with the normal distributions used in Chapter 5. These distributions are displayed in Table 6.1.

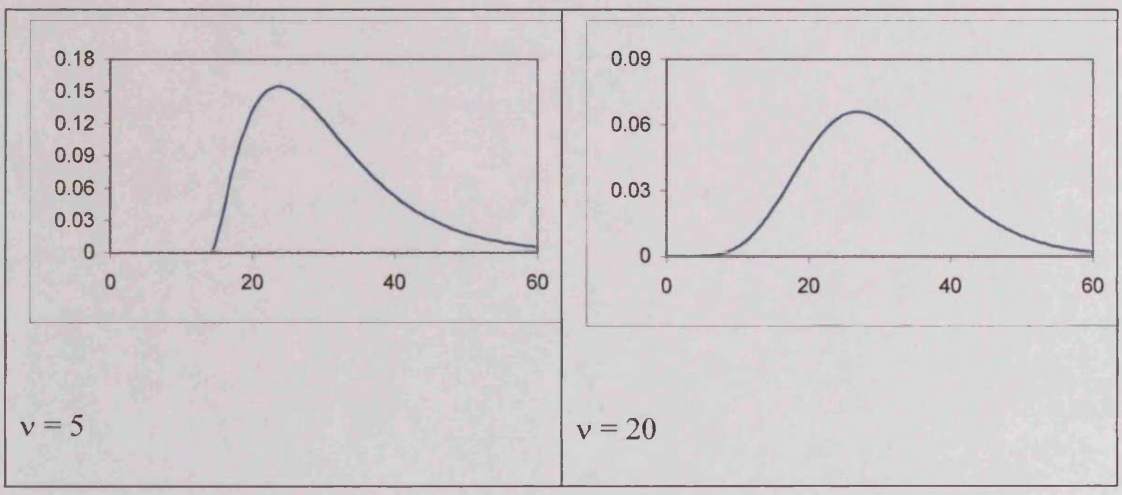


Table 6.1: Scaled chi-squared distributions

9) The observed Z values were generated according to a classical measurement error model, $Z = X + V$, where V was also assumed to be Normally distributed $V \sim (0, \sigma_v^2)$

10) The disease status Y was generated conditional on X for the case where $\alpha_t = -3$ and $\beta_t = 0.1$, prevalence of disease 0.5 and $Y = 1$ with probability

$$P(Y = 1|X) = \frac{\exp(\alpha_t + \beta_t X)}{1 + \exp(\alpha_t + \beta_t X)}$$

11) The measurement error standard deviation σ_v ranged from 0 to 4, with steps of 0.5.

12) Each simulation was run 13000 times.

When studying the effects of measurement error in the explanatory variable the parameter of interest is the associated regression coefficient; in the simple case this corresponds to the parameter β_t. Therefore, the following study will look at the

effects of measurement error on estimating β_t, that is, how well each method estimates β_t as σ_v is increased.

The results for each sample size and degrees of freedom are displayed in the following form:

- Table displaying the expected value for $\hat{\beta}_t$ and the associated empirical standard error, for both the cases where X follows a normal distribution and where X follows the relevant chi-squared distribution.
- Table displaying the coverage terms for $\hat{\beta}_t$, for both the cases where X follows a normal distribution and where X follows the relevant chi-squared distribution.

N=100 Degrees of Freedom=5

In chapter 5, for all cases when $\sigma_v = 0$, all the methods had expected values of $\hat{\beta}_t$ that were positively biased. For this case, Table 6.1 shows that the same is true when $\sigma_v = 0$.

As σ_v increased, for the ordinary logistic regression method, the expected values of $\hat{\beta}_t$ became negatively biased, with the bias slightly larger at $\sigma_v = 4$ than for the same case where X is normally distributed. When $\sigma_v = 0$, the estimates from the Reeves method show that the expected values of $\hat{\beta}_t$ had a positive bias that reduced slightly as σ_v was increased. For the conditional score method, the positive bias increased as σ_v increased, with the bias generally being larger than that observed when X is normally distributed.

In terms of the empirical standard errors, when $\sigma_v = 0$, the standard errors are approximately the same size as those observed previously. As σ_v increased, the ordinary logistic regression standard errors reduced in size, which again were approximately the same size as when X was normally distributed. For the Reeves method, the same was true in terms of the size of the standard errors, however, they increased in size as σ_v increased. The results from the conditional score method show that the standard errors are marginally larger than those previously observed.

In terms of the coverage by each of the methods, Table 6.3, the ordinary logistic regression method reached a low of 87% which is less than the case where X is

normally distributed. For the Reeves and conditional score methods both maintained a coverage level of approximately 95%.

In comparison, for this case, the Reeves results show that these estimates are the least biased in comparison to the conditional score method, despite X not being normally distributed. As expected, for the ordinary logistic regression method, the change in the distribution of the X values has not affected the estimates though they are still very biased in comparison to the other methods.

	OLR $X \sim Normal$		OLR $X \sim ChiSquared$		Reeves $X \sim Normal$		Reeves $X \sim ChiSquared$		Conditional Score $X \sim Normal$		Conditional Score $X \sim ChiSquared$	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10443	0.02858	0.10419	0.02981	0.10443	0.02858	0.10419	0.02981	0.10443	0.02858	0.10419	0.02981
0.5	0.1041	0.0283	0.1042	0.02944	0.10443	0.02842	0.10453	0.02957	0.10445	0.02843	0.1046	0.02961
1	0.10295	0.02838	0.1029	0.02997	0.10426	0.02887	0.10423	0.0305	0.10431	0.02891	0.10449	0.03065
1.5	0.10129	0.02778	0.10066	0.02916	0.10419	0.02888	0.10358	0.03033	0.10432	0.02897	0.10415	0.03067
2	0.09884	0.0271	0.0985	0.02836	0.10387	0.02901	0.10358	0.03037	0.10409	0.02917	0.10459	0.031
2.5	0.09717	0.02695	0.09609	0.02827	0.10495	0.02996	0.10386	0.03147	0.10529	0.03022	0.10542	0.03251
3	0.0936	0.02634	0.09267	0.02748	0.10439	0.03055	0.10349	0.03201	0.10489	0.03097	0.10566	0.03345
3.5	0.09063	0.02586	0.08905	0.02679	0.10489	0.03159	0.10319	0.03277	0.10561	0.03232		
4	0.08679	0.02502	0.08556	0.02599	0.10467	0.03238	0.10344	0.0338				

Table 6.2: N=100 Summary results for $\hat{\beta}_t$

σ_v	OLR $X \sim Normal$	OLR $X \sim ChiSquared$	Reeves $X \sim Normal$	Reeves $X \sim ChiSquared$	Conditional Score $X \sim Normal$	Conditional Score $X \sim ChiSquared$
0	95.45	95.32	95.45	95.32	95.45	95.32
0.5	95.42	95.71	95.45	95.75	95.47	95.75
1	95.23	94.92	95.3	95.01	95.22	94.98
1.5	95.45	94.76	95.75	95.04	95.36	94.96
2	94.68	94.36	95.27	95.11	95.18	94.82
2.5	93.98	93.33	95.27	95.15	94.95	94.88
3	92.69	91.88	95.04	94.84	95.12	94.52
3.5	91.1	89.85	94.96	94.65	94.78	
4	88.25	87.34	94.28	94.83		

Table 6.3: N=100 Coverage results for $\hat{\beta}_t$

N=500 Degrees of Freedom=5

Table 6.4 displays the expected mean values of $\hat{\beta}_t$ and associated standard errors for both the case where X is normally distributed and where X has a chi-squared distribution with 5 degrees of freedom.

When $\sigma_v = 0$, the bias associated with all the methods are larger than those observed when X is normally distributed. For the ordinary logistic regression method as σ_v is increased, the same trend for the mean estimates is observed but with a larger bias. In this case, the Reeves method expected values also have a downward trend as σ_v is increased becoming negative when $\sigma_v > 2.5$. For the conditional score method, the size of the bias, though larger than previously observed, remains at approximately the same size for all values of σ_v.

In terms of the size of the empirical standard errors, when $\sigma_v = 0$, all the standard errors are larger than previously observed. For the ordinary logistic regression method, the size of the standard error decreases as σ_v is increased with the standard errors significantly smaller than those observed for the Reeves and conditional score methods. The Reeves method standard errors are significantly larger than those observed for the case when X is normally distributed reflecting a greater degree of variation in the estimates of $\hat{\beta}_t$ than previously observed. Overall, the conditional score method has the largest standard errors for all values of σ_v.

In terms of the coverage, the ordinary logistic regression reaches a low of 62% when $\sigma_v = 4$, compared to a low of 66%, when X is normally distributed, reflecting the increase in the bias and the unrealistic size of the standard errors. The Reeves and conditional score coverage levels are approximately the same at 93.5%.

In comparison, the ordinary logistic regression method is significantly affected by the change in distribution for X and the small sample size. For the conditional score method, the method does not make the assumption that the X values are normally distributed and the results show that this methods' estimates are slightly affected by the change in distribution but not by as much as the Reeves method. Therefore, for this sample size, it must be concluded that the conditional score method should be used in such situations.

	OLR $X \sim Normal$		OLR $X \sim ChiSquared$		Reeves $X \sim Normal$		Reeves $X \sim ChiSquared$		Conditional Score $X \sim Normal$		Conditional Score $X \sim ChiSquared$	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10068	0.01189	0.10102	0.01264	0.10068	0.01189	0.10102	0.01264	0.10068	0.01187	0.10102	0.01264
0.5	0.10056	0.0118	0.10071	0.01269	0.10086	0.01185	0.10101	0.01274	0.10087	0.01189	0.10107	0.01276
1	0.0995	0.01193	0.09936	0.01259	0.10068	0.01212	0.10055	0.01279	0.10073	0.01191	0.10079	0.01285
1.5	0.09824	0.01183	0.09781	0.01241	0.10087	0.01225	0.10044	0.01285	0.10098	0.01197	0.10098	0.01299
2	0.09618	0.01159	0.09547	0.01213	0.10076	0.01231	0.10002	0.0129	0.10094	0.01203	0.10096	0.01313
2.5	0.09371	0.01138	0.0925	0.01183	0.10068	0.01249	0.09939	0.013	0.10095	0.01212	0.10079	0.01337
3	0.09089	0.01113	0.08951	0.01147	0.10063	0.01271	0.0991	0.01309	0.10102	0.01224	0.10107	0.01363
3.5	0.08785	0.011	0.08599	0.0112	0.10067	0.01313	0.09848	0.01334	0.10118	0.01239	0.10103	0.01406
4	0.08459	0.01061	0.08248	0.0109	0.10072	0.01332	0.09813	0.01367	0.10138	0.01257	0.10132	0.01459

Table 6.4: N=500 Summary results for $\hat{\beta}_t$

σ_v	OLR $X \sim Normal$	OLR $X \sim ChiSquared$	Reeves $X \sim Normal$	Reeves $X \sim ChiSquared$	Conditional Score $X \sim Normal$	Conditional Score $X \sim ChiSquared$
0	94.92	94.92	94.92	94.92	94.92	94.92
0.5	95.28	94.72	95.37	94.78	95.25	94.77
1	94.75	94.66	95.02	94.84	95.09	94.81
1.5	94.48	93.78	95.12	94.95	94.77	94.88
2	92.96	92.75	94.91	94.82	94.55	94.6
2.5	89.96	88.74	94.75	94.64	94.43	94.27
3	84.6	83.05	94.68	94.75	94.02	94.37
3.5	77.32	73.32	94.43	94.1	93.39	93.63
4	65.82	61.87	93.92	93.62	93.44	93.44

Table 6.5: N=500 Coverage results for $\hat{\beta}_t$

N=1000 Degrees of Freedom=5

The same trends in the estimates observed for n=500 are again observed for n=1000, Table 6.6, with the Reeves method estimates being more affected by the increase in the sample size. For the Reeves estimates, the expected values of $\hat{\beta}_1$ are negatively biased when $\sigma_v > 1$, whereas the conditional score method estimates remain with the same size of bias as σ_v increased. In terms of the coverage, Table 6.7, the Reeves and conditional score methods have a low of 93% whereas the ordinary logistic regression the low reaches approximately 35%.

For the larger sample size, the effect of the X values coming from a chi-squared distribution on the methods' estimates can be seen. In this case, the best method to use is that of the conditional score method as the Reeves method could under-estimate any relationship.

	OLR $X \sim Normal$		OLR $X \sim ChiSquared$		Reeves $X \sim Normal$		Reeves $X \sim ChiSquared$		Conditional Score $X \sim Normal$		Conditional Score $X \sim ChiSquared$	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10051	0.00847	0.10032	0.00883	0.10051	0.00847	0.10032	0.00883	0.10051	0.00836	0.10032	0.00883
0.5	0.10009	0.0084	0.09995	0.00881	0.10039	0.00844	0.10025	0.00885	0.1004	0.00837	0.10031	0.00886
1	0.09916	0.00829	0.09903	0.0088	0.10033	0.00842	0.1002	0.00894	0.10038	0.00839	0.10044	0.00898
1.5	0.09768	0.00826	0.09735	0.00868	0.10027	0.00855	0.09994	0.00898	0.10037	0.00842	0.10048	0.00908
2	0.09565	0.00803	0.09508	0.00851	0.10017	0.00853	0.09957	0.00903	0.10034	0.00846	0.10051	0.0092
2.5	0.0934	0.00797	0.09229	0.00825	0.1003	0.00874	0.09908	0.00905	0.10057	0.00853	0.10049	0.0093
3	0.09043	0.00787	0.0892	0.00816	0.10004	0.00897	0.09865	0.00929	0.10041	0.0086	0.10059	0.00966
3.5	0.08744	0.00769	0.08574	0.00781	0.10008	0.00915	0.09808	0.00928	0.10059	0.00871	0.10061	0.00976
4	0.08402	0.00752	0.08205	0.00766	0.09989	0.0094	0.09744	0.00957	0.10052	0.00882	0.10056	0.01021

Table 6.6: N=1000 Summary results for $\hat{\beta}_t$

σ_v	OLR $X \sim Normal$	OLR $X \sim ChiSquared$	Reeves $X \sim Normal$	Reeves $X \sim ChiSquared$	Conditional Score $X \sim Normal$	Conditional Score $X \sim ChiSquared$
0	94.94	95.34	94.94	95.34	94.94	95.34
0.5	94.86	94.88	94.95	95.06	94.85	95.01
1	94.41	94.84	94.7	95.08	94.98	94.96
1.5	93.8	92.85	94.94	94.78	94.87	94.66
2	91.05	90.19	95.02	94.63	95.01	94.6
2.5	84.93	83.36	94.75	94.63	94.6	94.4
3	74.82	71.93	94.32	94.16	94.19	93.78
3.5	59.63	55.56	94.2	93.66	93.48	93.62
4	41.98	35.31	94.05	92.82	93.62	92.92

Table 6.7: N=1000 Coverage results for $\hat{\beta}_t$

N=100 Degrees of Freedom=20

For this case, the number of degrees of freedom has been increased to 20. The effect is that the distribution of the X values is more symmetrical than the previous case and hence tending towards a normal distribution.

For all three methods, the size of the bias associated with the expected values of $\hat{\beta}_1$ is the same for this case as was observed for the case when X is described by a normal distribution, Table 6.8. The trends in the mean values, as σ_v is increased, are also the same. With respect to the size of the empirical standard errors, these too are the same size as those observed when X is described by a normal distribution. For the ordinary logistic regression method, the standard errors are smaller than those estimated by the Reeves and conditional score methods. These results are reflected in the coverage levels, Table 6.9, which are again the same as previously observed. For the ordinary logistic regression method, the coverage is approximately 88% whereas for the Reeves and conditional score methods the coverage is 94% reflecting the larger standard errors and less bias.

These results show that when the explanatory variable can be described by a slightly skewed chi-squared distribution and the sample size is small, the Reeves estimates are not affected in terms of bias and size of standard errors. In this case the Reeves and conditional score methods are comparable.

	OLR $X \sim Normal$		OLR $X \sim ChiSquared$		Reeves $X \sim Normal$		Reeves $X \sim ChiSquared$		Conditional Score $X \sim Normal$		Conditional Score $X \sim ChiSquared$	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10443	0.02858	0.10426	0.02892	0.10443	0.02858	0.10426	0.02892	0.10443	0.02858	0.10426	0.02892
0.5	0.1041	0.0283	0.10406	0.02902	0.10443	0.02842	0.10439	0.02915	0.10445	0.02843	0.10441	0.02916
1	0.10295	0.02838	0.10299	0.0283	0.10426	0.02887	0.1043	0.0288	0.10431	0.02891	0.10439	0.02886
1.5	0.10129	0.02778	0.10137	0.02823	0.10419	0.02888	0.10427	0.02935	0.10432	0.02897	0.10448	0.02947
2	0.09884	0.0271	0.09937	0.02779	0.10387	0.02901	0.10444	0.02974	0.10409	0.02917	0.10483	0.02998
2.5	0.09717	0.02695	0.09687	0.0274	0.10495	0.02996	0.1046	0.03047	0.10529	0.03022	0.10522	0.03089
3	0.0936	0.02634	0.09347	0.02636	0.10439	0.03055	0.10421	0.03062	0.10489	0.03097	0.10505	0.03116
3.5	0.09063	0.02586	0.09015	0.02622	0.10489	0.03159	0.10437	0.03207	0.10561	0.03232	0.10553	0.03296
4	0.08679	0.02502	0.08703	0.02545	0.10467	0.03238	0.10501	0.03314				

Table 6.8: N=100 Summary results for $\hat{\beta}_t$

σ_v	OLR $X \sim Normal$	OLR $X \sim ChiSquared$	Reeves $X \sim Normal$	Reeves $X \sim ChiSquared$	Conditional Score $X \sim Normal$	Conditional Score $X \sim ChiSquared$
0	95.45	95.27	95.45	95.27	95.45	95.27
0.5	95.42	95.35	95.45	95.37	95.47	95.38
1	95.23	95.28	95.3	95.42	95.22	95.37
1.5	95.45	94.87	95.75	95.24	95.36	95.14
2	94.68	94.54	95.27	95.25	95.18	95.13
2.5	93.98	93.87	95.27	95.08	94.95	94.96
3	92.69	92.52	95.04	94.82	95.12	94.65
3.5	91.1	90.51	94.96	94.68	94.78	94.42
4	88.25	88.20	94.28	94.23		

Table 6.9: N=100 Coverage results for $\hat{\beta}_t$

N=500 Degrees of Freedom=20

The results for this sample size are displayed in Table 6.10 and Table 6.11. The biases associated with the expected values of $\hat{\beta}_1$ have all reduced in size for all the methods. For the ordinary logistic regression method, the trend of the expected values of $\hat{\beta}_1$ as σ_v is increased is as previously observed for all the other cases. The Reeves method estimates have a positive bias for the case when X is described by a normal distribution. For this case, the bias is negative when $\sigma_v > 3.5$, though the size of the bias is approximately the same. For the conditional score method, the bias is marginally larger when $\sigma_v = 0$ for the chi-squared distribution case than for the normal distribution case. As σ_v is increased, the size of the bias stays at approximately the same level, whereas for the normal distribution case, the size of the bias increases slightly for large σ_v values.

On examination of the empirical standard errors for the ordinary logistic regression and Reeves methods, the sizes of the standard errors are approximately the same for both the chi-squared and the normal results. For the conditional score method, the chi-squared results show that the standard errors are slightly larger than previously observed. In comparison though, the Reeves and conditional score standard errors are approximately the same size for all σ_v values.

When $\sigma_v = 4$, the ordinary logistic regression method has a coverage level of 64% reflecting the size of the bias and the small standard error. This shows that for the ordinary logistic regression method, the small standard error does not reflect the effect of

the measurement error whereas the conditional score and Reeves methods' have a coverage level of approximately 93% reflecting the increased size of the standard errors and the smaller bias.

For this case, the estimates from the Reeves and conditional score methods' are comparable.

	OLR $X \sim Normal$		OLR $X \sim ChiSquared$		Reeves $X \sim Normal$		Reeves $X \sim ChiSquared$		Conditional Score $X \sim Normal$		Conditional Score $X \sim ChiSquared$	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10068	0.01189	0.10095	0.01222	0.10068	0.01189	0.10095	0.01222	0.10068	0.01187	0.10095	0.01222
0.5	0.10056	0.0118	0.10043	0.01205	0.10086	0.01185	0.10073	0.01209	0.10087	0.01189	0.10076	0.0121
1	0.0995	0.01193	0.09966	0.01187	0.10068	0.01212	0.10085	0.01205	0.10073	0.01191	0.10094	0.01207
1.5	0.09824	0.01183	0.09799	0.01206	0.10087	0.01225	0.10062	0.01248	0.10098	0.01197	0.10082	0.01253
2	0.09618	0.01159	0.09602	0.01164	0.10076	0.01231	0.1006	0.01237	0.10094	0.01203	0.10096	0.01246
2.5	0.09371	0.01138	0.09341	0.0115	0.10068	0.01249	0.10035	0.01262	0.10095	0.01212	0.10089	0.01276
3	0.09089	0.01113	0.09058	0.0112	0.10063	0.01271	0.10028	0.01279	0.10102	0.01224	0.10105	0.01299
3.5	0.08785	0.011	0.08711	0.01095	0.10067	0.01313	0.09978	0.01308	0.10118	0.01239	0.10079	0.01335
4	0.08459	0.01061	0.08374	0.01069	0.10072	0.01332	0.09966	0.0134	0.10138	0.01257	0.10092	0.01377

Table 6.10: N=500 Summary results for $\hat{\beta}_t$

σ_v	OLR $X \sim Normal$	OLR $X \sim ChiSquared$	Reeves $X \sim Normal$	Reeves $X \sim ChiSquared$	Conditional Score $X \sim Normal$	Conditional Score $X \sim ChiSquared$
0	94.92	94.88	94.92	94.88	94.92	94.88
0.5	95.28	95.05	95.37	95.14	95.25	95.13
1	94.75	94.93	95.02	95.2	95.09	95.15
1.5	94.48	93.86	95.12	94.95	94.77	94.82
2	92.96	92.88	94.91	95.06	94.55	94.84
2.5	89.96	89.49	94.75	94.71	94.43	94.34
3	84.6	84.52	94.68	94.83	94.02	94.36
3.5	77.32	76.08	94.43	94.15	93.39	93.55
4	65.82	64.06	93.92	93.91	93.44	93.24

Table 6.11: N=500 Coverage results for $\hat{\beta}_t$

N=1000 Degrees of Freedom=20

Table 6.12 shows that the increase in the sample size has reduced the associated bias and standard errors for the expected values of $\hat{\beta}_1$ for all the methods except for the ordinary logistic regression method. The same trends in the estimates that were observed for n=500 are again observed for n=1000.

For the Reeves method, the negative bias has reduced as σ_v increased. For the conditional score method, the positive bias has also reduced in size. For the Reeves and ordinary logistic regression methods, the standard errors are approximately the same size as those observed for when X is normally distributed. For the conditional score method, the standard errors are slightly larger than previously observed.

For this case, the Reeves and logistic regression methods estimates are affected by the change in the distribution. The results show that the conditional score method produced the least biased results.

	OLR $X \sim Normal$		OLR $X \sim ChiSquared$		Reeves $X \sim Normal$		Reeves $X \sim ChiSquared$		Conditional Score $X \sim Normal$		Conditional Score $X \sim ChiSquared$	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10051	0.00847	0.10045	0.00848	0.10051	0.00847	0.10045	0.00848	0.10051	0.00836	0.10045	0.00848
0.5	0.10009	0.0084	0.10011	0.00849	0.10039	0.00844	0.1004	0.00853	0.1004	0.00837	0.10042	0.00853
1	0.09916	0.00829	0.09914	0.00839	0.10033	0.00842	0.10031	0.00852	0.10038	0.00839	0.1004	0.00853
1.5	0.09768	0.00826	0.0976	0.00841	0.10027	0.00855	0.10019	0.00871	0.10037	0.00842	0.10039	0.00874
2	0.09565	0.00803	0.09561	0.0082	0.10017	0.00853	0.10013	0.00871	0.10034	0.00846	0.10049	0.00878
2.5	0.0934	0.00797	0.09298	0.00807	0.1003	0.00874	0.09984	0.00885	0.10057	0.00853	0.10038	0.00894
3	0.09043	0.00787	0.09022	0.00793	0.10004	0.00897	0.0998	0.00903	0.10041	0.0086	0.10055	0.00917
3.5	0.08744	0.00769	0.08681	0.00769	0.10008	0.00915	0.09933	0.00914	0.10059	0.00871	0.10033	0.00934
4	0.08402	0.00752	0.08351	0.00756	0.09989	0.0094	0.09924	0.00945	0.10052	0.00882	0.1005	0.00971

Table 6.12: N=1000 Summary results for $\hat{\beta}_t$

σ_v	OLR $X \sim Normal$	OLR $X \sim ChiSquared$	Reeves $X \sim Normal$	Reeves $X \sim ChiSquared$	Conditional Score $X \sim Normal$	Conditional Score $X \sim ChiSquared$
0	94.94	95.15	94.94	95.15	94.94	95.15
0.5	94.86	94.85	94.95	94.83	94.85	94.81
1	94.41	94.71	94.7	94.93	94.98	94.87
1.5	93.8	93.37	94.94	94.81	94.87	94.68
2	91.05	90.9	95.02	94.83	95.01	94.62
2.5	84.93	84.34	94.75	94.55	94.6	94.31
3	74.82	74.07	94.32	94.54	94.19	94.06
3.5	59.63	58.57	94.2	94.15	93.48	93.62
4	41.98	39.96	94.05	93.89	93.62	93.18

Table 6.13: N=1000 Coverage results for $\hat{\beta}_t$

6.2.3 Conclusion for where X does not follow a Normal Distribution

In this study we have investigated the case where X follows a chi-squared distribution with 5 and 20 degrees of freedom respectively. In the first case the distribution is highly skewed, for the latter, the distribution is tending towards a normal distribution. In terms of the methods investigated, the ordinary logistic regression was investigated to see whether a change in the values of X would affect the overall estimates of the model parameters. For the correction methods, the Reeves method assumes that the X values are normally distributed whereas the conditional score method does not make such a model assumption.

In terms of the logistic regression method, for both distributions and all the sample sizes, the effect on the bias was only marginal, with a large measurement error standard deviation causing a larger negative bias. For the standard errors, these were approximately the same size as those observed for the normal case for medium and large sample sizes. Overall, when the X values are not normally distributed and there is measurement error in the explanatory variable the logistic regression method will under-estimate the relationship between the explanatory variable and the disease status.

Throughout chapter 5, the Reeves and conditional score methods produced comparable results thereby enabling us to recommend the Reeves method as it is simpler to implement. This study was conducted in order to see whether these methods are robust to a change in the distribution of the explanatory variable. When the X values had a chi-squared distribution with 5 degrees of the freedom, the Reeves method was not as robust

as previously observed. When the sample size was small, the estimates were comparable to those produced by the conditional score method. As the sample size increased, so the estimates had a trend towards zero as σ_v increased. This behaviour was not observed for the same case in chapter 5. This trend resulted in a large negative bias when the measurement error standard deviation was large. In comparison, the conditional score method estimates maintained the same size bias as σ_v increased. The bias was slightly larger than observed previously however, this again reduced as the sample size was increased.

The trends of the estimates observed for the Reeves and conditional score methods for the chi-squared distribution with 5 degrees of freedom were observed for the other chi-squared distribution. In this case, the effect of the distribution not being as skewed, is that the bias observed in relation to the Reeves method had reduced. However, the conditional score method still produced the best results.

In conclusion, when it cannot be assumed that the explanatory variable is normally distributed then the conditional score method is recommended for all similar cases.

6.3 Case where the variation from the validation study is taken into account

6.3.1 Introduction

When an explanatory variable is known to be measured with error, it is common to undertake a subsidiary study in order to estimate the parameters of the error distribution associated with measuring the explanatory variable. In section 3.2.6 subsidiary studies are covered in more detail. In practice it cannot be assumed that the measurement error standard deviation is known exactly, and that the estimate could be subject to sampling variation especially if the sample size is small. To understand the effect that such a variation could have on the estimates of the regression parameters and the associated standard errors a simulation study was conducted.

6.3.2 Simulation study

To investigate and to compare the methods of Reeves and conditional score a simulation study was conducted according to the model described in section 6.2.2 except that a validation sample was generated to estimate the measurement error standard deviation. This estimate was then used in the methods of Reeves and conditional score.

The purpose of the validation sample is to provide an estimate for σ_v. This can be estimated either by having a repeated measure on the same individuals, in which case σ_v is estimated from the within-subject differences, or else by having a gold standard to measure the true value X; this can be compared to the usual observed on Z and the

differences are the errors V. For these purposes, a sample of values of V were generated in the simulation and the usual unbiased estimator of the error variance was derived. This was used in estimating the model parameters but the true value was used for generating the data.

N=500 Reeves results

Table 6.14 displays the results of the Reeves method. The table compares the expected values of $\hat{\beta}_t$ when σ_v is known, and when it is estimated from a validation study size of 25, 50 and 100 respectively.

When there is a small validation size that is, m=25, there is less information in order to estimate σ_v, and therefore it would be expected that σ_v will not be estimated precisely. When there is no measurement error, $\sigma_v = 0$, the effect on the Reeves method, is that the bias associated with the expected value of $\hat{\beta}_t$ is larger than that of the case when the measurement error is known. As σ_v is increased, the bias reduces in size such that the bias is less than that observed for the case where σ_v is known precisely, though the reduction in bias is only marginal.

As the validation size is increased to 50 and 100 respectively, the bias associated with the expected mean values of $\hat{\beta}_t$ is reduced to approximately the same size as the case when σ_v is known precisely. This suggests that for these sizes of validation study, the estimates

of σ_v are approximately the same as the true values and therefore are not overly affecting the expected values of $\hat{\beta}_t$.

When considering the empirical standard errors associated with the expected values of $\hat{\beta}_t$, Table 6.15, for a small validation study size it would be expected that there would be more fluctuation in the estimates of β_t. When $\sigma_v = 0$, the standard errors are approximately the same size for both the case when the measurement error standard deviation is known precisely and also when it is estimated from a validation study size of 25. The fluctuations in the estimates of β_t are seen as the measurement error standard deviation is increased with an increase in the size of the empirical standard errors. As the validation study is increased to 50 and 100 respectively, the size of the empirical standard errors decrease in-line with the decrease in bias. As with the size of the bias, the difference in the size of the empirical standard errors is only marginal compared to when σ_v is known.

		m=25	m=50	m=100
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$
0	0.10068	0.10107	0.10079	0.10065
0.5	0.10086	0.10085	0.10105	0.10068
1	0.10068	0.10058	0.1008	0.10077
1.5	0.10087	0.10077	0.10071	0.10056
2	0.10076	0.10057	0.10059	0.10056
2.5	0.10068	0.1006	0.10038	0.1007
3	0.10063	0.10035	0.10075	0.10052
3.5	0.10067	0.1002	0.10054	0.10065
4	0.10072	0.1002	0.10008	0.10023

Table 6.14: N=500 Reeves $\hat{\beta}_t$

		m=25	m=50	m=100
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.01189	0.01188	0.01195	0.01193
0.5	0.01185	0.01197	0.01201	0.01202
1	0.01212	0.01198	0.01211	0.01201
1.5	0.01225	0.01222	0.01206	0.01207
2	0.01231	0.01228	0.01231	0.01226
2.5	0.01249	0.01262	0.01241	0.01249
3	0.01271	0.01302	0.013	0.0127
3.5	0.01313	0.0134	0.01329	0.01325
4	0.01332	0.01432	0.01374	0.01349

Table 6.15: N=500 Reeves $SE_E(\hat{\beta}_t)$

N=500 Conditional Score results

Table 6.16 shows the results for the Conditional Score method. The bias for m=25 is larger than when the measurement error standard deviation is known precisely. For the conditional score method, the size of the bias stays at approximately the same size as σ_v is increased. When this size of bias is compared to the results for the case when σ_v is known precisely, the bias is smaller. As the validation study size is increased, the size of the bias remains at approximately the same size for all σ_v values.

Table 6.17 displays the empirical standard errors. For the cases when $\sigma_v = 0$, all the validation sizes, the standard errors are approximately the same size and are comparable with those when σ_v is known. This is true up until $\sigma_v = 2$. When $\sigma_v > 2$, for m=25, the size of the standard errors are larger than when σ_v is known. As the validation size is increased, this difference decreases in size. The size of the standard errors is therefore affected when the validation study is small.

		m=25	m=50	m=100
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$
0	0.10068	0.10107	0.10079	0.10065
0.5	0.10087	0.10086	0.10106	0.10069
1	0.10073	0.10062	0.10084	0.10082
1.5	0.10098	0.10087	0.10081	0.10066
2	0.10094	0.10074	0.10076	0.10074
2.5	0.10095	0.10087	0.10065	0.10097
3	0.10102	0.10073	0.10113	0.10091
3.5	0.10118	0.10069	0.10104	0.10117
4	0.10138	0.10086	0.10072	0.10088

Table 6.16: N=500 Conditional Score $\hat{\beta}_t$

		m=25	m=50	m=100
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.01189	0.01188	0.01195	0.01193
0.5	0.01185	0.01197	0.01201	0.01203
1	0.01213	0.01199	0.01212	0.01202
1.5	0.01228	0.01225	0.01209	0.0121
2	0.01236	0.01233	0.01236	0.01231
2.5	0.01257	0.01272	0.01249	0.01258
3	0.01283	0.01316	0.01313	0.01281
3.5	0.01329	0.0136	0.01346	0.01343
4	0.01354	0.01465	0.01399	0.01372

Table 6.17: N=500 Conditional Score $SE_E(\hat{\beta}_t)$

N=500 Comparison between the Reeves and conditionals core methods

When there is a small validation study size and a large measurement error standard deviation, the Reeves method is more affected than the conditional score method. This is reflected in the increasing negative bias associated with the expected values of $\hat{\beta}_t$ as the measurement error standard deviation is increased. For the conditional score method, the size of the bias remains approximately the same as the measurement error standard deviation is increased. Comparing the associated standard errors, the size of these errors are approximately the same for the two methods. From these results, it appears that there is no real difference between the two methods.

N=1000 Reeves Results

The sample size has been increased to a 1000 with validation study sizes of 50, 100 and 200 respectively. Table 6.18 displays the results for the expected values of $\hat{\beta}_t$ for all three validation study sizes and the case where the measurement error standard deviation is known. For this case, the size of the validation study and therefore the estimate of σ_v does not have an effect on the size of the bias. The results show that for all the measurement error standard deviation values, the size of the bias is approximately the same as those found for when σ_v is known precisely. These results suggest that for a large sample size with a 5% validation study size, the validation study has enough information in order to approximate the true value of σ_v.

With respect to the associated standard errors, Table 6.19, for a small validation study size, they are slightly larger than those observed for when σ_v is known. This difference reduces as the validation study size is increased. Any differences are marginal and do not overly affect the results.

		m=50	m=100	m=200
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$
0	0.10051	0.10044	0.10035	0.10031
0.5	0.10039	0.10039	0.1003	0.10054
1	0.10033	0.10035	0.10044	0.10039
1.5	0.10027	0.10036	0.10031	0.10026
2	0.10017	0.10024	0.10029	0.10024
2.5	0.1003	0.10009	0.10023	0.10004
3	0.10004	0.10004	0.10002	0.10025
3.5	0.10008	0.09983	0.0998	0.10004
4	0.09989	0.09974	0.09991	0.09974

Table 6.18: N=1000 Reeves $\hat{\beta}_t$

		m=50	m=100	m=200
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.00847	0.00829	0.00834	0.00842
0.5	0.00844	0.00842	0.00837	0.00842
1	0.00842	0.00829	0.0084	0.00839
1.5	0.00855	0.00861	0.00868	0.00847
2	0.00853	0.00871	0.00852	0.00863
2.5	0.00874	0.00887	0.00884	0.00863
3	0.00897	0.00921	0.00897	0.00895
3.5	0.00915	0.00953	0.00938	0.00927
4	0.0094	0.0101	0.00981	0.00948

Table 6.19: N=1000 Reeves $SE_E(\hat{\beta}_t)$

N=1000 Conditional Score Results

Table 6.20 displays the expected values of $\hat{\beta}_t$ for the case where σ_v is known and for the validation study sizes of 50, 100 and 200 respectively. As was observed for the Reeves method, the expected values of $\hat{\beta}_t$ are approximately the same size for all four cases and for all σ_v values. The empirical standard errors in Table 6.21 are also the same size in comparison to when σ_v is known.

Overall, these results suggest that the validation study sizes are sufficient in order to estimate σ_v, and that the conditional score method estimates are not affected by this estimation as opposed to knowing the measurement error standard deviation precisely.

		m=50	m=100	m=200
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$	$\hat{\beta}_t$
0	0.10051	0.10044	0.10035	0.10031
0.5	0.1004	0.1004	0.10031	0.10055
1	0.10038	0.10039	0.10048	0.10043
1.5	0.10037	0.10046	0.1004	0.10035
2	0.10034	0.10041	0.10046	0.10041
2.5	0.10057	0.10035	0.10049	0.1003
3	0.10041	0.10041	0.10039	0.10063
3.5	0.10059	0.10032	0.10029	0.10053
4	0.10052	0.10037	0.10055	0.10038

Table 6.20: N=1000 Conditional Score $\hat{\beta}_t$

		m=50	m=100	m=200
	σ_v known	$\hat{\sigma}_v$	$\hat{\sigma}_v$	$\hat{\sigma}_v$
σ_v	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$	$SE_E(\hat{\beta}_t)$
0	0.00847	0.00829	0.00834	0.00842
0.5	0.00844	0.00843	0.00837	0.00842
1	0.00843	0.0083	0.00841	0.00839
1.5	0.00857	0.00863	0.00869	0.00849
2	0.00856	0.00875	0.00856	0.00866
2.5	0.0088	0.00893	0.00889	0.00868
3	0.00905	0.0093	0.00905	0.00903
3.5	0.00926	0.00967	0.00949	0.00938
4	0.00955	0.0103	0.00998	0.00964

Table 6.21: N=1000 Conditional Score $SE_E(\hat{\beta}_t)$

6.3.3 Conclusion

In this case, the measurement error standard deviation was estimated from a validation study. The aim was to investigate if there would be an effect on the estimates by the Reeves and conditional score methods. Two sample sizes of 500 and 1000 were chosen with respective 5%, 10% and 25% validation study sizes. The effects on the method estimates of $\hat{\beta}_t$ were considered by observing the level of associated bias on the expected values and the size of the empirical standard errors.

The effect of estimating the measurement error standard deviation was only observed when the sample size was 500 and the validation study was 25. The effect was an increase in the size of the bias for both the methods. As the true value of the measurement error standard deviation was increased, the estimate of σ_v affected the expected values of $\hat{\beta}_t$ for the Reeves method, where a negative bias was observed. This affect was not observed for the conditional score method. Overall, this effect was only marginal and the two methods produced comparable results.

For all other sample sizes and validation study sizes, no real differences in size of bias and standard errors were observed. In conclusion, there is no practical difference between the two methods for the case A that has been investigated. Therefore, both methods are robust to using an estimate of σ_v rather than it being known precisely.

6.4 Discussion

In chapter 5, the Reeves and conditional score methods were comparable in all cases. As the Reeves method is simpler to use, this method was recommended. In order to implement the Reeves method a number of modelling assumptions have to be made, including that the X values are normally distributed. In comparison the conditional score method does not make such a model assumption. However, to implement both the methods an estimate of the measurement error standard deviation is required. This chapter was designed to investigate the robustness of the methods to a change in the model assumptions with regard to the explanatory variable, as well as an estimate of the measurement error standard deviation being used in the methods as opposed to it being known precisely, the latter assumption implying a need for a certain size validation study.

With regard to a change in the distribution of the explanatory variable, highly skewed and skewed chi-squared distributions were investigated. For both distributions it was found that the simpler to implement Reeves method estimates were affected by the non-normal distribution. Therefore, it is recommended that the conditional score method should be used.

In terms of using an estimate of the measurement error standard deviation in the Reeves and conditional score methods, a validation study size of 10% of the sample was found to be sufficient in order to produce the same level of biased results as when the measurement error standard deviation is known. A smaller validation size could mean that the estimate of the measurement error standard deviation could affect the Reeves and conditional score estimates. A validation study size of 20% did not affect

the size of the bias or the standard errors and therefore do not provide any more information than that obtained from the 10% validation study size. Therefore, in order to truly understand the size of the measurement error associated with the explanatory variable a validation study size of approximately 10% should be sufficient.

Overall, when considering the planning of any study, the distribution of the explanatory variable must be determined in order to choose the best correction method and a sufficiently sized validation study must also be included.

Chapter 7

7 Bayesian Analysis

7.1 An Introduction to Bayesian Analysis

Within a classical framework as previously considered in this study, inference about model parameters is based entirely on the data. These model parameters are regarded as fixed, but unknown, constants. In a Bayesian framework, both the data and the model parameters are considered to be random variables and both are described through probability distributions. The aim is to estimate the distributions of the model parameters conditional on the observed data, and to then draw inferences about the model parameters from these so-called posterior distributions.

Within a Bayesian approach, it is possible to incorporate any previous knowledge about the model parameters θ within the approach. This is included in the form of a prior distribution $p(\theta)$. If no prior information is available then this prior can be

chosen to be uninformative, usually by giving it a large variance. If there is real information about the parameters then this can be built into the specification of the prior distribution. The prior is then combined with the data to obtain a joint distribution of the parameters and the data and then the conditional distribution of the parameter θ given the data is derived using Bayes' theorem. This is called the posterior distribution of the parameters, given the observed data. Specifically by Bayes theorem

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \qquad (7.1)$$

where

$$f(x) = \int f(x|\theta)p(\theta)d\theta \qquad (7.2)$$

Therefore, the posterior distribution can be determined if the integral can be evaluated.

In many practical situations, an analytical evaluation of the integral in (7.2) is not possible. It may be possible to change the specification of the prior distribution into what is called a conjugate prior, which often leads to an analytical expression for the posterior distribution. For example with data following a binomial distribution the conjugate prior is a beta distribution. If the prior distribution for the probability θ is taken to be a beta distribution then the posterior distribution is a different beta distribution and the technical problems largely disappear. It is often not possible or appropriate to do this and so an alternative approach is to use a numerical method for estimating the value of the integral. A variety of algorithms for this have been derived but the methods are cumbersome. The advent of Markov Chain Monte Carlo methods has revolutionised the situation.

7.2 An introduction to Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) is a Monte Carlo simulation based on a Markov Chain, which is constructed in such a way that its states represent possible values of the posterior distribution and sample values from the simulation of the chain form a sample from the posterior distribution. This section provides a brief introduction to the technique; a more comprehensive discussion of the technique and issues surrounding its use is contained in Gamerman (1997).

A Markov Chain is a probabilistic model consisting of a set of states and a set of probabilities of transitions between these states. It has the special property that future movements between states depend only on the current state occupied and not on previously visited states. It is often described as having no memory for that reason. Two concepts are particularly important. The stationary distribution π is a probability distribution of the states which does not change with time. Thus if π_i is the probability that state i is occupied, this does not change as the chain evolves. The limiting distribution is a distribution of the states which may be reached after the chain has run for a long time. It can be shown that under certain fairly general conditions (Gamerman (1997)) a Markov Chain has a limiting distribution which coincides with the stationary distribution. This is crucial for MCMC methods.

MCMC methods involve the construction of a Markov chain which satisfies the conditions referred to above and whose stationary distribution is the posterior distribution required. This means that if the chain is simulated for a long time, the probabilities the states are occupied should tend to the limiting distribution and this is

actually the stationary distribution and hence the posterior distribution. Hence an appropriate simulation will provide a sample from this required posterior distribution and properties can be derived from it.

There are two main problems, therefore. The first and most important is the construction of an appropriate transition matrix between the states. The second more practical one is the determination of a stage in the simulation by which the stationary distribution has been reached, so that we can be sure we are sampling from the posterior distribution.

There are a number of ways of determining this transition matrix. The Metropolis-Hastings is a general method which has Gibbs sampling as a special case. As the software used in this work is based on the latter method, both are outlined here; full details are in Gammerman(1997).

The Metropolis-Hastings algorithm uses a proposal distribution to generate a possible next state, called a candidate state, from the current one. An acceptance probability, which depends on the proposal distribution and the current state, is derived and the candidate state is accepted with this probability. If rejection occurs the next state is the current state – the chain does not move. Provided the acceptance probability is properly constructed it can be shown (Gammerman (1997)) that the chain has the appropriate stationary distribution, regardless of the form of the proposal distribution. There are conflicting demands on this choice. A proposal distribution which leads to a candidate state near the current one will have a high probability of acceptance, but it may take a very long simulation run to reach the limiting, and hence stationary

distribution. One which makes larger jumps from the current state will have a much smaller acceptance probability. The choice of proposal distribution is therefore important in practice.

Another form of the Metropolis-Hastings algorithm is that of the Single Component Metropolis Hastings algorithm which provides a computationally more efficient algorithm when the problem is a multivariate one, that is there is more than one parameter. A state now consists of a k-dimensional point, where k is the number of parameters, and a new state may be found by changing components one at a time or changing several simultaneously.

Gibbs Sampling technique is a special case of the Single Component Metropolis Hastings algorithm. It requires the derivation of the full conditional distribution for each component given the values of the remaining components. This is then used as the proposal distribution. The acceptance probability is 1 and so all proposed points are accepted. This can have an effect on the convergence of the chain as well as on the way the sampled value moves around the posterior distribution.

This leads us to the second problem, that of assessing convergence. By convergence we mean that the distribution of the states occupied has reached the limiting distribution of the chain, that is the required posterior distribution. This issue has been referred to implicitly in the above discussion of the methods for constructing the chain and many modifications to the basic ideas above have been suggested with a view to accelerating convergence, for further details see Gammerman (1997

If a method can be found for determining when the limiting distribution has been reached, the values produced by the simulation before this stage has been reached are discarded so as not to influence the estimation of the posterior distribution. This period is known as a burn-in period and any values drawn during this period are discarded. The practical problem is therefore to determine how long a burn-in period is required. It is possible to derive some simple diagnostics during a simulation which can cast some light on the rate of convergence. For example the autocorrelation function for a given parameter examines correlations within the sequence of values of that parameter produced by the simulation. If the autocorrelations remain high at substantial lags it suggests that successive sampled values are very close together and this may mean that convergence is slow; the chain is said to be mixing slowly.

There are two principal ways to address the issue of convergence. The first is to run a number of chains with different starting values and to explore when the chains converge to the same distribution to thus determine a burn-in period. Alternatively a single chain can be run for an increased length of time and different parts of the sequence of values produced can be compared to see if they can be regarded as coming from the same distribution.

Gelman and Rubin proposed a test, subsequently modified by Brooks, based on the first principle. Several chains are run, starting from over-dispersed initial values, that is these are chosen to have greater variance than will be expected of the posterior distribution. The chains are monitored, both separately and when pooled. An analysis of variance type of method is employed to compare the variation within individual chains with that in the pooled chain and when these are sufficiently similar it is

concluded that convergence has occurred. Geweke's test by contrast is based on a single chain and takes a number of different sections of different lengths from the sequence of values produced. A succession of tests is performed on these to assess the state of convergence.

Other methods have been proposed; a full discussion is beyond the scope of this work but the question of convergence will be discussed in the context of specific examples later. For further information see Gelman & Rubin (1992b) and Geweke (1989) and (1992).

7.3 BUGS software

BUGS (Bayesian Inference using Gibbs Sampling) is a piece of software which implements MCMC methods using Gibbs sampling and which is freely available from the MRC Biostatistics Unit, Cambridge. It provides a general framework for defining models and constructs automatically the conditional distributions required for the implementation of the Gibbs sampling. Further details of the many features of the software can be found at the following website http://www.mrc-bsu.cam.ac.uk/bugs. For the purposes of this study, the BUGS software provided all the features that were required to model the logistic regression measurement error model in a Bayesian framework and thus will be used throughout this chapter.

7.4 Current literature associated with the Bayesian Analysis of the Measurement Error Model

A number of authors have addressed the issues surrounding the Bayesian analysis of the logistic regression measurement error model and how such a model can be

implemented using a Bayesian analysis approach with MCMC. Richardson and Gilks (1993) discuss the logistic regression measurement error problem in terms of conditional independence models. They construct a model of the measurement error problem in such a way that different types of study and validation study designs can be incorporated within the model. That is, once the measurement error model has been determined as well as the distributions associated with the explanatory variable and the measurement error then the full conditional distribution can be described as the product of these model distributions. The flexibility of this design means that the authors can incorporate multiple measuring instruments or repeated measurements, as well as a 'gold standard', by defining different models for the unknown risk factor and associated distributions depending on the design of the validation study. As the measurement error model parameters are defined by their prior distributions, the more information that is available concerning the measurement error model can be reflected in the precision of the prior distribution. If little information is available for the measurement error model parameters then a non-informative prior can be used, thus the model can reflect the accuracy associated with the measurement error model.

The authors conclude their study with some general relative merits of the conditional independence modelling technique.

1. The method is conceptually straightforward.
2. This technique follows the exact structure of the measurement error model without having to make any further model assumptions (for example, in the Rosner method, it is assumed that the disease is rare).
3. This technique also uses all the information that is specified by the model rather than just the first and second moments.

4. This technique allows any uncertainty associated with the distributions of the random variables to be specified through the prior distributions and to be reflected in the estimated posterior distribution.
5. The technique can also take into account both continuous and discrete variables.
6. The technique can also handle missing data as well as any ancillary data that is available.

The authors show that this method is both very flexible and accommodating however, they do mention that this method is computationally expensive and do not consider the issues surrounding the use of MCMC methods. This paper was written before the software BUGS was made available.

Dellaportas and Stephens (1995) also look at Bayesian methods with MCMC to estimate the model parameters but do not supply a strategy on how to actually implement the method and do not discuss the effect that a large measurement error standard deviation has on the estimation process. Dellaportas and Stephens do point out that in the Bayesian formulation that they have described it does not matter whether the measurement error model is Classical or Berkson; both can be accommodated as in the Bayesian framework they are just random variables with different distributions.

Richardson and LeBlond (1997) take the work of Richardson and Gilks (1993) one step further and consider the problem of when the exposure model, the distribution of the unknown risk factor, is miss-specified. The authors investigate how the estimation

of the regression parameters of interest can be influenced by the misspecification of the parametric form for the prior distribution of X. The result of this study shows that miss-specification of this distribution showed a moderate bias in the estimates and increased posterior standard deviations. This investigation looked at various values for the regression parameter as opposed to increasing size of the standard deviation associated with the measurement error.

Our study will be looking at the effect of increasing the measurement error standard deviation and the effect this has on the estimation of the regression parameters and their associated standard deviations through the estimation of the posterior distribution.

7.5 The Logistic Regression Errors-in-Variables Model in the BUGS formulation

To model the logistic regression errors-in-variables problem the three conditional independence models namely, the disease model, the measurement error model an the exposure model must be defined. For this study we will look at the simple case of a Berkson measurement error model with a single covariate measured with error in line with the work that was conducted in Chapter 5. One of the ways to understand how this model is built in the BUGS formulation is to use a directed acyclic graph (DAG). The idea of a DAG is to show the direction the conditionals are calculated in for the Gibbs Sampling and so the problem is specified in the correct way for the BUGS software to deal with the model. Figure 7.1 depicts the logistic regression measurement error model where x_i denotes the individual true values, y the outcome, Z_i the observed values, β the logistic regression model parameters, π the distribution

of the true values and λ a model parameter specifying the relationship between the true and observed values that is, the measurement error model.

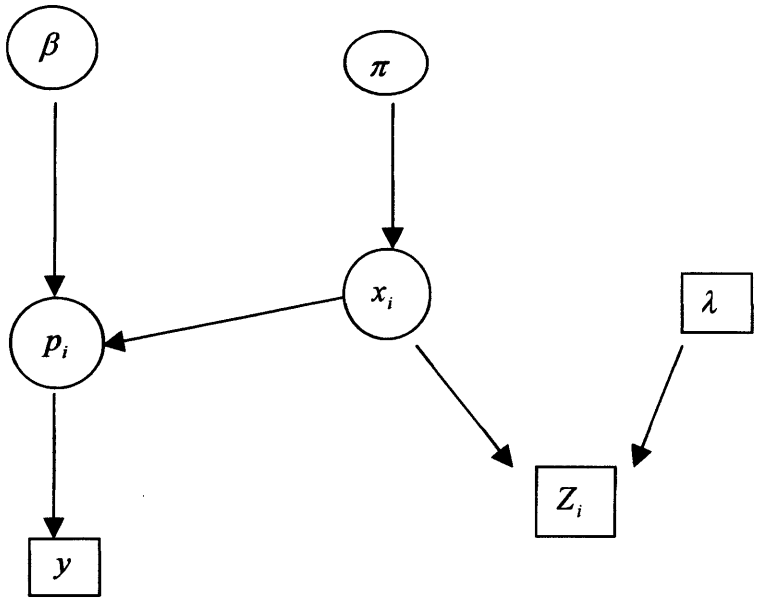


Figure 7.1: DAG for the logistic regression errors in variables problem

There are two types of data depicted within the DAG, one is from the main study data set and the other is from the validation study data set. Any known parameter from the main study data set or that can be estimated from the validation study data set is surrounded by a box within the graph that is, Y and Z are observed from the main study data set and λ is estimated from the validation study. Any parameter that must be estimated is surrounded by a circle within the graph and is referred to as a node. In the case of the measurement error model, the unknown parameters to be estimated are formed by the distributions for β.

The Gibbs Sampling works by updating each node at each iteration of the algorithm. Each node depends on its prior distribution and on the likelihood of its children and their co-parents. The directed graph shows how the joint distribution is factored into a number of components allowing the required conditional distributions to be

constructed for the Gibbs sampling. Therefore, if there is little information with regard to λ then a non-informative prior distribution must be used and so there is less information to estimate the unknown parameters β.

The structure given by Richardson and Gilks (1993) define:

- *The disease model*, $[Y_i|X_i,\beta]$;
- *The measurement model*, $[Z_i|X_i,\lambda]$;
- *The exposure model*, $[X_i|\pi]$.

The full conditional Distribution model can be defined as:

$$[\beta][\lambda][\pi]\prod_i[X_i|\pi]\prod_i[Z_i|X_i,\lambda]\prod[Y_i|X_i,\beta]$$

In this case, vague prior distributions are given for the true model parameters:

- *The disease model*, $P(Y=1|X)=\frac{\exp(\alpha_t+\beta_t X)}{1+\exp(\alpha_t+\beta_t X)}$ where β is defined by α_t and β_t which are given vague priors such as $\alpha_t \sim N(0,100)$ and $\beta_t \sim N(0,100)$
- *The measurement model*, using the Berkson measurement error model, $X=Z+V$ where $V \sim N(0,\sigma_v^2)$
- *The exposure model*, $X \sim N(\mu_x,\sigma_x^2)$.

7.6 Simulation Study

7.6.1 Introduction

Using the BUGS software this study aims to investigate the effect of increasing the standard deviation of the measurement error on the estimation of the regression parameters using a Bayesian analysis approach. A comparison between this Bayesian approach and the relevant classical methods discussed in Chapter 4 will also be

investigated, to see whether the flexibility of a Bayesian approach is accompanied by a method that corrects for measurement error as effectively as the classical methods.

In section 7.2 we discussed the problem of deciding whether the Markov Chain had converged to the stationary distribution and as a result the length of the burn-in period that must be determined before sampling values for the posterior distribution. One of the suggested ways of making this choice is to consider multiple as well as single chains. Therefore, before any simulation study can be conducted, an exploratory analysis is required to decide when convergence has taken place. After this exploratory analysis, the Bayesian approach to the logistic regression measurement error model is studied through simulation. Results of using the Bayesian methods are then compared to those classical methods applicable to the Berkson measurement error model, that is those of Rosner and Reeves, as well as the ordinary logistic regression method which ignores measurement error.

To make this simulation study comparable with those previously conducted in Chapters 5 and 6 the following assumptions are also made:

1. The sample size N, was chosen to be 100, 500 and 1000.
1. The true X values were assumed to follow a Normal Distribution with parameters $X \sim N(30,10^2)$
2. The disease status Y was generated conditional on X for the case where $\alpha_t = -3$ and $\beta_t = 0.1$ and Y=1 with probability

$$P(Y=1|X)=\frac{\exp(\alpha_t+\beta_t X)}{1+\exp(\alpha_t+\beta_t X)}$$

3. The measurement error model is described by the Berkson measurement error model that is, $X=Z+V$ where Z and V are independent and $V \sim N(0,\sigma_v^2)$
4. The measurement error standard deviation σ_v ranged from 0 to 4, increasing by 1, and is therefore assumed to be known.
5. Each simulation was run 20 times. The limited number of simulations is due to the current state of technology that means a simulation run must be conducted manually. Given this, a suitable simulation size was chosen.

For the Bayesian formulation the model parameters α_t and β_t have been given vague prior distributions in the following form $\alpha_t \sim N(0,100)$ and $\beta_t \sim N(0,100)$.

7.6.2 The Initial Exploratory Analysis

Richardson and Gilks (1993) reported that there is good convergence behaviour of the Gibbs sampler in logistic regression measurement error problems. To ensure convergence and therefore determine the burn-in period required for the stationary distribution to be reached an exploratory analysis was conducted. This determined whether the model had good convergence and therefore a simulation study could be conducted based on the burn-in period determined by this study or whether the determination of the burn-in period would have to be conducted for each level of the measurement error standard deviation.

There are a number of issues that should be considered when deciding on a burn-in period and hence convergence to the stationary distribution. When using a Bayesian approach there are two considerations that are under control of the analyst. These are the prior distributions associated with the model parameters and the starting values that are given. A further consideration is the effect of the sample size of the data.

Therefore, the effect of differing model parameter prior distributions, their starting values and the measurement error standard deviation on the convergence to the respective stationary distributions must be investigated.

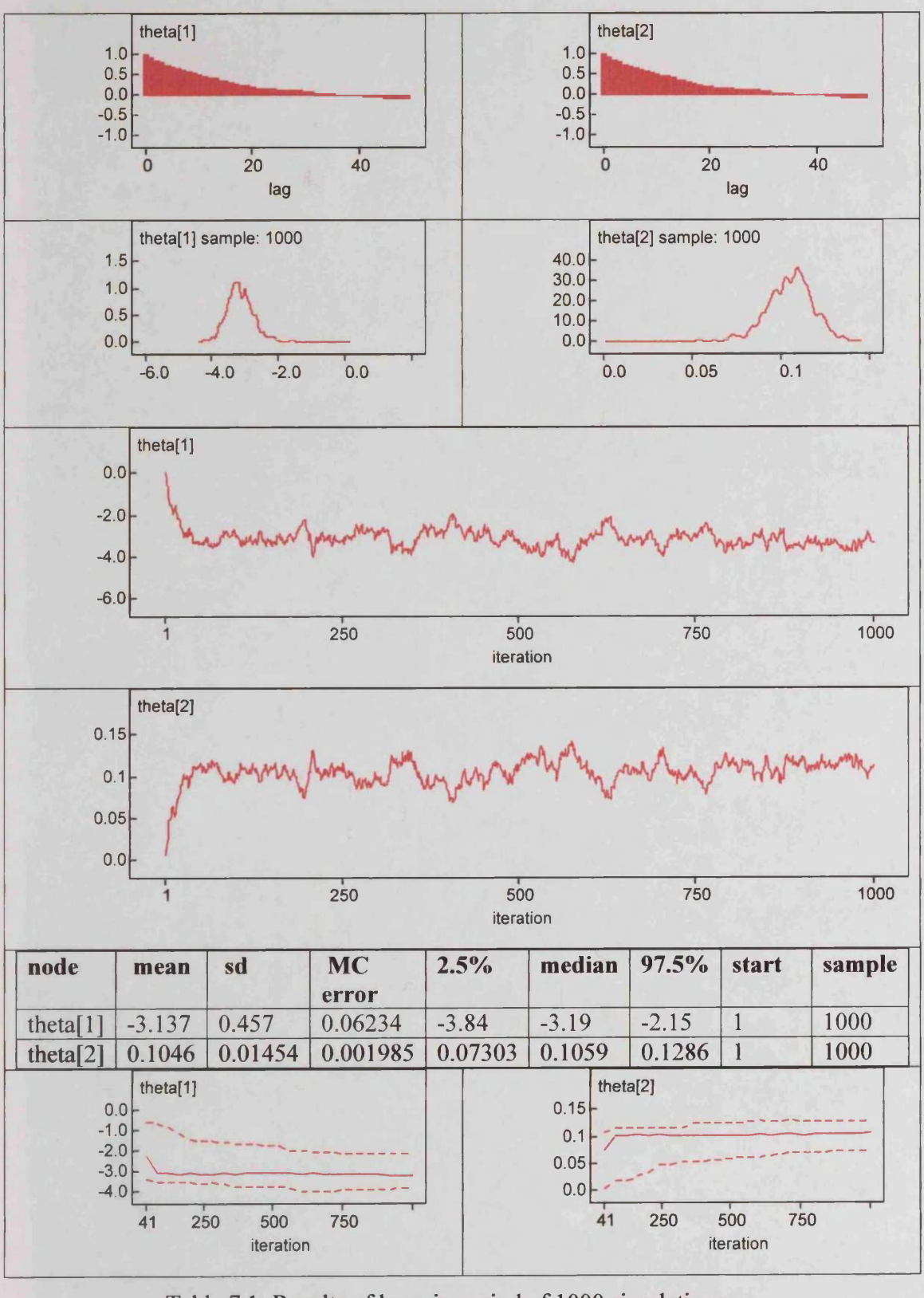
For this exploratory analysis, a sample size of 500 was generated according to the Case A model that is $\alpha_t = -3$ and $\beta_t = 0.1$. However, to keep with the BUGS notation, within this piece of work, α_t is known as theta[1] and β_t is known as theta[2]. The measurement error standard deviation was set to 4. The results of a 1000 burn-in period are presented as well as the results of a 10000 run after the results from the burn-in period have been discarded. Different starting values, as well as differing uninformative and informative prior distributions are also considered.

In each of these cases, the convergence to the stationary distribution is considered from both formal and informal information. Formally, the Gelman-Rubin statistic for multiple chains as described in section 7.2 is used to determine whether convergence has taken place. Other simple diagnostics are also considered including the autocorrelation function, the kernel densities of the model parameters and the individual parameter trace. The posterior distributions are also summarised using simple statistics.

7.6.2.1 Burn in period 1000

The results from the burn-in period of 1000 simulations is displayed in Table 7.1. The first boxes display the auto-correlation function for each of the model parameters. These show that the auto-correlation fades away at about 20-40 lags suggesting that the chain is mixing well. The kernel densities of the model parameters display the

variance associated with the posterior distributions. The variance is made up of two components, the underlying variance of the posterior distribution and the MC error arising from the simulation. From these results it can be seen that the estimated posterior distributions are not smooth suggesting that a longer run is required to reduce the MC error. The parameter traces show the effect of the starting values of the chain for each model parameter, in this case 0.00001 and 0.00001 for theta[1] and theta[2] respectively. In both cases by about lag 50 both traces are ranging around the true values suggesting that the effect of the starting values is minimal. These initial results from a 1000 burn-in period all suggest that the chain needs to be run for a longer period.

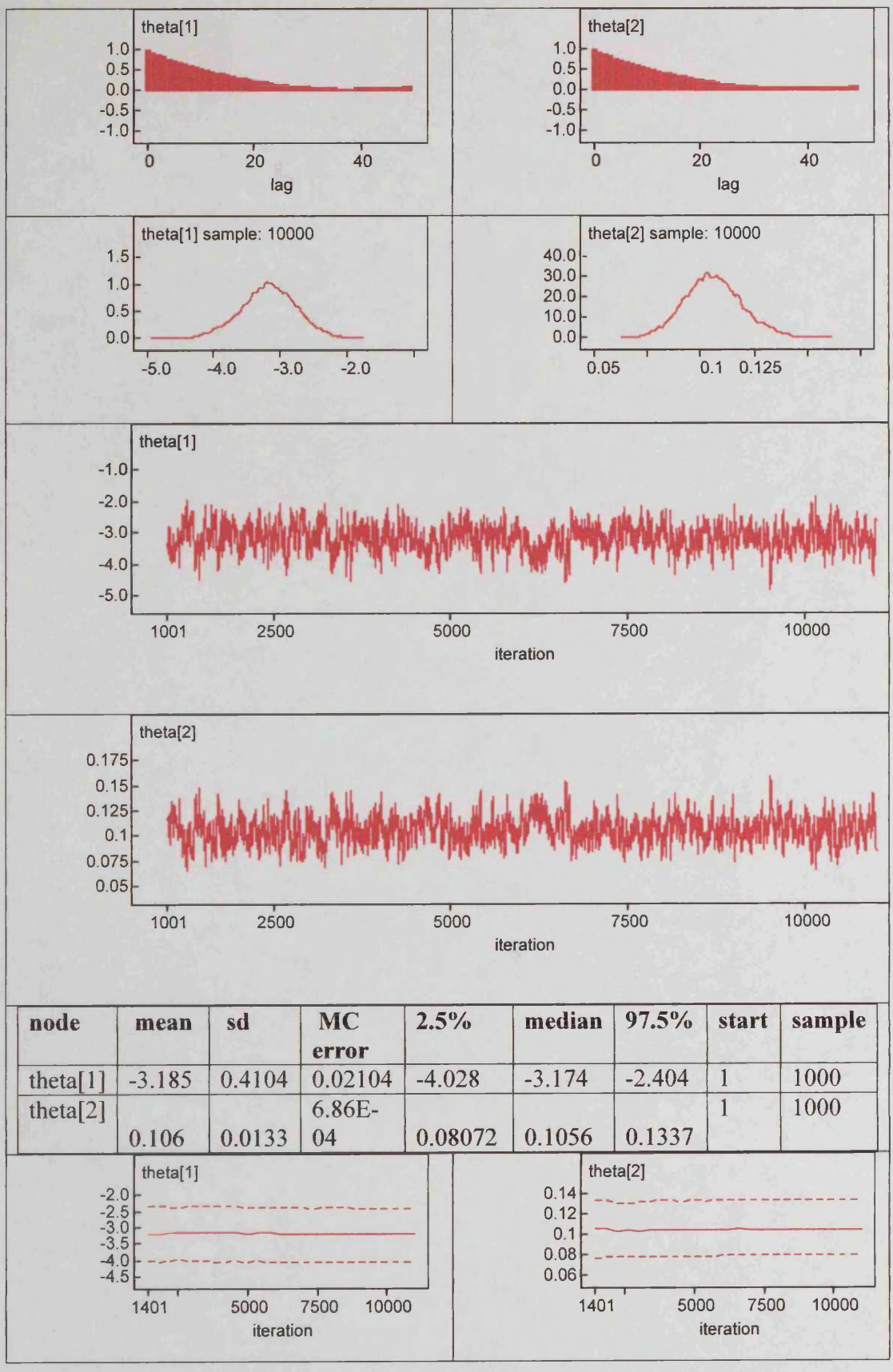


node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-3.137	0.457	0.06234	-3.84	-3.19	-2.15	1	1000
theta[2]	0.1046	0.01454	0.001985	0.07303	0.1059	0.1286	1	1000

Table 7.1: Results of burn-in period of 1000 simulations

Results after burn-in period is discarded and run for 10000 iterations

Table 7.2 displays the results from when the chain was run for 10000 iterations after the initial burn-in period has been discarded. Again the auto-correlation function has been reduced to effectively zero by lag 40. The kernel densities have been smoothed though from comparing the summary statistics, the mean estimates are approximately the same for both the burn-in period and the longer run though the associated standard deviations have decreased.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-3.185	0.4104	0.02104	-4.028	-3.174	-2.404	1	1000
theta[2]	0.106	0.0133	6.86E-04	0.08072	0.1056	0.1337	1	1000

Table 7.2: Results for 10000 run after burn-in period has been discarded

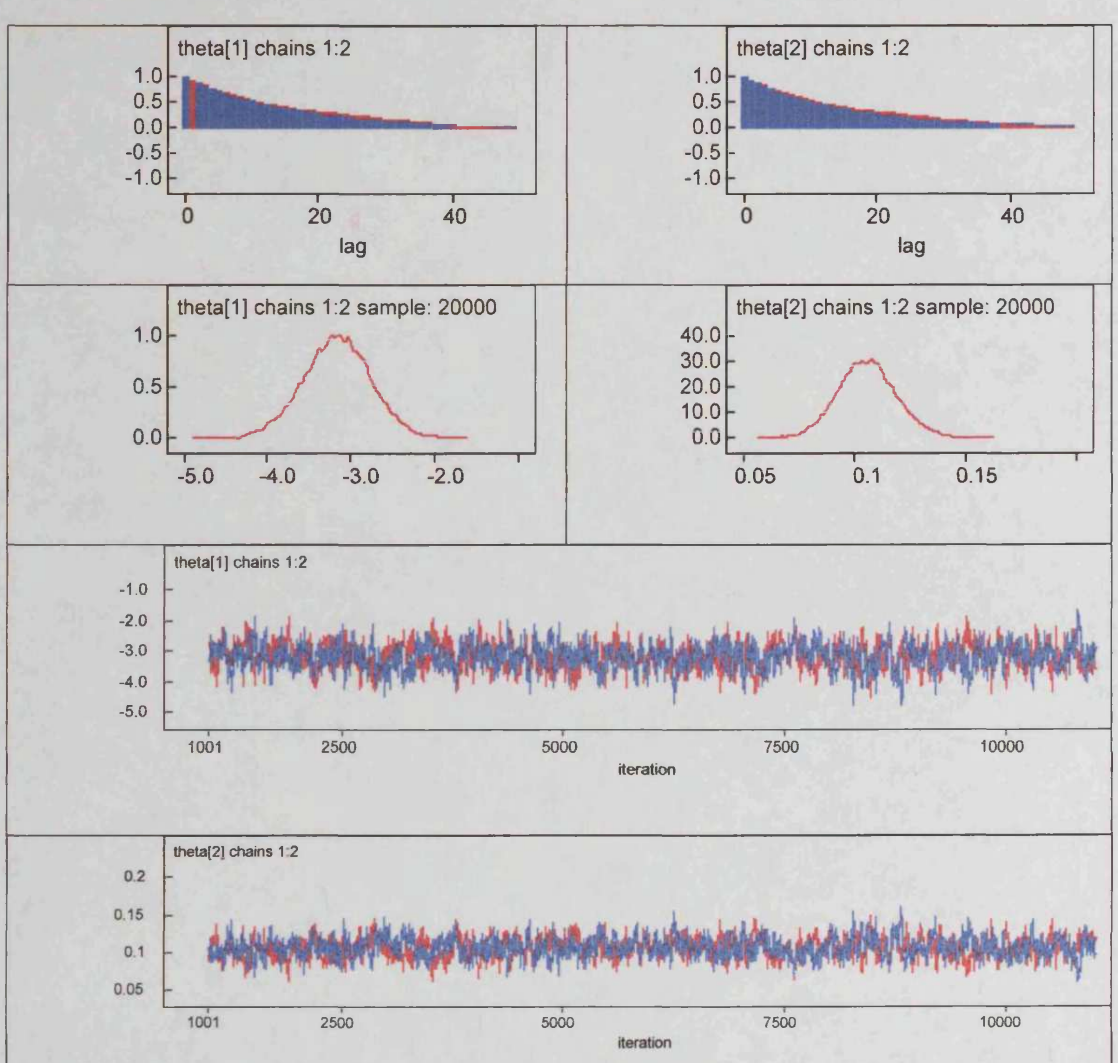
7.6.2.2 Multiple chains after initial 1000 burn-in period discarded and run for 1000 iterations

To see whether the chain has converged, two chains were run so the Gelman-Rubin statistic could be examined. As described earlier the Gelman-Rubin statistic looks at the variation between the two chains compared to the variation within a chain to determine whether convergence to the limiting, and hence stationary, distribution has been established.

There are three lines displayed within the Gelman-Rubin convergence statistic graphs. The first line (green) explains the overall variance of the two chains by displaying the width of the central 80% interval of the pooled runs. The second line (blue) displays the average width of the 80% intervals within the individual runs. Both width intervals have been normalized to have an overall maximum of one. The Gelman-Rubin convergence statistic graph then displays a visual comparison of the overall variance compared to the within variance. The Gelman-Rubin convergence statistic is the ratio of these two quantities. Therefore, to determine convergence to the limiting and hence the stationary distributions, convergence to a ratio of 1 as well as the convergence of both the overall and within interval widths to a stable line is required.

For this exploratory analysis starting values for each run namely (0.1, 0.1) and (0.00001, 0.00001) respectively were chosen. From examination of the green and blue lines it can be seen that by about iteration 4000 the lines are approximately stable. When considering the ratio, this has converged to approximately 1 by about iteration 2500. This convergence experience is in line with that of Richardson & Gilks (1993) who reported that there is good convergence behaviour of the method in logistic regression. Once convergence has been assumed then the model parameters have

converged to their limiting distributions and hence from these stationary distributions, inferences can then be made about the posterior distributions.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-3.176	0.4051	0.015	-3.996	-3.169	-2.395	1000	20002
theta[2]	0.1057	0.0131	4.87E-04	0.08043	0.1055	0.1323	1000	20002

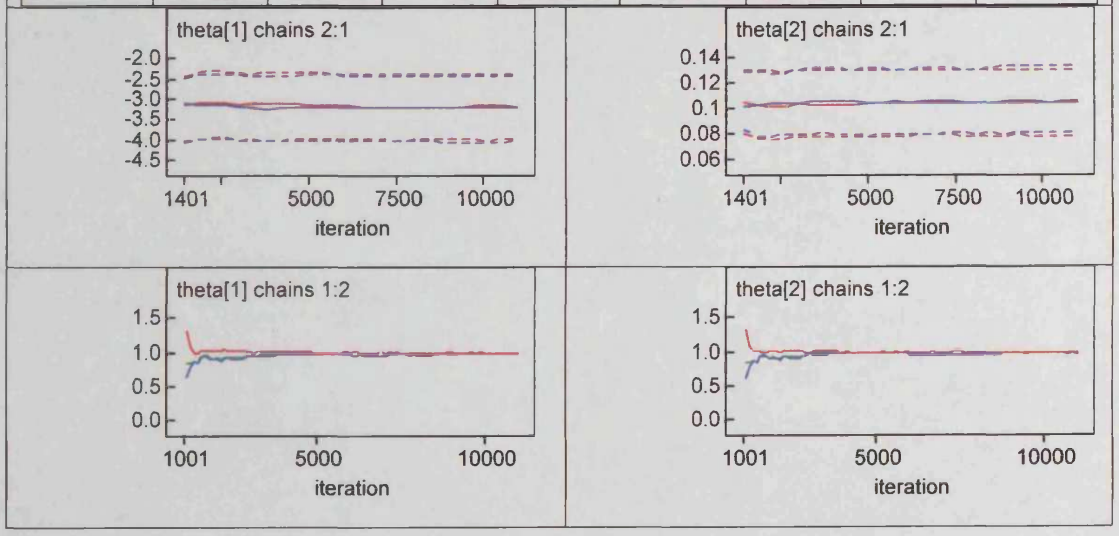


Table 7.3: Results of two chains with the Gelman-Rubin statistic

7.6.2.3 Using Informative Prior Distributions for the Model Parameters

One of the key characteristics of a Bayesian approach to measurement error problems is that it can incorporate any previous information with regard to the model parameters. In the previous analyses, uninformative prior distributions were used. Within the section we consider more informative prior distributions that could have been determined from previous similar studies. The two cases that we considered are:

1. $\theta_1 \sim N(-4,25)$ $\theta_2 \sim N(0.15,1)$
2. $\theta_1 \sim N(-3,4)$ $\theta_2 \sim N(0.1,0.01)$

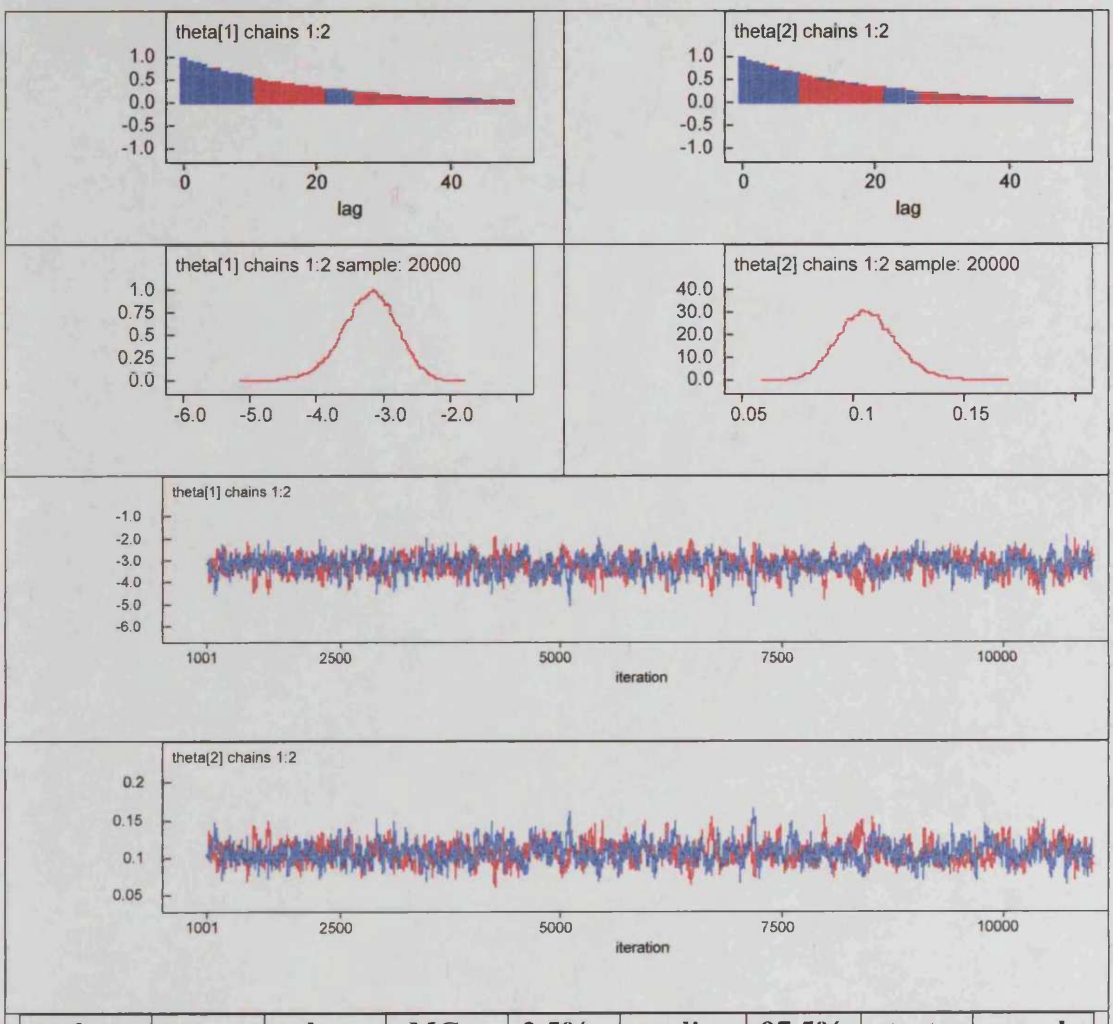
Though the second case is unrealistic with respect to the confidence associated with the prior distributions, it is considered within this analysis simply to see the effect on the posterior distributions of the model parameters and the previously seen large variances associated with the posterior distributions.

A burn-in period of 1000 was run and discarded. The posterior distribution inferences were taken from 10000 iterations.

Case 1 $\theta_1 \sim N(-4,25)$ $\theta_2 \sim N(0.15,1)$

Table 7.4 displays the multiple chain results for Case 1. In comparison with the results obtained from a non-informative prior distributions (Table 7.3) the summary statistics show that even with the increased information with regard to the model parameters this has not had any effect in reducing the associated standard deviations. The pattern of the Gelman-Rubin convergence statistics graphs also show that more informative prior distributions have also not effected the convergence of the chain.

In this case the prior distributions have been over-ridden by the data showing that for this particular problem the posterior distributions are not very sensitive to the prior distributions.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-3.211	0.42	0.01679	-4.101	-3.191	-2.44	1000	20002
theta[2]	0.1069	0.01358	5.44E-04	0.082	0.1062	0.1358	1000	20002

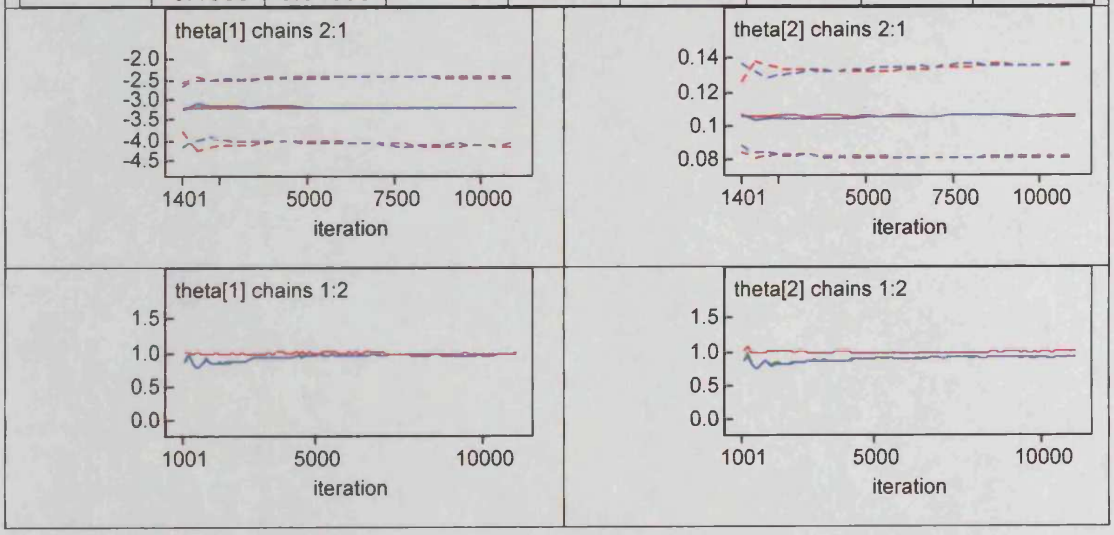
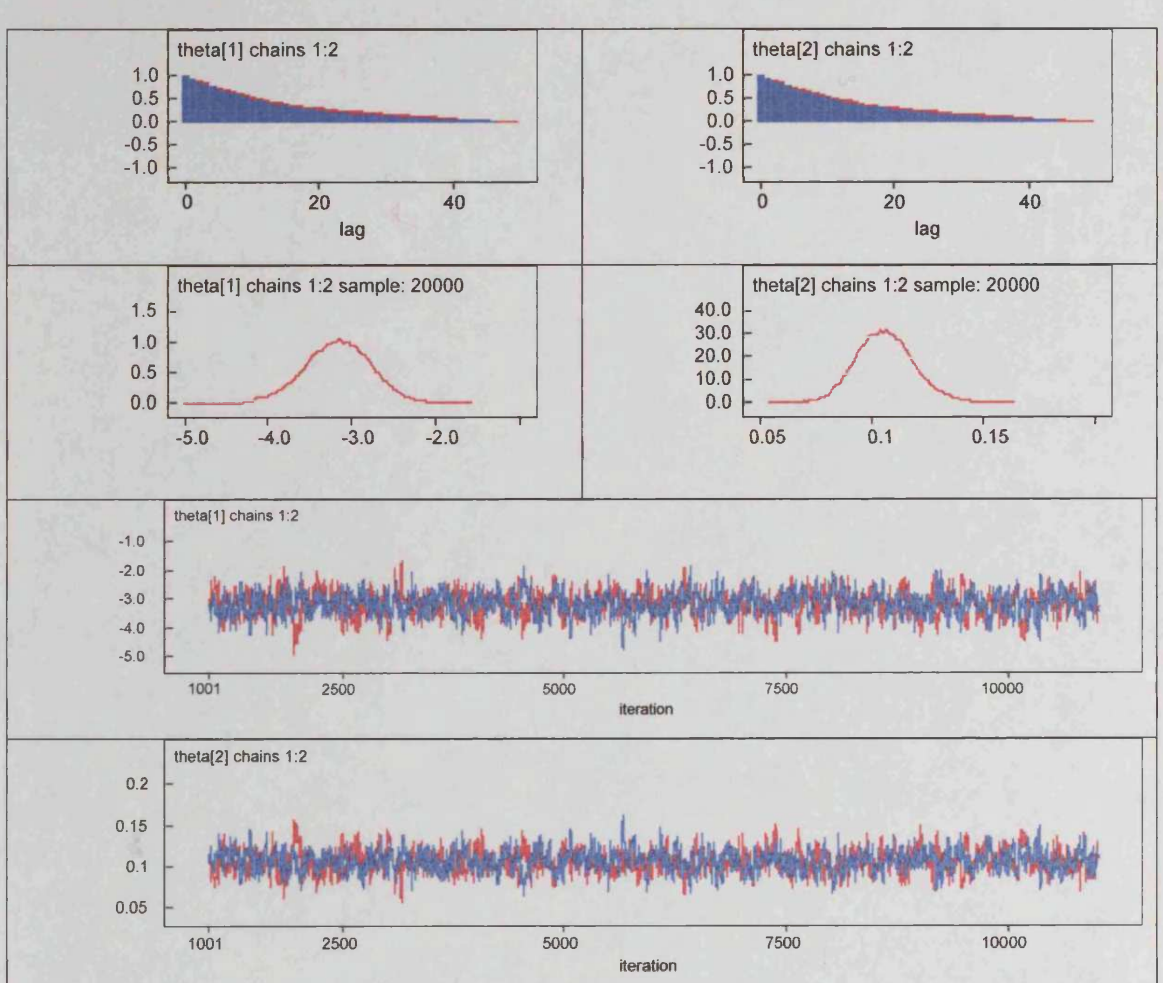


Table 7.4: Case 1 Results

Case 2 $\theta_1 \sim N(-3,4)$ $\theta_2 \sim N(0.1,0.01)$

In this case, the prior distributions are centered on the true values for the model parameters. Again it is observed (Table 7.5) that this extra information does not provide any further precision in the posterior distributions. As we observed for case 1, the model is not very sensitive to the prior distributions and so vague prior distributions for this particular problem in the simulation study can be used.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-3.177	0.3985	0.01503	-3.991	-3.167	-2.426	1000	20002
theta[2]	0.1058	0.01292	4.89E-04	0.08171	0.1055	0.1325	1000	20002

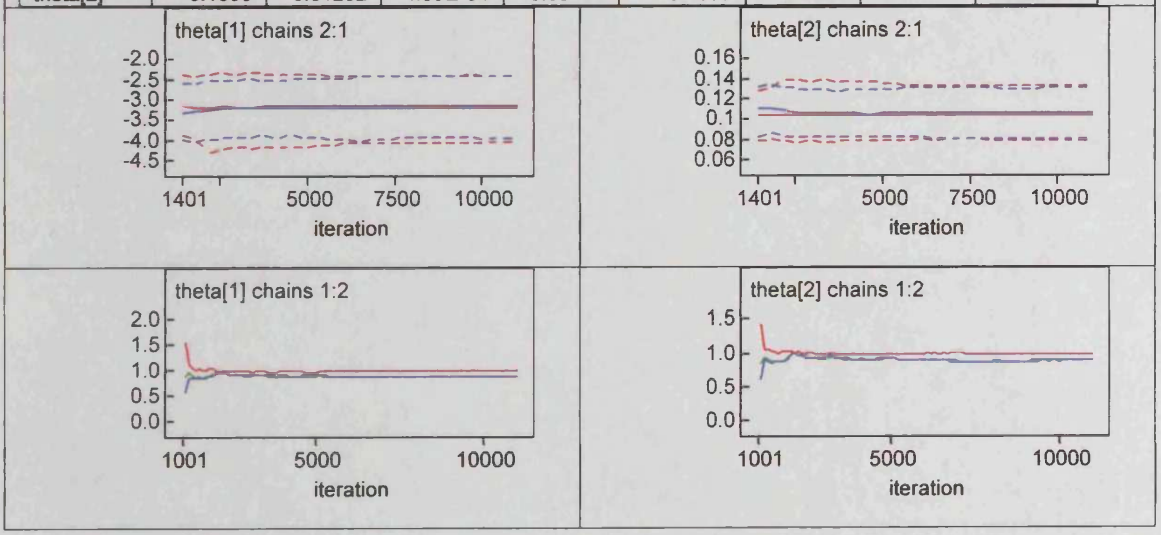


Table 7.5: Case 2 Results

7.6.3 Conclusion of the exploratory analysis

Overall this exploratory analysis has shown that a burn-in period of 1000 is sufficient within this context and that convergence to the stationary distribution could be assumed. Using different starting values has little impact on the convergence to the stationary distribution and therefore the overall inferences made from the posterior distributions are not affected substantially by the starting values. Further to this, little precision is gained in the posterior distributions when more informative prior distributions are used for the model parameters. However, as having access to more informative prior distributions is unrealistic, for the purposes of this study we will assume that only uninformative prior distributions for the model parameters would be available.

7.6.4 Simulation Study Results

The MCMC approach gives a sample from the posterior distribution. This can be summarized in a number of ways. The mean or median of this distribution can be used as a point estimate of the parameter. In addition a 95% credible interval can be derived. This is an interval which contains 95% of the posterior distribution and is usually obtained as the interval between the 2.5^{th} and 97.5^{th} percentiles of the posterior distribution. To illustrate how the Bayesian approach works as the measurement error standard deviation is increased, the program was run 20 times for each value of the measurement error standard deviation. The following results therefore, are the mean values of those 20 simulations for each summary statistic.

N=100

For a sample size of 100, the BUGS mean values of $\hat{\beta}_t$ are positively biased when $\sigma_v = 0$. This positive bias decreases and becomes a negative bias when $\sigma_v = 4$. The associated standard deviation also increased as σ_v increased.

SD of error	Mean	Sd	MC error	2.50%	median	97.50%	Sample
0	-3.45915	0.893425	0.04509	-5.30235	-3.42215	-1.8103	10000
2	-3.43545	0.890275	0.043893	-5.29295	-3.3966	-1.79632	10000
4	-2.9747	0.9216	0.048094	-4.93915	-2.9223	-1.31971	10000

Table 7.6: Case A Sample Size 100 Simulations results for $\hat{\alpha}_t$

SD of error	Mean	Sd	MC error	2.50%	median	97.50%	Sample
0	0.11545	0.02882	0.001455	0.062339	0.114287	0.175185	10000
2	0.113182	0.028372	0.001399	0.06106	0.111959	0.172514	10000
4	0.098081	0.030045	0.001569	0.044415	0.096263	0.162257	10000

Table 7.7: Case A Sample size 100 Simulations results for $\hat{\beta}_t$

N=500

With the increase in sample size, the bias associated with the mean value of $\hat{\beta}_t$ has reduced, as have the standard deviations. However, there is no pattern to the direction of this bias but this could be due to the limited number of simulations. The size of the credible intervals have also reduced along with the MC error.

SD of error	Mean	Sd	MC error	2.50%	median	97.50%	Sample
0	-2.9864	0.36575	0.017524	-3.7251	-2.9799	-2.28855	10000
2	-3.14015	0.37569	0.018304	-3.89625	-3.13205	-2.42285	10000
4	-3.04665	0.404945	0.021222	-3.87425	-3.03585	-2.28475	10000

Table 7.8: Case A Sample Size 500 Simulations results for $\hat{\alpha}_t$

SD of error	Mean	Sd	MC error	2.50%	median	97.50%	Sample
0	0.099484	0.011706	0.000561	0.077208	0.099299	0.123265	10000
2	0.104426	0.012049	0.000586	0.081573	0.10415	0.12873	10000
4	0.101495	0.013077	0.000685	0.076942	0.101126	0.1283	10000

Table 7.9: Case A Sample size 500 Simulations results for $\hat{\beta}_t$

N=1000

The increase in sample size from 500 to 1000 can be seen in the standard deviation values. These have again reduced and as a result so have the credible intervals. The MC error has also reduced. The bias in the mean values of $\hat{\beta}_t$ has also not changed but again this could be due to the number of simulations.

SD of error	Mean	Sd	MC error	2.50%	median	97.50%	Sample
0	-3.14815	0.262585	0.012422	-3.6694	-3.14455	-2.64105	10000
2	-2.93375	0.261335	0.012993	-3.45415	-2.9307	-2.43025	10000
4	-3.03535	0.281895	0.014547	-3.6076	-3.0283	-2.501	10000

Table 7.10: Case A Sample Size 1000 Simulations results for $\hat{\alpha}_t$

SD of error	Mean	Sd	MC error	2.50%	median	97.50%	Sample
0	0.104894	0.008439	0.000399	0.088596	0.104803	0.12169	10000
2	0.097461	0.008395	0.000418	0.081294	0.097361	0.114215	10000
4	0.101616	0.009051	0.000467	0.084464	0.101392	0.119989	10000

Table 7.11: Case A Sample size 1000 Simulations results for $\hat{\beta}_t$

Overall, an increase in sample size in a Bayesian approach, reduced the associated bias in the mean values of $\hat{\beta}_t$, the size of the standard errors and associated credible intervals. Further to this, the MC error had also reduced in size.

We shall now go on to compare the Bayesian approach with the Berkson measurement error models investigated in section 5.2.3 in terms of bias and associated standard errors.

7.6.5 Comparison of BUGS estimators with methods of Ordinary Logistic Regression as well as the methods by Rosner and Reeves

To see how well the Bayesian method works in comparison to the other methods that we have considered in Chapter 5, the Bayesian results are displayed in comparison with the methods of Ordinary Logistic Regression, Reeves and Rosner. Comparisons are made with respect to the mean values of $\hat{\alpha}_t$ and $\hat{\beta}_t$, and the confidence intervals calculated from the empirical standard error associated with the simulation estimates for β_t.

For this investigation, the sample sizes are as in previous simulation studies namely, 100, 500 and 1000. The measurement error standard deviation was taken to have the values 0, 2 and 4. The simulation results are comparing 13000 simulation runs for the ordinary logistic regression, Reeves and Rosner results and the 20 simulation runs for the BUGS results. Again, the limited number of BUGS simulation runs is due to the current inability to run the analysis as a simulation exercise. As in previous display of the results the following tables and graphs are given:

- Table displaying mean value for $\hat{\alpha}_t$ and the empirical standard errors for each method.
- Table displaying mean value for $\hat{\beta}_t$ and the empirical standard errors for each method.
- Graph comparing each methods mean estimates for $\hat{\alpha}_t$
- Graph comparing each methods mean estimates for $\hat{\beta}_t$

The following graphs are displayed with different axis scales so that the detail for each method can be displayed. For a comparison of the methods, the reader is referred to the previous graph.

- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the ordinary logistic regression method.
- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the Reeves method.
- Graph displaying the mean value for $\hat{\beta}_t$ and the two confidence intervals for the Rosner method.
- Graph displaying the mean value for $\hat{\beta}_t$ and the credible interval for the Bayesian method.

N=100

In comparison with the other three methods, the BUGS mean values of $\hat{\beta}_t$ are more positively biased when $\sigma_v = 0$ with a larger standard deviation, Table 7.7. As σ_v increased the level of bias reduced such that there is a negative bias when $\sigma_v = 4$. In comparison with the other methods the Bayesian approach produced the least biased mean value of $\hat{\beta}_t$ when $\sigma_v = 4$. The pattern and direction of the bias associated with the BUGS results follow the same pattern as is observed for the ordinary logistic regression and Rosner results. However, due to the number of simulations that were run for the Bayesian approach the confidence intervals (Figure 7.7) are much larger in size compared to the other methods though they do follow the pattern of the mean values of $\hat{\beta}_t$ explaining that the bias is real.

Figure 7.3, shows that this limited number of simulations has shown that in comparison to the other methods the Bayesian approach would not be used for a small sample size. Further simulation runs may effect this result.

	Olr		Reeves		Rosner		Bayesian	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.1348	0.8551	-3.13477	0.85514	-3.13477	0.85514	-3.45915	1.040664
2	-2.9764	0.82888	-3.09566	0.85264	-3.01336	0.85195	-3.43545	0.866053
4	-2.5993	0.76334	-2.97132	0.84824	-2.70215	0.84919	-2.9747	0.883648

Table 7.12: Sample Size 100 $\hat{\alpha}_t$ summary statistics

	Olr		Reeves		Rosner		Bayesian	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10442	0.02844	0.10442	0.02844	0.10442	0.02844	0.11545	0.032286
2	0.0992	0.02753	0.10426	0.02948	0.10042	0.02856	0.113182	0.025447
4	0.08663	0.02478	0.10446	0.03207	0.09003	0.02779	0.098081	0.033234

Table 7.13: Sample Size 100 $\hat{\beta}_t$ summary results

Figure 7.2: Sample size 100 $\hat{\alpha}_t$ method comparison

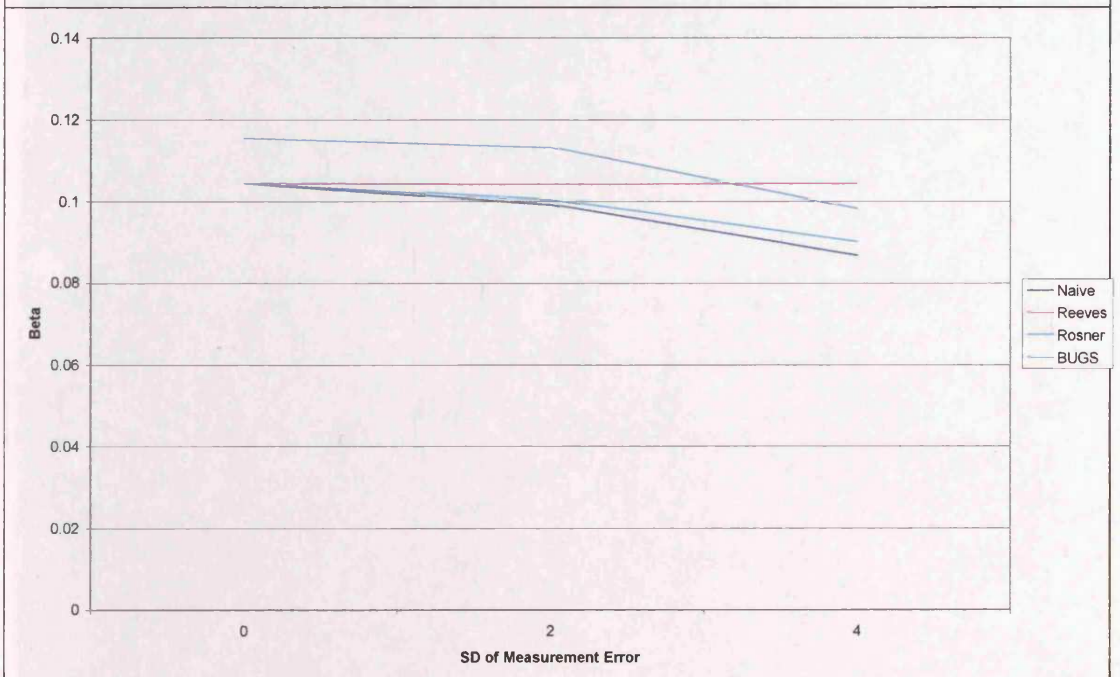


Figure 7.3: Sample size 100 $\hat{\beta}_t$ method comparison

Figure 7.4: Sample size 100 $\hat{\beta}_t$ Olr Confidence Interval comparison

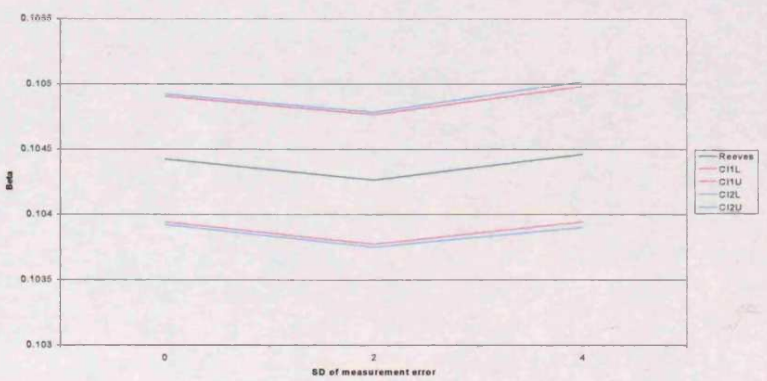


Figure 7.5: Sample size 100 $\hat{\beta}_t$ Reeves Confidence Interval comparison

Figure 7.6: Sample size 100 $\hat{\beta}_t$ Rosner Confidence Interval comparison

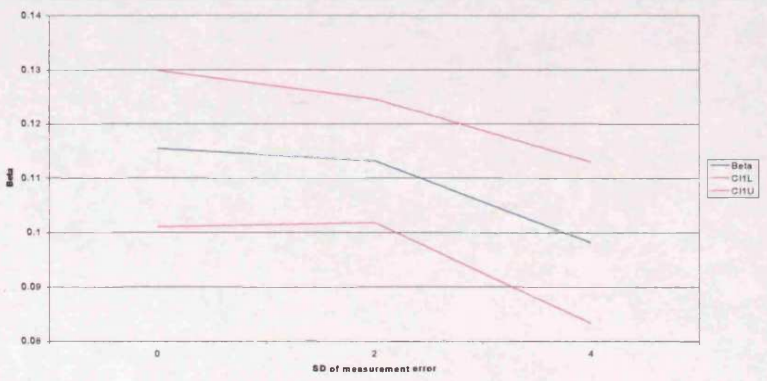


Figure 7.7: Sample size 100 $\hat{\beta}_t$ Bayesian Confidence Interval

N=500

Figure 7.9, shows that with the increased sample size, the bias in the mean values of $\hat{\beta}_t$ from the BUGS results, is now less than the results from the ordinary logistic regression and Rosner methods. The standard deviation is also smaller in comparison to the other methods, though this may just be due to the nature of the datasets within the 20 simulation sample. Figure 7.13 shows a scaled graph of the Bayesian mean values of $\hat{\beta}_t$ and associated confidence interval. As previously observed, the pattern of the confidence intervals mirrors that of the mean values of $\hat{\beta}_t$.

From this small simulation study of 20 runs, the Bayesian approach produced less biased results for a sample size of 500, than the ordinary logistic regression and Rosner methods. However, the Reeves method still produced the least biased results.

	Olr		Reeves		Rosner		Bayesian	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.0242	0.36963	-3.02418	0.36965	-3.02418	0.36965	-2.9864	0.331477
2	-2.8824	0.35878	-2.99403	0.36832	-2.91038	0.36795	-3.14015	0.317157
4	-2.5272	0.33164	-2.87869	0.36562	-2.60469	0.36628	-3.04665	0.397194

Table 7.14: Sample Size 500 $\hat{\alpha}_t$ summary statistics

	Olr		Reeves		Rosner		Bayesian	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10086	0.01199	0.10086	0.01199	0.10086	0.01199	0.099484	0.010298
2	0.09605	0.01151	0.10062	0.01223	0.09698	0.01185	0.104429	0.009076
4	0.08424	0.01058	0.10029	0.01328	0.08681	0.01156	0.101495	0.012142

Table 7.15: Sample Size 500 $\hat{\beta}_t$ summary results

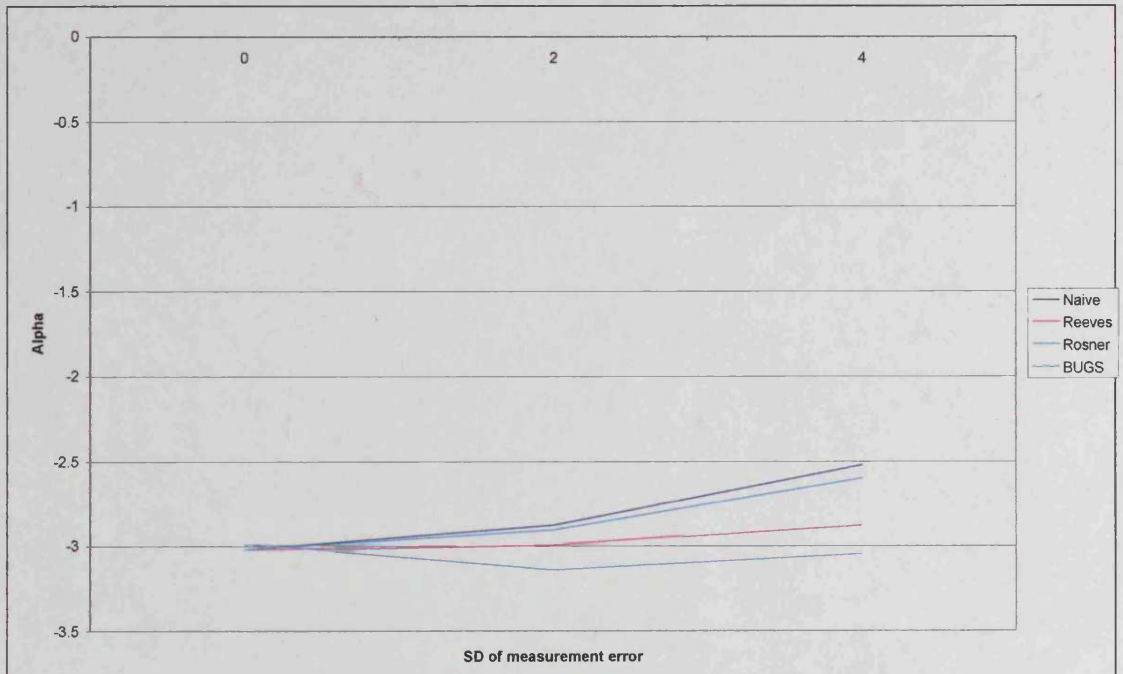


Figure 7.8: Sample size 500 $\hat{\alpha}_t$ method comparison

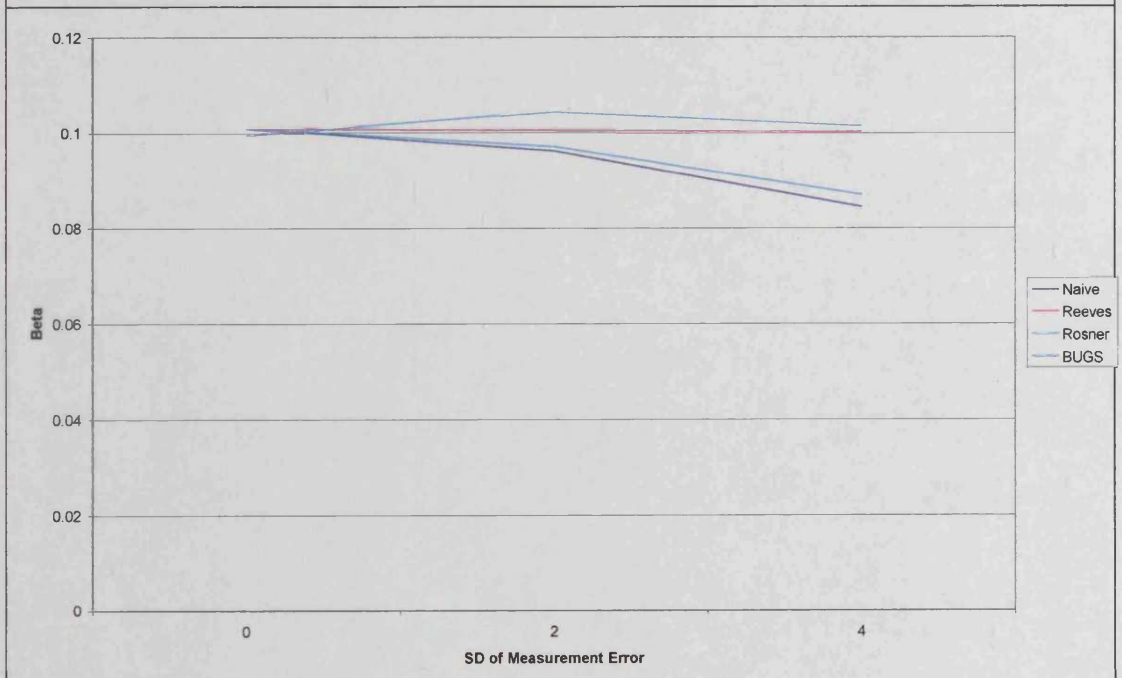


Figure 7.9: Sample size 500 $\hat{\beta}_t$ method comparison

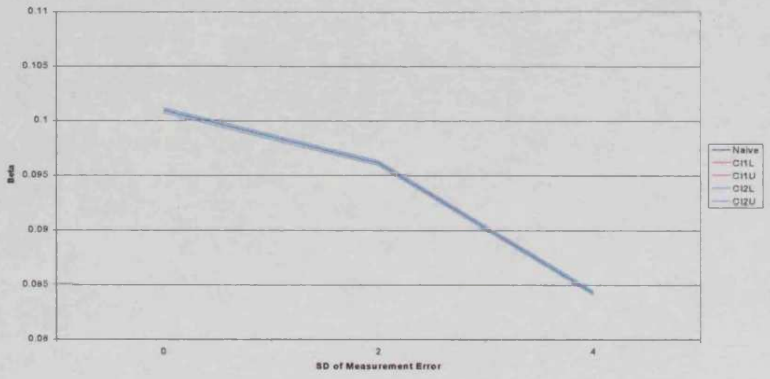


Figure 7.10: Sample size 500 $\hat{\beta}_t$ Olr Confidence Interval comparison

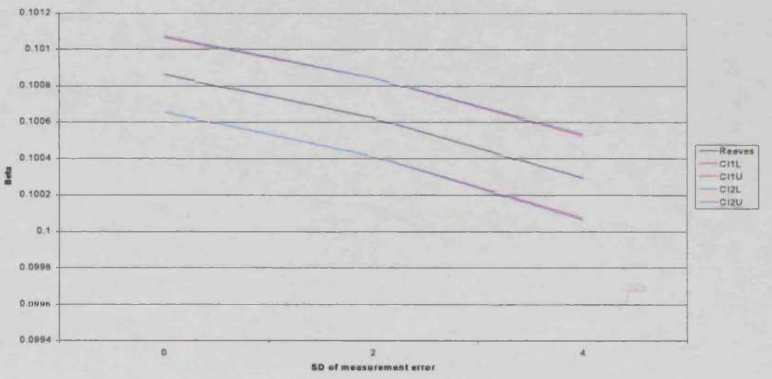


Figure 7.11: Sample size 500 $\hat{\beta}_t$ Reeves Confidence Interval comparison

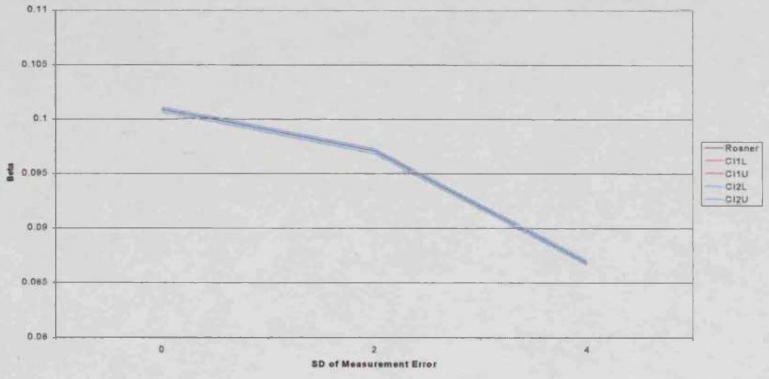


Figure 7.12: Sample size 500 $\hat{\beta}_t$ Rosner Confidence Interval comparison

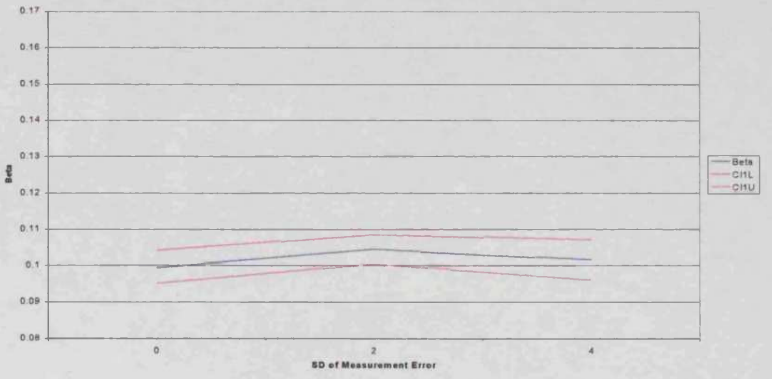


Figure 7.13: Sample size 500 $\hat{\beta}_t$ Baysian Confidence Interval

N=1000

Figure 7.15 shows that with the increased sample size the Bayesian method had mean values of $\hat{\beta}_t$ that are less biased than for the previous sample sizes. The direction of the bias, whether it is positive or negative, changes with the value of σ_v though this may be a feature of the small simulation size.

In comparison, the Reeves method produced the least biased results for this sample size whereas, though the Bayesian approach results were slightly more positively biased, there was little difference between the two methods. A longer simulation run for the Bayesian approach may show that theses methods are comparable.

	Olr		Reeves		Rosner		Bayesian	
σ_v	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$	$\hat{\alpha}_t$	$SE(\hat{\alpha}_t)$
0	-3.0164	0.26048	-3.01642	0.26049	-3.01642	0.26049	-3.14815	0.29155
2	-2.8711	0.25269	-2.98182	0.25936	-2.89814	0.25909	-2.93375	0.220327
4	-2.5171	0.23362	-2.86577	0.25732	-2.5918	0.2578	-3.03535	0.347355

Table 7.16: Sample Size 1000 $\hat{\alpha}_t$ summary statistics

	Olr		Reeves		Rosner		Bayesian	
σ_v	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE_E(\hat{\beta}_t)$
0	0.10053	0.0085	0.10053	0.0085	0.10053	0.0085	0.104894	0.008897
2	0.09569	0.00809	0.1002	0.00859	0.09659	0.00832	0.097461	0.006354
4	0.08392	0.00747	0.09975	0.00934	0.08639	0.00814	0.101616	0.011035

Table 7.17: Sample Size 1000 $\hat{\beta}_t$ summary results

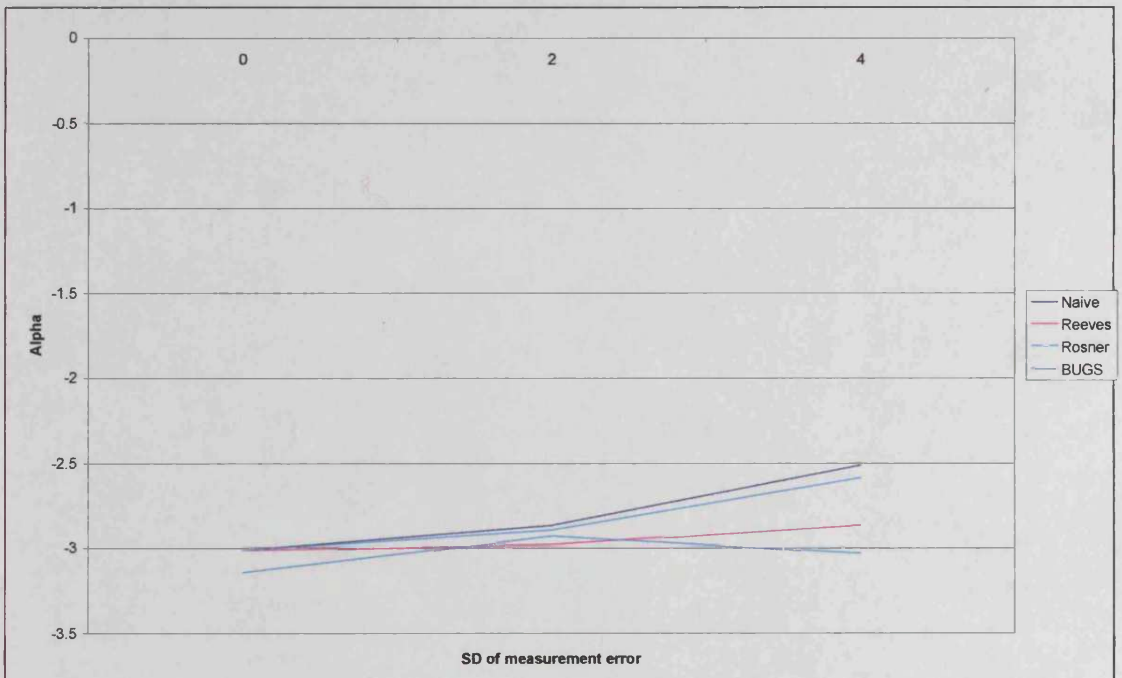


Figure 7.14: Sample size 1000 $\hat{\alpha}_t$ method comparison

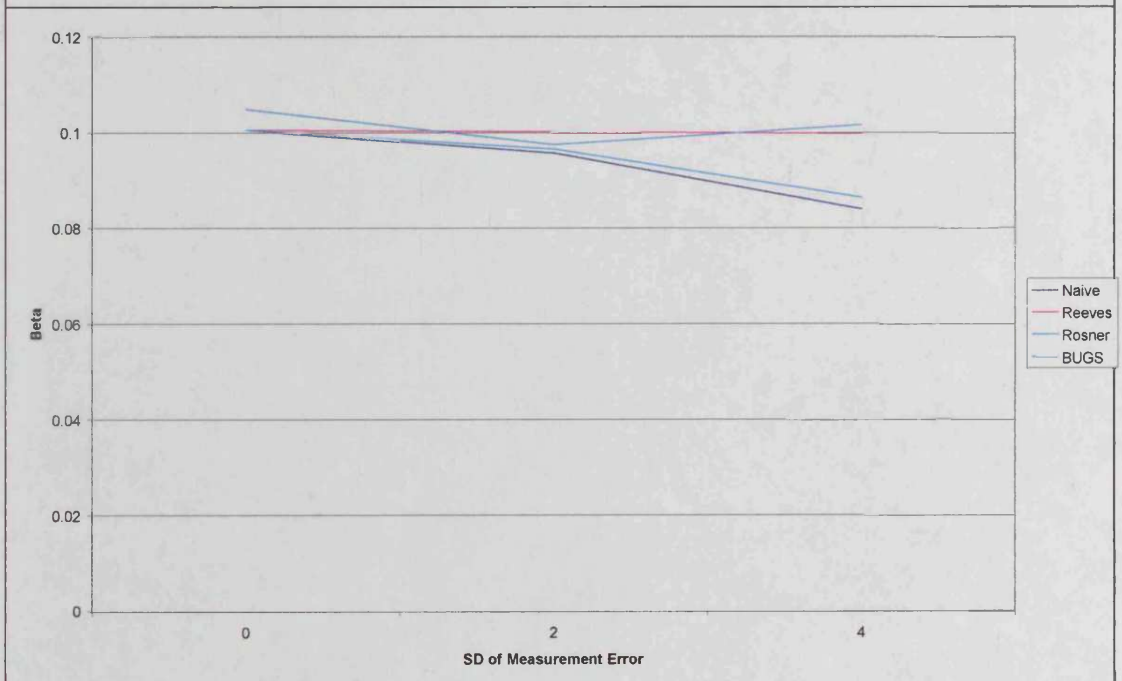


Figure 7.15: Sample size 1000 $\hat{\beta}_t$ method comparison

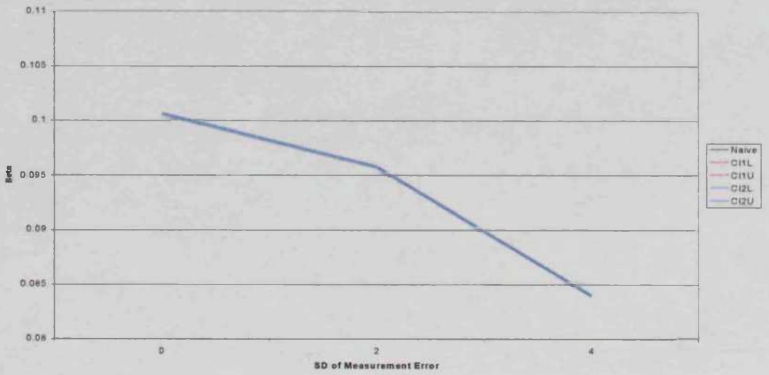


Figure 7.16: Sample size 1000 $\hat{\beta}_t$ Olr Confidence Interval comparison

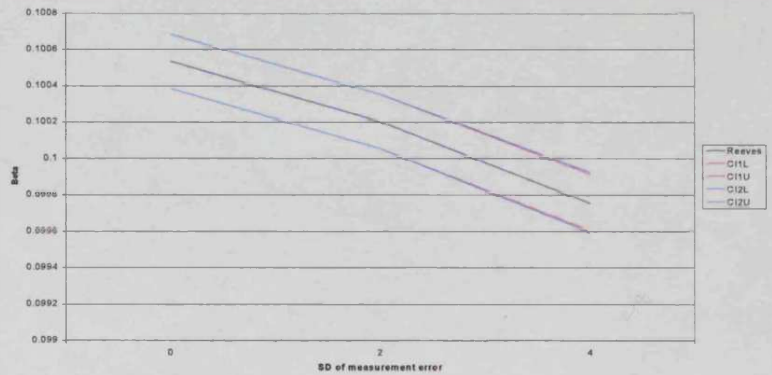


Figure 7.17: Sample size 1000 $\hat{\beta}_t$ Reeves Confidence Interval comparison

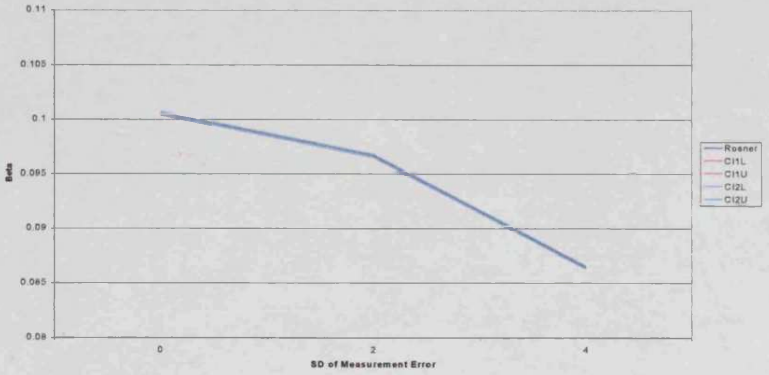


Figure 7.18: Sample size 1000 $\hat{\beta}_t$ Rosner Confidence Interval comparison

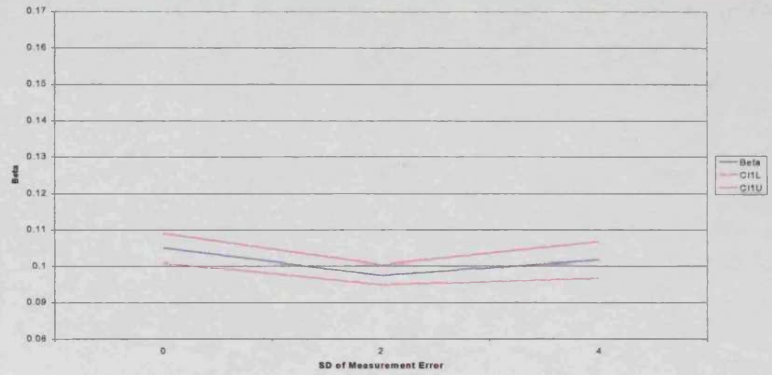


Figure 7.19: Sample size 1000 $\hat{\beta}_t$ Bayesian Confidence Interval

7.6.6 Discussion

For a small sample size, comparing the mean estimates for β_t across the methods, the Bayesian method produced mean values for $\hat{\beta}_t$ that are more biased compared to the other methods. As the sample size increased this level of bias reduced such that, the Bayesian approach produced less biased results than the methods of logistic regression and Rosner.

Due to the technical difficulty in running BUGS in a batch situation only a limited number of simulation runs could be conducted so this simulation study was not comparing comparable results. However, this study has shown that for a sample size of 500 and above, the Bayesian approach was nearly as good at estimating the model parameters as the Reeves approach. A larger simulation study would have to be conducted in order to compare the two methods further. As a result, this simulation study found that the Reeves method still produces the best estimates of the model parameters.

7.7 Conclusion

With the advent of the BUGS software, the Bayesian approach to statistical problems has come into the mainstream making it more accessible to practitioners. In terms of the logistic regression measurement error problem, the model can be described in a Bayesian formulation and posterior distributions for the model parameters can be obtained using the BUGS software. This lead to the question as to whether the Bayesian approach is viable option to use in correcting for measurement error in medical problems.

This chapter has introduced at a high level the concepts associated with the Bayesian approach and the main issues that must be understood in order to carry out a Bayesian analysis. These included the determination of the convergence of the chain to the stationary distribution and hence the limiting distribution as well the effect of informative and uninformative prior distributions for the model parameters.

In the case of the logistic regression measurement error problem, the markov chain mixed well and seemed to have good convergence even when there was large measurement error associated with the measurement of the explanatory variable. Plus, for a sample size of 500, the posterior distributions of the model parameters did not seem to be sensitive to the prior distributions of the model parameters.

The BUGS software was used to run a small simulation study in order to compare the Bayesian approach with the ordinary logistic regression, Rosner and Reeves methods, in order to see which method would be best to use in the measurement error problem. Despite the small number of simulations, the Bayesian approach proved to almost

produce as unbiased mean estimates of the model parameters as the Reeves method. However, from this small simulation study it must be concluded that the Reeves method produced the least biased results.

Chapter 8

8 Identifiability of the logistic regression errors in variables model

8.1 Introduction

Throughout the previous chapters we have investigated the various methods that are available to correct for measurement error in the logistic regression model. This enabled us to form conclusions as to which methods are the best at estimating the model parameters in a number of situations. Foremost in these investigations is the assumption that the measurement error distribution could be estimated from a validation or reproducibility study. However, in some situations it is not always possible to incorporate such a study, or perhaps the study under investigation had already been conducted without a validation study. Indeed a 'gold standard' may not

always be available. In such situations none of the previous methods could be used and therefore a way of estimating the measurement error distribution, as well as the model parameters α_t and β_t, from a single study data set would be desirable.

In this chapter we consider the logistic regression measurement error problem in terms of identification where the term identifiability means that all the model parameters, including those associated with the measurement error, can be estimated uniquely from a single study data set. In the case of the linear regression measurement error problem, it was not identifiable if both the measurement error standard deviations were unknown. For the logistic regression problem we shall show that the problem is identifiable. We also consider the practical implementation of this theory. Further investigation is given to the estimation of the model parameters through the maximisation of the log-likelihood using both a classical and Bayesian approach.

8.2 Identification of the logistic regression model parameters

8.2.1 Classical Model

Work has been conducted on the theory of estimating all the model parameters including the measurement error variance from a single study data set. Effectively this would remove the need for a separate validation study to estimate the measurement error variance and hence would reduce the need and associated costs of conducting such a study. The paper by Kuchenhoff (1994) shows that all the model parameters in the simple logistic model namely α_t, β_t, μ_x, σ_x^2 and σ_v^2 are estimable from the main study data set. However, it does not give a method for implementing this theory.

Kuchenhoff defines the simple logistic regression model to be

$$P(Y=1|X=x)=G(\alpha_t+\beta_t x)$$

$$Z=X+V$$

$$V\sim N(0,\sigma_v^2)$$

$$X\sim N(\mu_x,\sigma_x^2)$$

(X,Y) and V are independent

$$G(t)=(1+\exp(-t))^{-1}$$

The following theorem is quoted from Kuchenhoff's paper.

Theorem: *In the simple logistic regression model with errors in the variables all parameters are identifiable if* $\sigma_x \neq 0$ *and* $\beta_t \neq 0$ *from the main study data set, where identifiability means that the following mapping is injective*

$$(\sigma_v,\sigma_x,\mu_x,\alpha_t,\beta_t)\rightarrow D(Y,Z)$$

and D(Y,Z) is the joint distribution of Y, Z generated by the model.

Kuchenhoff omitted some of the details in his proof; they are supplied here.

From equation (3.6) the function $Q(Z)$ is defined to be

$$Q(Z)=P(Y=1|Z)=\int P(Y=1|X)f(X|Z)dX$$

The conditional distribution of X given Z is normal with the parameters

$$\mu_{X|z}=\frac{\sigma_x^2}{\sigma_z^2}z+\frac{\sigma_v^2}{\sigma_z^2}\mu_x$$

$$\sigma_{X|z}^2=\frac{\sigma_x^2\sigma_v^2}{\sigma_z^2}$$

Substituting into $Q(Z)$, which is then re-arranged, the joint distribution of Y and Z can be written in the form

$$Q(z) = \int G\left(\alpha_t + \beta_t \left(\frac{\sigma_x \sigma_v}{\sigma_z} u + \frac{\sigma_x^2}{\sigma_z^2} z + \frac{\sigma_v^2}{\sigma_z^2} \mu_x \right) \right) \varphi(u) du$$

where $\varphi(.)$ is the pdf of a standard normal distribution.

We write this as

$$Q(z) = \int G\left(c_1 + c_2 z + \sqrt{c_3} u\right) \varphi(u) du$$

where

$$c_1 = \alpha_t + \beta_t \frac{\sigma_v^2}{\sigma_z^2} \mu_x$$

$$c_2 = \beta_t \frac{\sigma_x^2}{\sigma_z^2}$$

$$c_3 = \beta_t^2 \frac{\sigma_x^2 \sigma_v^2}{\sigma_z^2}$$

and $\int$ stands for $\int_{-\infty}^{\infty}$ unless otherwise stated.

There are five parameters to be estimated namely α_t, β_t, μ_x, σ_x^2 and σ_v^2.

Rearranging the above equations we obtain

$$\sigma_v^2 = \frac{c_3 \sigma_z^2}{c_2^2 \sigma_z^2 + c_3}$$

$$\beta_t = c_2 + \frac{c_3}{c_2 \sigma_z^2}$$

$$\alpha_t = c_1 - \frac{c_3}{c_2 \sigma_z^2} \mu_x$$

Also, from the measurement error model,

$$\mu_x = \mu_z$$

$$\sigma_x^2 = \sigma_z^2 - \sigma_v^2$$

Since μ_z and σ_z are easily estimable from the observed data, then provided c_1, c_2, and c_3 can be estimated, the model parameters are identifiable.

We will now show that the values c_1, c_2, and c_3 can be uniquely determined by the function Q in the following way.

$$c_2 = \lim_{z\to\infty} \frac{Q'(z)}{Q(z)} + \lim_{z\to-\infty} \frac{Q'(z)}{Q(z)} \tag{8.1}$$

$$c_1 = -Q^{-1}(0.5)c_2 \tag{8.2}$$

$$c_3 = \left(K^{-1}Q\left(c_2^{-1}(1-c_1)\right)\right)^2 \tag{8.3}$$

where the function $K: \Re_0^+ \to \Re$ is defined by:

$$K(t) = \int G(1+tu)\phi(u)du$$

The following proof for (8.1) will assume that $c_2 > 0$. A similar argument can be constructed for the case of $c_2 < 0$. If $Q(z)$ is differentiated with respect to z then

$$Q'(z) = \int c_2 G'\left(c_1 + c_2 z + \sqrt{c_3}u\right)\phi(u)du$$

$$= c_2 \int G'(w)\phi(u)du$$

where

$$w = c_1 + c_2 z + \sqrt{c_3}u$$

Now $G(w)$ satisfies $G'(w) = G(w)(1-G(w))$. Hence

$$Q'(z) = c_2 \int G(w)(1-G(w))\phi(u)du$$

As $z \to \infty$, for $c_2 > 0$, $w \to \infty$ so $G(w) \to 1$ and $1-G(w) \to 0$. Since G is bounded, interchanging the order of integration and taking the limit

$$Q'(z) \to 0 \qquad \text{as } z \to \infty$$

Similarly

$$Q(z) \to 1 \qquad \text{as } z \to \infty$$

and hence

$$\lim_{z\to\infty} \frac{Q'(z)}{Q(z)} = 0$$

It is also noted that

$$\frac{Q'(z)}{Q(z)} = \frac{\int c_2 G(c_1 + c_2 z + \sqrt{c_3}u)\{1 - G(c_1 + c_2 z + \sqrt{c_3}u)\}\varphi(u)du}{\int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du} \tag{8.4}$$

$$= \frac{c_2 \int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du - c_2 \int G^2(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}{\int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}$$

$$= c_2 - c_2 \frac{\int G^2(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}{\int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}$$

Since $G(t) \le e^t$,

$$\int G^2(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du \le \int \exp[2(c_1 + c_2 z + \sqrt{c_3}u)]\varphi(u)du$$

$$= \int \exp(2c_1 + 2c_2 z + 2\sqrt{c_3}u)\varphi(u)du$$

$$= \exp(2c_1 + 2c_2 z)\int \exp(2\sqrt{c_3}u)\varphi(u)du \tag{8.5}$$

The MGF of a standard normal distribution is

$$M(t) = E(\exp(tu)) = \int \exp(tu)\varphi(u)du$$

$$= \exp\left(\frac{1}{2}t^2\right)$$

Hence, the integral in equation (8.5) can be written as

$$\int \exp\left(2\sqrt{c_3}u\right)\varphi(u)du = \exp\left(\frac{1}{2}\left(2\sqrt{c_3}\right)^2\right) = \exp(2c_3) \quad \text{setting } t = 2\sqrt{c_3}$$

Therefore,

$$\int G^2\left(c_1 + c_2 z + \sqrt{c_3}u\right)\varphi(u)du \le \exp(2c_1 + 2c_2 z + 2c_3)$$

Now if $t<0$ then $G(t) \ge \frac{1}{2}e^t$. If $z < -c_1 c_2^{-1} = -\frac{c_1}{c_2}$ then $c_1 + c_2 z < 0$ and

$$\int_{-\infty}^{0} G\left(c_1 + c_2 z + \sqrt{c_3}u\right)\varphi(u)du \ge \frac{1}{2}\int_{-\infty}^{0} \exp\left(c_1 + c_2 z + \sqrt{c_3}u\right)\varphi(u)du$$

$$= \frac{1}{2}\exp(c_1 + c_2 z)\int_{-\infty}^{0} \exp\left(\sqrt{c_3}u\right)\varphi(u)du$$

Now

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} \exp(au)\exp\left(-\frac{1}{2}u^2\right)du$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} \exp\left(-\frac{1}{2}\left(u^2 - 2au + a^2\right)\right)\exp\left(\frac{1}{2}a^2\right)du$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} \exp\left(-\frac{1}{2}(u-a)^2\right)du \exp\left(\frac{1}{2}a^2\right)$$

By taking $w = u - a$ then

$$= \exp\left(\frac{1}{2}a^2\right)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-a} \exp\left(-\frac{1}{2}w^2\right)dw$$

$$= \exp\left(\frac{1}{2}a^2\right)\Phi(-a)$$

On substitution of $a = \sqrt{c_3}$

$$\int_{-\infty}^{0} \exp\left(\sqrt{c_3}u\right)\varphi(u)du = \exp\left(\frac{1}{2}c_3\right)\Phi\left(-\sqrt{c_3}\right)$$

Therefore,

$$\int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du \geq \frac{1}{2}\exp\left(c_1 + c_2 z + \frac{1}{2}c_3\right)\Phi\left(-\sqrt{c_3}\right)$$

where Φ is the standard normal distribution function.

$$0 \leq \lim_{z\to-\infty} \frac{\int G^2(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}{\int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}$$

$$\leq \lim_{z\to-\infty} \frac{\exp(2c_1 + 2c_2 z + 2c_3)}{\exp\left(c_1 + c_2 z + \frac{1}{2}c_3\right)\frac{1}{2}\Phi\left(-\sqrt{c_3}\right)} \to 0 \qquad \text{as } z \to -\infty \text{ since } c_2 > 0$$

Hence,

$$\lim_{z\to-\infty} \frac{\int G^2(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du}{\int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du} = 0$$

So in (8.1)

$$\lim_{z\to-\infty} \frac{Q'(z)}{Q(z)} = c_2 \qquad \text{and} \qquad \lim_{z\to\infty} \frac{Q'(z)}{Q(z)} = 0$$

Therefore,

$$c_2 = \lim_{z\to-\infty} \frac{Q'(z)}{Q(z)} + \lim_{z\to\infty} \frac{Q'(z)}{Q(z)}$$

which proves (8.1).

To prove (8.2) let $a = \sqrt{c_3}$ then

$$\int_{-\infty}^{\infty} G(au)\varphi(u)du$$

$$= \int_{-\infty}^{0} G(au)\varphi(u)du + \int_{0}^{\infty} G(au)\varphi(u)du$$

$$= \int_{0}^{\infty} G(-au)\varphi(-u)du + \int_{0}^{\infty} G(au)\varphi(u)du$$

$$= \int_0^\infty G(-au)\varphi(u)du + \int_0^\infty G(au)\varphi(u)du \qquad \text{by symmetry of } \varphi$$

$$= \int_0^\infty (G(-au)+G(au))\varphi(u)du$$

$$= \int_0^\infty \left(\frac{1}{1+\exp(-au)} + \frac{1}{1+\exp(au)}\right)\varphi(u)du$$

$$= \int_0^\infty \frac{2+\exp(au)+\exp(-au)}{2+\exp(au)+\exp(-au)}\varphi(u)du = \frac{1}{2}$$

Therefore,

$$\int G(c_3u)\varphi(u)du = \frac{1}{2}$$

Taking $z = \dfrac{-c_1}{c_2}$ in $Q(z) = \int G(c_1 + c_2 z + \sqrt{c_3}u)\varphi(u)du$

we have

$$Q\left(\frac{-c_1}{c_2}\right) = \int G(c_1 - c_1 + \sqrt{c_3}u)\varphi(u)du$$

$$= \int G(\sqrt{c_3}u)\varphi(u)du = \frac{1}{2}$$

Therefore,

$$Q\left(\frac{-c_1}{c_2}\right) = \frac{1}{2}$$

As $Q'>0$, Q is injective and therefore has an inverse. Hence,

$$-\frac{c_1}{c_2} = Q^{-1}\left(\frac{1}{2}\right)$$

so that equation (8.2) is true.

To prove the final result of equation (8.3)

$$c_3 = \left(K^{-1}Q\left(c_2^{-1}(1-c_1)\right)\right)^2$$

we must show that the function K

$$K(t) = \int G(1+tu)\varphi(u)du$$

is invertible on $\Re_0^+$. Differentiating the function K with respect to t

$$K'(t) = \int G'(1+tu)u\varphi(u)du$$

$G'(v)$ is symmetric about 0 and is monotonic decreasing for $tv > 0$ and monotonic increasing for $v < 0$. So if, $x > y > 0$, $G'(x) < G'(y)$. If, $x > |y|$ and $y < 0$, $G'(x) < G'(|y|) = G'(y)$. Therefore, if $x > |y|$, $G'(x) < G'(y)$. So for $u,t > 0$

$$|1+tu| > |1-tu| \Rightarrow G'(1+tu) < G'(1-tu)$$

Hence,

$$\int_0^\infty G'(1+tu)u\varphi(u)du < \int_0^\infty G'(1-tu)u\varphi(u)du$$

$$= \int_0^{-\infty} G'(1+tw)(-w)\varphi(w)-dw \qquad \text{where } w = -u$$

$$= \int_0^{-\infty} G'(1+tw)w\varphi(w)dw$$

$$= -\int_{-\infty}^0 G'(1+tw)w\varphi(w)dw$$

Therefore,

$$\int_0^\infty G'(1+tu)u\varphi(u)du < -\int_{-\infty}^0 G'(1+tu)u\varphi(u)du$$

$$\int_0^\infty G'(1+tu)u\varphi(u)du + \int_{-\infty}^0 G'(1+tu)u\varphi(u)du < 0$$

$$K'(t) = \int_{-\infty}^{\infty} G'(1+tu)u\varphi(u)du < 0$$

So $K'(t) < 0$ for all t and so $K(t)$ is monotonic decreasing and therefore has an inverse. Now

$$Q\left(\frac{1-c_1}{c_2}\right) = \int G\left(c_1 + c_2\left(\frac{1-c_1}{c_2}\right) + \sqrt{c_3}u\right)\varphi(u)du$$

$$= \int G(1+\sqrt{c_3}u)\varphi(u)du$$

$$= K(\sqrt{c_3})$$

Therefore,

$$\sqrt{c_3} = K^{-1}\left(Q\left(\frac{1-c_1}{c_2}\right)\right)$$

and so

$$c_3 = \left[K^{-1}\left(Q\left(\frac{1-c_1}{c_2}\right)\right)\right]^2 \qquad (8.3)$$

Therefore, the three values of c_1, c_2, and c_3 can be uniquely determined by the function Q and the observed model parameters and hence the problem is identifiable.

However, it leaves the practical problem of estimating these three quantities. Kuchenhoff states that though in theory the model parameters can be identified he had found no practical way of estimating them. This point will be discussed later.

8.2.1.1 The multiple logistic regression model

The previous section showed that theoretically the simple logistic regression model is identifiable. This leads to the question as to whether the multiple logistic regression model is also identifiable.

It is assumed that a vector $\boldsymbol{X}$ contains a set of d covariates, where the first d_1 covariates are measured with error and therefore the remaining covariates are assumed to be measured precisely. Hence, the model is

$$P(Y=1|\boldsymbol{Z}=\boldsymbol{z})=\mathrm{G}(\alpha_t+\boldsymbol{x}\boldsymbol{\beta}_t)$$

$$\boldsymbol{Z}=\boldsymbol{X}+\boldsymbol{V} \quad \dim \boldsymbol{X}=d$$

$$\boldsymbol{V}=(\boldsymbol{V}_1',0')' \quad \boldsymbol{V}_1 \sim \mathrm{N}(0,\Sigma_{v1}) \quad \Sigma_{v1} \text{ regular} \quad \dim \boldsymbol{V}_1=\boldsymbol{d}_1 \le \boldsymbol{d}$$

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}_x,\Sigma_x) \quad \Sigma_x \text{ regular}$$

$(\boldsymbol{X},\boldsymbol{Y})$ and $\boldsymbol{V}$ are assumed to be independent

This model implies that for $\mathrm{Q}(\boldsymbol{z})=P(Y=1|\boldsymbol{Z}=\boldsymbol{z})$

$$Q(\boldsymbol{z})=\int G(\alpha_t+\boldsymbol{z}'\Sigma_z^{-1}\Sigma_x\boldsymbol{\beta}_t+\boldsymbol{\mu}_x'\Sigma_z^{-1}\Sigma_v\boldsymbol{\beta}_t+(\boldsymbol{\beta}_t'\Sigma_x\Sigma_z^{-1}\Sigma_v\boldsymbol{\beta}_t)^{\frac{1}{2}}\boldsymbol{u})\,\varphi(\boldsymbol{u})\,\mathrm{d}(\boldsymbol{u})$$

Kuchenhoff proves that without further information concerning the measurement error matrix Σ_{v1}, this model is not identifiable. The system of equations produced by this model that is

$$P(Y=1|\boldsymbol{Z}=\boldsymbol{z})=Q(\boldsymbol{z})=\int \mathrm{G}\ (\boldsymbol{c}_1+\boldsymbol{z}'\boldsymbol{c}_2+\sqrt{\boldsymbol{c}_3}\boldsymbol{u})\,\varphi(\boldsymbol{u})\ \mathrm{d}(\boldsymbol{u})$$

$$\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{\mu}_z,\Sigma_z)$$

$$\Sigma_z=\Sigma_x+\Sigma_v$$

$$\boldsymbol{\mu}_z=\boldsymbol{\mu}_x$$

$$\mathbf{c}_1=\alpha_t+\boldsymbol{\mu}_x'\Sigma_z^{-1}\Sigma_v\boldsymbol{\beta}_t$$

$$\boldsymbol{c}_2=\Sigma_z^{-1}\Sigma_x\boldsymbol{\beta}_t$$

$$c_3 = \beta_t' \Sigma_x \Sigma_z^{-1} \Sigma_v \beta_t$$

results in more parameters to be estimated than there are equations if $d_1 > 1$. However, if Σ_v is a known constant then there are enough equations to estimate the remaining model parameters and hence the model is identifiable.

8.2.2 Berkson Model – An extension to the Kuchenhoff paper

The theory by Kuchenhoff was derived for the classical measurement error model. A natural extension to this work is to consider the Berkson measurement error model, to see if this case is identifiable. In the Kuchenhoff notation, the Berkson measurement error model can be described in the following terms.

$$Q(Z) = P(Y = 1|Z) = \int P(Y = 1|X) f(X|Z) dX$$

where the conditional distribution of X given Z is normal with the parameters

$$\mu_{x|z} = \mu_z$$
$$\sigma_{x|z}^2 = \sigma_v^2 + \sigma_z^2$$

Substituting into $Q(Z)$, which is then re-arranged, the joint distribution of Y and Z can be written in the form

$$Q(z) = \int G(\alpha + \beta x) \frac{1}{\sqrt{\sigma_v^2 + \sigma_z^2}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2(\sigma_v^2 + \sigma_z^2)} (x - z)^2 \right) dx$$

If we make the substitution

$$u = \frac{x - z}{\sqrt{\sigma_v^2 + \sigma_z^2}}$$

then

$$x = z + \sqrt{\sigma_v^2 + \sigma_z^2} u$$

Therefore, $Q(z)$ can be re-arranged to

$$Q(z) = \int G\left(\alpha + \beta z + \beta\sqrt{\sigma_v^2 + \sigma_z^2}\,u\right)\varphi(u)du \qquad (8.6)$$

Using Kuchenhoff's notation (8.6) can be written in the form

$$Q(z) = \int G\left(c_1 + c_2 z + \sqrt{c_3}\,u\right)\varphi(u)du$$

where

$$c_1 = \alpha$$
$$c_2 = \beta$$
$$c_3 = \beta^2\left(\sigma_v^2 + \sigma_z^2\right)$$

For both the classical and Berkson measurement error models there are five parameters to be estimated namely α_t, β_t, μ_x, σ_x^2 and σ_v^2. For the Berkson measurement error model, in terms of c_1, c_2 and c_3, α_t, β_t, and σ_v^2 can be written as

$$\alpha = c_1$$
$$\beta = c_2$$
$$\sigma_v^2 = \frac{c_3}{c_2^2} - \sigma_z^2$$

where σ_z^2 can be estimated from the observed data.

Therefore, the Berkson logistic regression measurement error model is identifiable if c_1, c_2 and c_3 can be estimated.

For the Berkson measurement error model it can be shown that, like the classical measurement error model, c_1, c_2, and c_3 can be uniquely determined by the function Q in the following way.

$$c_2 = \lim_{z\to\infty}\frac{Q'(z)}{Q(z)} + \lim_{z\to-\infty}\frac{Q'(z)}{Q(z)} \qquad (8.1)$$

$$c_1 = -Q^{-1}(0.5)c_2 \qquad (8.2)$$

$$c_3 = \left(K^{-1}Q\left(c_2^{-1}(1-c_1)\right)\right)^2 \tag{8.3}$$

where the function $K: \Re_0^+ \to \Re$ is defined by:

$$K(t) = \int G(1+tu)\varphi(u)du$$

As the proof for the Kuchenoff theory depends on the quantities c_1, c_2, and c_3 and the form of $Q(z)$, the same proof for the Classical measurement error model holds for the Berkson measurement error model. That is, the Berkson logistic regression measurement error model is identifiable if the three values of c_1, c_2, and c_3 can be uniquely determined by the function Q and the observed model parameters.

8.2.3 Conclusion to the theoretical side of identification of the errors-in-variables model

For the logistic regression errors-in-variables model there is only one error variance to be estimated. Kuchenhoff has shown that the model is identifiable if the classic measurement error model holds. We have taken this work a step further by looking at the case where the measurement error model follows a Berkson model and shown that the model is also identifiable in this case.

8.3 Analytical methods of implementing the Kuchenhoff theory

8.3.1 Estimate Q(z) by the logistic model

To implement these results we need to provide an estimate of $Q(z)$ in a different way from calculating the likelihood. An analytical form is needed so that the limits can be evaluated. The obvious type of function is the logistic.

$$Q(z) = \frac{\exp(\alpha_s + \beta_s z)}{1+\exp(\alpha_s + \beta_s z)} \tag{8.7}$$

Then

$$Q'(z) = \frac{\beta_s \exp(\alpha_s + \beta_s z)}{(1+\exp(\alpha_s + \beta_s z))^2}$$

and hence

$$\frac{Q'(z)}{Q(z)} = \frac{\beta_s}{1+\exp(\alpha_s + \beta_s z)}$$

If $\beta_s \geq 0$ then the limits are

$$\lim_{z\to\infty} \frac{Q'(z)}{Q(z)} = 0$$

$$\lim_{z\to-\infty} \frac{Q'(z)}{Q(z)} = \beta_s$$

If $\beta_s < 0$ then the limits are

$$\lim_{z\to\infty} \frac{Q'(z)}{Q(z)} = \beta_s$$

$$\lim_{z\to-\infty} \frac{Q'(z)}{Q(z)} = 0$$

Therefore, $c_2 = \beta_s$.

From (8.2)

$$c_1 = -Q^{-1}(0.5)c_2$$

On re-arrangement

$$Q\left(-\frac{c_1}{c_2}\right) = \frac{1}{2}$$

Now $Q^{-1}(0.5) = -\frac{\alpha_s}{\beta_s}$. Hence, $c_1 = \alpha_s$.

Both of the above estimates can be made from the main study set. This leaves a value for c_3 to be determined.

The value of c_3 can be uniquely determined by the function Q and the function K defined by (8.3) and (8.4). This cannot be evaluated analytically but an approximation can be attempted. Substituting the values already obtained for c_1 and c_2 in (8.3),

$$Q\left(\frac{1-c_1}{c_2}\right)=G\left(\alpha_s+\beta_s\left(\frac{1-c_1}{c_2}\right)\right)=G(1)=\frac{e}{1+e}$$

Hence we require c_3 such that

$$K\left(\sqrt{c_3}\right)=G(1)$$

A Taylor expansion gives

$$K(t)\approx\int\left[G(1)+tuG'(1)+\frac{(tu)^2}{2}G''(1)+\ldots\right]\varphi(u)du$$

This integral is an odd function therefore, the coefficients of the odd powers of t are all zero. Hence, on expansion

$$K(t)=G(1)+G^{II}(1)\frac{t^2}{2}\int u^2\varphi(u)du+G^{IV}(1)\frac{t^4}{4!}\int u^4\varphi(u)du+G^{VI}\frac{t^6}{6!}\int u^6\varphi(u)du+\ldots$$

$$\approx G(1)+G^{II}(1)\frac{t^2}{2}+G^{IV}(1)\frac{t^4}{8}+G^{VI}(1)\frac{t^6}{48} \qquad (8.8)$$

Since we require t such that $K(t)=1$, then

$$0\approx G^{II}(1)\frac{t^2}{2}+G^{IV}(1)\frac{t^4}{8}+G^{VI}(1)\frac{t^6}{48}+\ldots$$

Taking the second order term only gives $t=0$, which then gives $c_3=0$ and $\sigma_v=0$. This is the same naïve solution that would be obtained by ignoring the measurement error.

Taking the fourth order term

$$G^{II}(1)+G^{IV}(1)\frac{t^2}{4}=0$$

Now

$$G^{II}(1)=\frac{e(1-e)}{(1+e)^3}=-0.09086$$

$$G^{IV}(1)=\frac{e(1-11e+11e^2-e^3)}{(1+e)^5}=0.123507$$

This leads to

$$t^2=c_3=2.94267$$

and

$$\sigma_v^2=\frac{2.94267\sigma_z^2}{\beta_s^2\sigma_z^2+2.94267}$$

Hence,

$$\beta_t=\beta_s+\frac{2.94267}{\beta_s\sigma_z^2}$$

Therefore, using the logistic model to estimate the function $Q(z)$ produces a correction factor method. This method is dependent on σ_z and therefore, in turn is dependent on σ_v hence, the correction factor part will change as σ_v changes. However, the application of this correction factor method did not produce realistic estimates hence, a further approximation was implemented.

Taking the sixth order term

$$G^{II}(1)+G^{IV}(1)\frac{t^2}{4}+G^{VI}(1)\frac{t^4}{24}=0$$

Now

$$G^{VI}(x)=\frac{e^{-x}\left(-1+34e^{-x}-99e^{-2x}+44e^{-3x}-2e^{-4x}\right)}{\left(1+e^{-x}\right)^{6}}-\frac{3e^{-2x}\left(-8+79e^{-x}-119e^{-2x}+33e^{-3x}-e^{-4x}\right)}{\left(1+e^{-x}\right)^{7}}$$

$$=\frac{e^{-x}\left(1-e^{-x}\right)\left(1-56e^{-x}+246e^{-2x}-56e^{-3x}+e^{-4x}\right)}{\left(1+e^{-x}\right)^{7}}$$

Hence,

$$G^{VI}(1)=-0.2834339$$

When this term is substituted into (8.8), there is no solution to t^2 and hence there is no solution to c_3.

Therefore, we have found no way of implementing Kuchenhoff's theoretical results if the logistic model is used to estimate $Q(z)$. This is in line with Kuchenhoff's results; he stated that he also found no practical way of implementing his theory, though no details were given of the methods that he considered.

8.3.2 The use of other approximations to $P(Y=1|Z)$ in the theory by Kuchenhoff

Aim: To see if the probit model can be used in the method by Kuchenhoff

It has already been shown that the method fails when the logistic model is used to estimate $Q(z)$. A possible alternative is to use the probit model.

Let

$$Q(z)=\Phi(\alpha_s+\beta_s z)$$

where $\Phi(.)$ is the normal cumulative distribution function.

$$Q(z)=\int_{-\infty}^{\alpha_s+\beta_s z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)du$$

and the derivative is

$$Q'(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\alpha_s + \beta_s z)^2}{2} \right) \beta_s$$

Hence

$$\frac{Q'(z)}{Q(z)} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\alpha_s + \beta_s z)^2}{2} \right) \beta_s}{\int_{-\infty}^{\alpha_s + \beta_s z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2} \right) du} \tag{8.9}$$

Therefore, taking the limits

$$\frac{Q'(z)}{Q(z)} = \frac{\phi(z)}{\Phi(z)}$$

By L'Hopital's Rule

$$\lim_{z \to -\infty} \frac{\phi(z)}{\Phi(z)} = \frac{\phi'(-\infty)}{\Phi'(-\infty)} = \frac{\phi'(-\infty)}{\phi(-\infty)}$$

Hence,

$$\frac{\phi(z)}{\Phi(z)} = \frac{-z \exp\left(-\frac{1}{2} z^2 \right)}{\exp\left(-\frac{1}{2} z^2 \right)} \to \infty \qquad \text{as } z \to -\infty$$

Therefore,

$$\lim_{z \to \infty} \frac{Q'(z)}{Q(z)} \to 0 \text{ and } \lim_{z \to -\infty} \frac{Q'(z)}{Q(z)} \to \infty$$

Therefore, if the probit model is used to approximate $Q(z)$ then $c_2 = 0$ or $c_2 = \infty$. As a result the probit model is also not useful in trying to estimate the measurement error distribution.

Aim: To see if there is a way of combining the methods of Kuchenhoff and Reeves

From Chapter 3, it was shown that Reeves had devised a method using the following approximation to $Q(z)$

$$Q(z) \approx G\left(\frac{\alpha_t + \beta_t \mu_x (1 - \gamma_{x.z}) + \beta_t \gamma_{x.z} z}{\left(1 + k^2 \beta_t^2 \sigma_{x.z}^2\right)^{\frac{1}{2}}} \right)$$

where $G(.)$ as usual denotes the logistic function. If this function is re-defined so that

$$\alpha_s = \frac{\alpha_t + \beta_t \mu_x (1 - \gamma_{t.s})}{\left(1 + k^2 \beta_t^2 \sigma_{x.z}^2\right)^{\frac{1}{2}}}$$

and

$$\beta_s = \frac{\beta_t \gamma_{x.z}}{\left(1 + k^2 \beta_t^2 \sigma_{x.z}^2\right)^{\frac{1}{2}}}$$

then $Q(z) \approx G(\alpha_s + \beta_s Z)$. Hence the result for c_2 is the same as in 4.3.1. Such that,

$$c_2 = \beta_s \qquad \text{and} \qquad c_1 = \alpha_s$$

Therefore the same theory applies as described in 4.3.1. That is, the Reeves method does not produce analytical estimates of the model parameters when the measurement error variance is unknown in the theory by Kuchenhoff.

8.3.3 Conclusion

The limited number of models, which have the right type of behaviour at the limits, used to approximate $Q(z)$ in this section, have proven that we have found no practical way of implementing the method by Kuchenhoff. However, given that the Kuchenhoff theorem depends on the behaviour as $x \to \infty$, from a single dataset little information could be provided with regard to the explanatory variable at the limits and therefore, this result is not surprising. As the probit model differs from the logistic model at the limits, the probit model is also considered for identifiability.

8.4 Identification of the probit regression model parameters

In Chapter 3 we explained the probit regression measurement error model. In order to estimate the model parameters, the log-likelihood

$$LogL = \sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - p_i)$$

is minimized where

$$p_i = P(Y = 1|z) = \Phi\left\{ \frac{\alpha_t + \beta_t \left(\frac{\sigma_x^2}{\sigma_z^2} z + \frac{\sigma_v^2}{\sigma_z^2} \mu_x \right)}{\left(1 + \beta_t^2 \frac{\sigma_v^2 \sigma_x^2}{\sigma_z^2} \right)^{\frac{1}{2}}} \right\}$$

Though the probit regression model is not so well used in practice, the minimization of the log-likelihood is simpler as it does not require numerical integration in order to determine $P(Y = 1|z)$.

In terms of identifying all the model parameters from a single data set, the probit model is claimed to be not identifiable (Carroll et al (1995)) though we have found no proof of this result.

As we discussed in chapter 2, the logistic and probit models are very similar through the middle range of probability values, **Figure 8.1**. The difference between the two models can be seen at the extreme values.

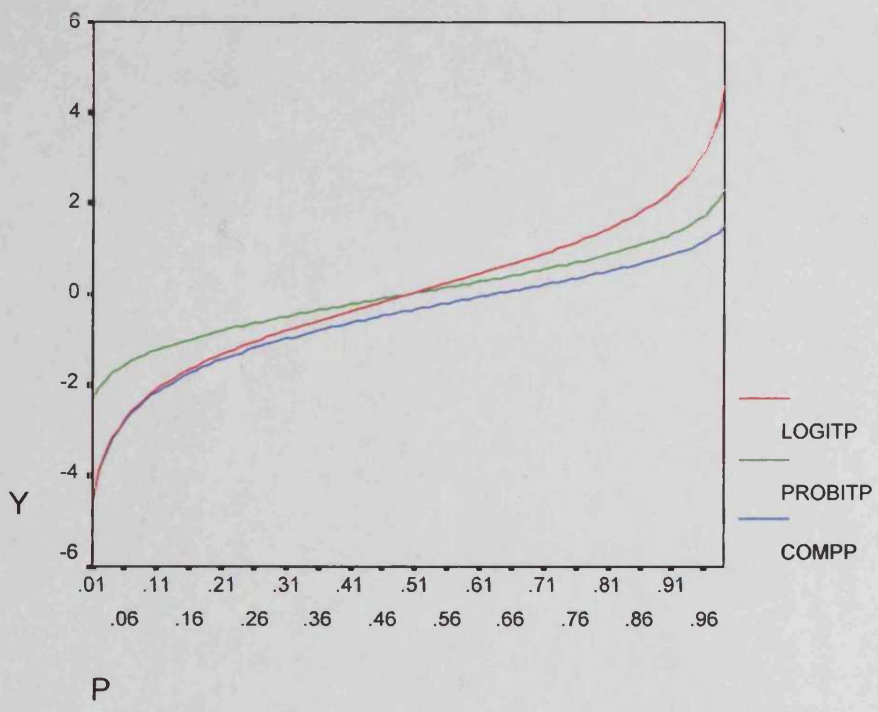


Figure 8.1: Graph to show logistic, probit and complementary log-log mean values across the range [0,1]

As the probit model is not identifiable and the Kuchenhoff theorem depends on the behaviour as $x \to \infty$, the identifiability of the logistic model must depend on a quite subtle behaviour at the limits. When considering a practical dataset, it is unlikely that there will be a huge amount of information at the limits especially as X is assumed to be Normally distributed. Therefore it is unlikely that there would be sufficient information to estimate all the model parameters using this theory.

8.5 Maximisation of the log likelihood to estimate α_t, β_t and σ_v^2

8.5.1 Introduction

It has been proven that for the logistic regression measurement error problem, all three of the model parameters can be estimated from a single data set. This chapter has also shown that there are no obvious ways of implementing this theory in practice. This section studies the associated log-likelihood and estimates of the model parameters when the log-likelihood is maximized to try to understand why the theory exists but no practical implementation is available.

8.5.2 Estimating the three model parameters

The theory by Kuchenhoff proved that all three model parameters could be estimated from a single study dataset. In this section, the log likelihood function is maximised in order to investigate the log likelihood function when all three parameters are to be estimated from a single dataset. The data was generated as before according to the following model:

13) Sample size N, was chosen to be 500 and 1000 respectively.

14) The true X values were assumed to follow a Normal Distribution with parameters

$$X \sim N(30,10^2)$$

15) The observed Z values were generated according to a classical measurement error model, $Z = X + V$, where V was also assumed to be Normally distributed $V \sim (0,\sigma_v^2)$

16) The disease status Y was generated conditional on X for the Case A where $\alpha_t = -3$ and $\beta_t = 0.1$, and the prevalence of the disease is 0.5 and $Y = 1$ with probability

$$P(Y = 1|X) = \frac{\exp(\alpha_t + \beta_t X)}{1 + \exp(\alpha_t + \beta_t X)}$$

17) The measurement error standard deviation σ_v ranged from 0 to 4, with steps of 1.

18) For each dataset, by each value of σ_v, the starting values for the maximisation of the log likelihood were the ordinary logistic regression model parameters, α_s and β_s, and σ_v was given the values 0, 2 and 4.

N=500

The results for N=500 are tabulated in Table 8.1 to Table 8.5. The likelihood value is displayed as –log(likelihood) so for these results a minimum is required.

When there is no measurement error in the explanatory variable, and the starting value for σ_v in the maximisation technique is 0, the estimates of α_t and β_t are very similar to the ordinary logistic regression method estimates. With respect to the estimate of σ_v, the maximisation technique estimates are all approximately 0. For all the datasets examined, this behaviour was consistent in all cases when the starting value for σ_v was 0.

When the starting value for σ_v is set to 2 and 4 respectively, in four of the datasets the maximisation technique found a maximum at $\hat{\sigma}_v > 8$, that is it found the scenario that $\hat{\sigma}_v \approx \hat{\sigma}_x$. These four datasets are colour coded as red in Table 8.1. This means that a maximum of the log likelihood has been found that corresponds to the case where all the variation in the observed Z values is measurement error, and there is little variation in the true X value. When the scenario of $\hat{\sigma}_v \approx \hat{\sigma}_x$ was found by the maximisation technique, the resulting estimate for β_t was a large positive value and correspondingly, the value of $\hat{\alpha}_t$ was large and negative. Comparing the log(likelihood) values the values have reduced by a small amount of the order of 0.0215 to 0.1, suggesting a very flat likelihood.

The results for the case where the true value in the measurement error model of σ_v was set to 1 are displayed in Table 8.2. As previously observed, when the starting value of σ_v

is set to 0, the maximisation technique estimates of α_t and β_t are the same as those estimated by the ordinary logistic regression method with an estimate of σ_v which is approximately 0. When the starting values for σ_v are set to 2 and 4 respectively, again the same four datasets have estimate of σ_v>8. So with the increase in the measurement error standard deviation, the same behaviour of a maximum being found at $\hat{\sigma}_v = 0$ when the starting value is zero, is observed. The same pattern of results are found for all values of the measurement error standard deviation.

These results suggest that the likelihood is very flat. In some cases there may be a local maxima at $\hat{\sigma}_v = 0$, in others, the maximum of the likelihood can be found when $\hat{\sigma}_v \approx \hat{\sigma}_x$. Which scenario is found by the maximisation technique seems to depend on the starting values of the model parameters suggesting that due to the nature of the likelihood the maximisation technique is unable to consistently find the maximum.

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.6074	0.0897	305.1648	-2.6074	0.0897	0.0001	305.1433	-59.720	2.0029	8.9993	305.1433	-52.589	1.7639	8.9871
-2.6744	0.0914	304.5228	-2.6744	0.0914	0.0000	304.5228	-2.6744	0.0914	0.0000	304.5228	-2.6744	0.0914	0.0000
-2.7176	0.0934	306.9432	-2.7176	0.0934	0.0000	306.9432	-2.7176	0.0934	0.0000	306.9432	-2.7176	0.0934	0.0000
-2.7123	0.0915	305.6784	-2.7123	0.0915	0.0001	305.5419	-122.87	4.1352	8.7002	305.5419	-108.95	3.6668	8.6973
-2.9720	0.0937	304.3945	-2.9720	0.0937	0.0000	304.3945	-2.9720	0.0937	0.0000	304.3945	-2.9720	0.0937	0.0000
-2.5083	0.0825	308.0989	-2.5083	0.0825	0.0000	308.0989	-2.5083	0.0825	0.0000	308.0989	-2.5083	0.0825	0.0000
-3.0041	0.1023	299.7220	-3.0041	0.1023	0.0001	299.6140	-109.48	3.6481	8.2765	299.6140	-97.450	3.2470	8.2727
-3.2039	0.1075	292.4982	-3.2039	0.1075	0.0001	292.4218	-81.406	2.7375	8.5564	292.4218	-64.599	2.1723	8.5419
-2.5981	0.0857	308.8058	-2.5981	0.0857	0.0000	308.8058	-2.5981	0.0857	0.0000	308.8058	-2.5981	0.0857	0.0000
-2.9318	0.0949	309.2396	-2.9318	0.0949	0.0000	309.2396	-2.9318	0.0949	0.0000	309.2396	-2.9318	0.0949	0.0000

Table 8.1: $\sigma_v = 0$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.5587	0.0879	305.7592	-2.5587	0.0879	0.0001	305.7490	-45.537	1.5235	9.0848	305.7490	-42.255	1.4138	9.0738
-2.6974	0.0921	303.4991	-2.6974	0.0921	0.0000	303.4991	-2.6974	0.0921	0.0000	303.4991	-2.6974	0.0921	0.0000
-2.6642	0.0917	308.0062	-2.6642	0.0917	0.0000	308.0062	-2.6642	0.0917	0.0000	308.0062	-2.6642	0.0917	0.0000
-2.7023	0.0911	306.1461	-2.7023	0.0911	0.0001	305.9916	-126.37	4.2498	8.7012	305.9917	-108.72	3.6559	8.6976
-2.9357	0.0924	304.8728	-2.9357	0.0924	0.0000	304.8728	-2.9357	0.0924	0.0000	304.8728	-2.9357	0.0924	0.0000
-2.4958	0.0823	308.1576	-2.4958	0.0823	0.0000	308.1576	-2.4958	0.0823	0.0000	308.1576	-2.4958	0.0823	0.0000
-2.8840	0.0981	302.4955	-2.8840	0.0981	0.0001	302.4386	-81.862	2.7244	8.3627	302.4387	-61.621	2.0510	8.3437
-3.1519	0.1059	293.5330	-3.1519	0.1059	0.0001	293.4454	-85.296	2.8733	8.6123	293.4454	-77.322	2.6047	8.6074
-2.4464	0.0804	312.1882	-2.4464	0.0804	0.0000	312.1882	-2.4464	0.0804	0.0000	312.1882	-2.4464	0.0804	0.0000
-2.8457	0.0920	310.9304	-2.8457	0.0920	0.0000	310.9304	-2.8457	0.0920	0.0000	310.9304	-2.8457	0.0920	0.0000

Table 8.2: $\sigma_v = 1$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.4579	0.0843	307.0878	-2.4579	0.0843	0.0000	307.0878	-2.4579	0.0843	0.0000	307.0878	-2.4579	0.0843	0.0003
-2.6576	0.0907	303.3700	-2.6576	0.0907	0.0000	303.3700	-2.6576	0.0907	0.0000	303.3700	-2.6576	0.0907	0.0000
-2.5516	0.0880	309.8626	-2.5516	0.0880	0.0000	309.8626	-2.5516	0.0880	0.0000	309.8626	-2.5516	0.0880	0.0000
-2.6372	0.0889	307.3798	-2.6372	0.0889	0.0001	307.2204	-109.98	3.6961	8.8017	307.2205	-100.89	3.3906	8.7992
-2.8377	0.0891	306.1797	-2.8377	0.0891	0.0000	306.1797	-2.8377	0.0891	0.0000	306.1797	-2.8377	0.0891	0.0000
-2.4296	0.0803	308.9843	-2.4296	0.0803	0.0000	308.9843	-2.4296	0.0803	0.0000	308.9843	-2.4296	0.0803	0.0000
-2.7021	0.0920	306.0516	-2.7021	0.0920	0.0001	306.0403	-6.9504	0.2329	6.9975	306.0403	-6.9499	0.2329	6.9973
-3.0231	0.1018	295.7039	-3.0231	0.1018	0.0001	295.6111	-106.82	3.6053	8.7991	295.6111	-74.751	2.5228	8.7845
-2.2554	0.0738	315.9445	-2.2554	0.0738	0.0000	315.9445	-2.2554	0.0738	0.0000	315.9445	-2.2554	0.0738	0.0000
-2.6927	0.0868	313.3703	-2.6927	0.0868	0.0000	313.3703	-2.6927	0.0868	0.0000	313.3703	-2.6927	0.0868	0.0000

Table 8.3: $\sigma_v = 2$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.3186	0.0794	308.9874	-2.3186	0.0794	0.0000	308.9874	-2.3186	0.0794	0.0000	308.9874	-2.3186	0.0794	0.0000
-2.5637	0.0876	304.0588	-2.5637	0.0876	0.0001	304.0063	-84.088	2.8022	9.2295	304.0063	-80.899	2.6959	9.2278
-2.3910	0.0827	312.3428	-2.3910	0.0827	0.0000	312.3428	-2.3910	0.0827	0.0000	312.3428	-2.3910	0.0827	0.0000
-2.5230	0.0850	309.2775	-2.5230	0.0850	0.0001	309.1251	-112.90	3.7913	9.0069	309.1251	-103.53	3.4767	9.0046
-2.6917	0.0841	308.1530	-2.6917	0.0841	0.0000	308.1530	-2.6917	0.0841	0.0000	308.1530	-2.6917	0.0841	0.0002
-2.3175	0.0768	310.4788	-2.3175	0.0768	0.0000	310.4788	-2.3175	0.0768	0.0000	310.4788	-2.3175	0.0768	0.0000
-2.4763	0.0843	310.1044	-2.4763	0.0843	0.0000	310.1043	-2.6294	0.0894	2.2155	310.1043	-2.6285	0.0894	2.2093
-2.8333	0.0956	298.7853	-2.8333	0.0956	0.0001	298.6927	-84.797	2.8674	9.0882	298.6927	-77.187	2.6101	9.0838
-2.0414	0.0665	319.8252	-2.0414	0.0665	0.0000	319.8252	-2.0414	0.0665	0.0000	319.8252	-2.0414	0.0665	0.0000
-2.4907	0.0801	316.3246	-2.4907	0.0801	0.0000	316.3246	-2.4907	0.0801	0.0000	316.3246	-2.4907	0.0801	0.0000

Table 8.4: $\sigma_v = 3$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.1556	0.0738	311.2683	-2.1556	0.0738	0.0000	311.2683	-2.1556	0.0738	0.0000	311.2683	-2.1556	0.0738	0.0000
-2.4293	0.0831	305.4110	-2.4293	0.0831	0.0001	305.3348	-39.259	1.3080	9.5000	305.3348	-39.259	1.3080	9.5000
-2.1981	0.0762	315.2270	-2.1981	0.0762	0.0000	315.2270	-2.1981	0.0762	0.0000	315.2270	-2.1981	0.0762	0.0000
-2.3705	0.0798	311.6777	-2.3705	0.0798	0.0001	311.5394	-128.34	4.3067	9.3089	311.5395	-109.23	3.6655	9.3053
-2.5140	0.0781	310.5849	-2.5140	0.0781	0.0000	310.5849	-2.5140	0.0781	0.0000	310.5849	-2.5140	0.0781	0.0000
-2.1723	0.0721	312.4774	-2.1723	0.0721	0.0000	312.4774	-2.1723	0.0721	0.0000	312.4774	-2.1723	0.0721	0.0000
-2.2268	0.0760	314.3535	-2.2268	0.0760	0.0000	314.3535	-2.2268	0.0760	0.0000	314.3535	-2.2268	0.0760	0.0000
-2.6041	0.0880	302.4722	-2.6041	0.0880	0.0001	302.3829	-72.816	2.4671	9.4858	302.3829	-72.562	2.4585	9.4856
-1.8200	0.0590	323.6109	-1.8200	0.0590	0.0000	323.6109	-1.8200	0.0590	0.0000	323.6109	-1.8200	0.0590	0.0000
-2.2611	0.0725	319.5289	-2.2611	0.0725	0.0000	319.5289	-2.2611	0.0725	0.0000	319.5289	-2.2611	0.0725	0.0000

Table 8.5: $\sigma_v = 4$

N=1000

The previous exercise was repeated for a sample size of 1000, to see if the results that were found were affected by the sample size. Again the likelihood value is given as –log(likelihood) and therefore a minimum value is required. Table 8.6 to Table 8.10 show the results for the different cases where the true value of σ_v in the measurement error model is set to 0 to 4 respectively.

The same patterns and results were identified as for the previous sample size. Again, the datasets where the maximisation technique found the scenario of $\hat{\sigma}_v \approx \hat{\sigma}_x$ have been high-lighted in red. These datasets show that the likelihood is very flat as the maximisation technique finds different estimates of $(\alpha_t, \beta_t, \sigma_v)$ with very similar values for –log(likelihood).

For dataset 8, high-lighted in blue throughout all the tables, it can be seen that a different scenario has been found with respect to the estimates of the model parameters. For this dataset, when the true value of σ_v has been set to zero in the simulation model, a maximum has been found when $\hat{\sigma}_v = 3.0650$. As the true value of σ_v is increased in the simulation model, so different scenarios are found. For the case where the true value of σ_v is set to 4 in the simulation model then the scenario of $\hat{\sigma}_v \approx \hat{\sigma}_x$ is found. From examination of the –log(likelihood) values, the values are very similar showing a flat likelihood.

This particular dataset again shows that the likelihood is very flat. As a result, there could be a number of local maxima which are not dependent on the true values of the model parameters. If that were the case then the identification of the local maxima would be dependent on the starting values for the model parameters rather than the data itself.

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.6737	0.0927	610.4740	-2.6737	0.0927	0.0001	610.2900	-92.352	3.0826	8.6362	610.2900	-104.87	3.5003	8.6406
-2.7388	0.0942	608.2727	-2.7388	0.0942	0.0000	608.2727	-2.7388	0.0942	0.0000	608.2727	-2.7388	0.0942	0.0000
-3.1419	0.1081	585.6652	-3.1419	0.1081	0.0000	585.6652	-3.1419	0.1081	0.0000	585.6652	-3.1419	0.1081	0.0000
-3.0313	0.1004	599.8206	-3.0313	0.1004	0.0001	599.7157	-65.315	2.1777	8.4048	599.7158	-59.698	1.9903	8.3972
-2.9897	0.0983	603.5049	-2.9897	0.0983	0.0000	603.5049	-2.9897	0.0983	0.0000	603.5049	-2.9897	0.0983	0.0000
-2.8743	0.0959	601.6861	-2.8743	0.0959	0.0000	601.6861	-2.8743	0.0959	0.0000	601.6861	-2.8743	0.0959	0.0000
-2.9384	0.0991	603.8778	-2.9384	0.0991	0.0001	603.8037	-56.370	1.8782	8.3581	603.8038	-49.516	1.6499	8.3431
-3.2777	0.1101	579.5750	-3.2777	0.1101	0.0001	579.5742	-3.6884	0.1239	3.0650	579.5742	-3.6884	0.1239	3.0650
-2.7231	0.0902	612.0035	-2.7231	0.0902	0.0000	612.0035	-2.7231	0.0902	0.0000	612.0035	-2.7231	0.0902	0.0000
-3.1750	0.1070	599.7937	-3.1750	0.1070	0.0000	599.7937	-3.1750	0.1070	0.0000	599.7937	-3.1750	0.1070	0.0000

Table 8.6: $\sigma_v = 0$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.6337	0.0913	611.0856	-2.6337	0.0913	0.0001	610.9066	-96.571	3.2185	8.7390	610.9067	-88.256	2.9414	8.7356
-2.7268	0.0938	607.8593	-2.7268	0.0938	0.0000	607.8593	-2.7268	0.0938	0.0000	607.8593	-2.7268	0.0938	0.0000
-3.1213	0.1074	585.9427	-3.1213	0.1074	0.0000	585.9427	-3.1213	0.1074	0.0000	585.9427	-3.1213	0.1074	0.0000
-3.0468	0.1008	598.7965	-3.0468	0.1008	0.0001	598.6561	-103.28	3.4417	8.4391	598.6562	-70.394	2.3456	8.4208
-2.9520	0.0969	604.1631	-2.9520	0.0969	0.0000	604.1631	-2.9520	0.0969	0.0000	604.1631	-2.9520	0.0969	0.0000
-2.8506	0.0952	602.3003	-2.8506	0.0952	0.0000	602.3003	-2.8506	0.0952	0.0000	602.3003	-2.8506	0.0952	0.0000
-2.8545	0.0962	607.6572	-2.8545	0.0962	0.0000	607.6175	-47.638	1.5858	8.4167	607.6176	-41.013	1.3654	8.3932
-3.2072	0.1079	582.1385	-3.2072	0.1079	0.0001	582.1326	-4.9350	0.1659	5.4669	582.1326	-4.9358	0.1659	5.4678
-2.6007	0.0860	616.7377	-2.6007	0.0860	0.0000	616.7377	-2.6007	0.0860	0.0000	616.7377	-2.6007	0.0860	0.0001
-3.1337	0.1057	600.8992	-3.1337	0.1057	0.0000	600.8992	-3.1337	0.1057	0.0000	600.8992	-3.1337	0.1057	0.0000

Table 8.7: $\sigma_v = 1$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.5334	0.0878	613.3416	-2.5334	0.0878	0.0001	613.1858	-82.226	2.7363	8.9466	613.1858	-89.161	2.9670	8.9501
-2.6493	0.0912	609.2773	-2.6493	0.0912	0.0000	609.2773	-2.6493	0.0912	0.0000	609.2773	-2.6493	0.0912	0.0000
-3.0299	0.1045	588.3475	-3.0299	0.1045	0.0000	588.3475	-3.0299	0.1045	0.0000	588.3475	-3.0299	0.1045	0.0000
-3.0004	0.0992	599.6074	-3.0004	0.0992	0.0001	599.4907	-67.539	2.2491	8.5347	599.4907	-76.151	2.5359	8.5424
-2.8515	0.0934	606.5866	-2.8515	0.0934	0.0000	606.5866	-2.8515	0.0934	0.0000	606.5866	-2.8515	0.0934	0.0000
-2.7578	0.0923	604.8673	-2.7578	0.0923	0.0000	604.8673	-2.7578	0.0923	0.0000	604.8673	-2.7578	0.0923	0.0000
-2.7087	0.0912	613.0515	-2.7087	0.0912	0.0001	613.0471	-6.1608	0.2059	6.7819	613.0471	-6.2039	0.2073	6.8002
-3.0602	0.1030	587.0356	-3.0602	0.1030	0.0001	587.0168	-14.060	0.4731	8.2314	587.0168	-14.017	0.4717	8.2284
-2.4301	0.0803	622.6831	-2.4301	0.0803	0.0000	622.6831	-2.4301	0.0803	0.0000	622.6831	-2.4301	0.0803	0.0000
-3.0152	0.1018	604.0570	-3.0152	0.1018	0.0000	604.0570	-3.0152	0.1018	0.0000	604.0570	-3.0152	0.1018	0.0000

Table 8.8: $\sigma_v = 2$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.3872	0.0828	616.9100	-2.3872	0.0828	0.0000	616.7867	-78.851	2.6199	9.2634	616.7867	-82.888	2.7540	9.2657
-2.5190	0.0868	612.2507	-2.5190	0.0868	0.0000	612.2507	-2.5190	0.0868	0.0000	612.2507	-2.5190	0.0868	0.0000
-2.8787	0.0994	592.5765	-2.8787	0.0994	0.0000	592.5765	-2.8787	0.0994	0.0000	592.5765	-2.8787	0.0994	0.0000
-2.8983	0.0957	602.0900	-2.8983	0.0957	0.0001	602.0418	-44.851	1.4925	8.7107	602.0418	-49.211	1.6377	8.7235
-2.7026	0.0884	610.4391	-2.7026	0.0884	0.0000	610.4391	-2.7026	0.0884	0.0000	610.4391	-2.7026	0.0884	0.0000
-2.6084	0.0875	609.0628	-2.6084	0.0875	0.0000	609.0628	-2.6084	0.0875	0.0000	609.0628	-2.6084	0.0875	0.0000
-2.5161	0.0848	619.6059	-2.5161	0.0848	0.0000	619.6059	-2.5161	0.0848	0.0000	619.6059	-2.5161	0.0848	0.0000
-2.8561	0.0963	593.7256	-2.8561	0.0963	0.0001	593.6945	-41.572	1.4007	9.0812	593.6945	-40.761	1.3733	9.0779
-2.2291	0.0735	629.3424	-2.2291	0.0735	0.0000	629.3424	-2.2291	0.0735	0.0000	629.3424	-2.2291	0.0735	0.0000
-2.8357	0.0958	608.8674	-2.8357	0.0958	0.0000	608.8674	-2.8357	0.0958	0.0000	608.8674	-2.8357	0.0958	0.0000

Table 8.9: $\sigma_v = 3$

Ordinary Logistic Regression		Starting Value $\sigma_v = 0$				Starting Value $\sigma_v = 2$				Starting Value $\sigma_v = 4$			
$\hat{\alpha}_t$	$\hat{\beta}_t$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$	-log (likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
-2.2120	0.0769	621.3765	-2.2120	0.0769	0.0000	621.2882	-25.461	0.8461	9.5000	621.2882	-25.461	0.8461	9.5000
-2.3525	0.0813	616.3734	-2.3525	0.0813	0.0001	616.3660	-8.3892	0.2829	8.3550	616.3660	-8.3880	0.2829	8.3548
-2.6849	0.0930	598.1644	-2.6849	0.0930	0.0000	598.1644	-2.6849	0.0930	0.0000	598.1644	-2.6849	0.0930	0.0000
-2.7516	0.0908	605.9314	-2.7516	0.0908	0.0000	605.9314	-2.7516	0.0908	0.0001	605.9314	-2.7516	0.0908	0.0000
-2.5223	0.0823	615.2927	-2.5223	0.0823	0.0000	615.2927	-2.5223	0.0823	0.0000	615.2927	-2.5223	0.0823	0.0000
-2.4209	0.0813	614.4100	-2.4209	0.0813	0.0000	614.4100	-2.4209	0.0813	0.0000	614.4100	-2.4209	0.0813	0.0000
-2.2948	0.0773	626.7967	-2.2948	0.0773	0.0000	626.7967	-2.2948	0.0773	0.0000	626.7967	-2.2948	0.0773	0.0000
-2.6180	0.0884	601.5485	-2.6180	0.0884	0.0001	601.5179	-39.025	1.3166	9.5000	601.5179	-39.025	1.3166	9.5000
-2.0149	0.0663	636.2352	-2.0149	0.0663	0.0000	636.2352	-2.0149	0.0663	0.0000	636.2352	-2.0149	0.0663	0.0000
-2.6169	0.0885	614.7913	-2.6169	0.0885	0.0000	614.7913	-2.6169	0.0885	0.0000	614.7913	-2.6169	0.0885	0.0000

Table 8.10: $\sigma_v = 4$

8.5.3 Conclusion

From this simulation study it appears that the solution found by the maximization technique is very dependent on the starting values for the model parameters. The main results are two scenarios, the first is that $\hat{\sigma}_v = 0$, and the second is that $\hat{\sigma}_v \approx \hat{\sigma}_x$. The difference in the value of the likelihood for these two different scenarios is marginal suggesting that the likelihood is very flat. Further to this, if the starting value for σ_v is zero, then the solution is an estimate for σ_v close to zero. Therefore, there is either a local maximum near to zero or else the optimization routine is unable to find a maximum as the likelihood is too flat. The following section investigates into these conclusions further.

8.5.4 Investigation into the logistic regression measurement error likelihood

The results from the previous section identified a number of questions. These included whether the results are from the inaccuracy in the numerical integration or whether local maxima were being identified by the maximisation technique instead of looking through the whole search space. Finally, is there a local maxima at $\sigma_v = 0$ despite the true value of σ_v. The following sections consider these points.

8.5.4.1 Accuracy in the programming

To conduct the previous investigation, a program was written in Fortran with the aid of NAG routines (www.nag.co.uk). Two NAG routines were used in particular, one for the numerical integration and the other for the maximization of the log-likelihood. The associated accuracy and stopping criteria of these routines are covered in the following sections.

Numerical Integration

The NAG routine, D01AMF, calculates an approximation to the integral of a function $f(x)$ over an infinite interval $[a,b]$:

$$I = \int_a^b f(x)dx$$

in this case

$$I = \int G\left(\alpha_t + \beta_t\left(\frac{\sigma_x\sigma_v}{\sigma_z}u + \frac{\sigma_x^2}{\sigma_z^2}z + \frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)\right)\varphi(u)du$$

The infinite integral is transformed to $[0,1]$ and an adaptive procedure is then employed on the transformed integral, see NAG website for further details. The accuracy of the routine is defined by

$$|I - result| \leq tol$$

where

$$tol = \max\{|EPSABS|, |EPSREL| \times |I|\}$$

and

I is the integral to solved

Result is the solution to the integral

EPSABS is the absolute accuracy, in this case 0.0

EPSREL is the relative accuracy, in this case 1.0×10^{-10}

To numerically solve the integral

$$I = \int G\left(\alpha_t + \beta_t\left(\frac{\sigma_x\sigma_v}{\sigma_z}u + \frac{\sigma_x^2}{\sigma_z^2}z + \frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)\right)\varphi(u)du$$

the stopping criteria determined by the accuracy of *ESPABS* and *EPSREL* were deemed to be sufficient in order to provide a good approximation to the integral.

The accuracy associated with this routine seems to suggest that the variable results were not due to the numerical integration procedure.

Maximisation Technique

The NAG routine, E04JYF, uses a quasi-Newton algorithm for finding a maximum of a function $F(x_1,...,x_n)$, subject to fixed upper and lower bounds of the independent variables $x_1, x_2,..., x_n$ without the use of the second derivatives of $F(x)$. For the logistic regression measurement error problem, the function to be minimized is

$$LogL = \sum_{i=1}^{n} y_i \log\{P(Y=1|Z)\} + \sum_{i=1}^{n}(1-y_i)\log\{1-P(Y=1|Z)\}$$

where

$$P(Y=1|z) = \int G\left(\alpha_t + \beta_t\left(\frac{\sigma_x\sigma_v}{\sigma_z}u + \frac{\sigma_x^2}{\sigma_z^2}z + \frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)\right)\varphi(u)du.$$

Using starting points supplied by the user, the routine estimates the gradient and the curvature of $F(x)$ and a sequence of feasible points is then made which is intended to converge to a local maximum. To determine convergence, two sets of criteria are used. The first requires the following conditions to hold:

- The size of the projected movement for the current point is less than a multiple of the machine precision
- The difference in function values must be less than a function of the machine precision and the current function value.
- The multiple of the positive-definite approximation of the matrix of second derivatives and the search direction has to be less than a function of the machine precision and the current function value

Or that the following condition holds:

- The multiple of the positive-definite approximation of the matrix of second derivatives and the search direction is less than a multiple of the machine precision.

This routine was considered to have a high-level of accuracy and therefore was used for this analysis. However, the results suggest that the nature of the likelihood being very flat may mean that this technique is unable to find the maximum.

8.5.4.2 Investigation

To investigate the question of whether there is a local maximum at $\hat{\sigma}_v = 0$, the properties of a number of the datasets considered in the previous section will be investigated.

If we take the case where the true value of $\sigma_v = 2$ in the simulation model and the first dataset, a maximum was found for all starting values of σ_v at

- $\hat{\alpha}_t$=-2.4579
- $\hat{\beta}_t$=0.0843
- -log(likelihood)=307.0878

To see whether a true maximum had been found, a profile likelihood was calculated such that, the value of σ_v was increased in small steps from zero, and the log likelihood maximised for the parameters α_t and β_t. In Table 8.11 the value of –log(likelihood) is displayed therefore, a minimum value is required. As σ_v is increased, so the value of –log(likelihood) increases, showing that for this case, a

maximum can be found when $\sigma_v = 0$. Table 8.11 also shows how flat the likelihood is, that is a small change in the value of σ_v results in a change in the value of the log(likelihood) in the 7th decimal place.

σ_v	-log(likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$
0	307.0878050003	-2.4579479754	0.0842842697
0.01	307.0878050076	-2.4579505621	0.0842843659
0.02	307.0878050294	-2.4579582798	0.0842846141
0.03	307.0878050658	-2.4579714068	0.0842850515
0.04	307.0878051166	-2.4579894954	0.0842856538
0.05	307.0878051821	-2.4580132433	0.0842864414
0.06	307.0878052620	-2.4580421355	0.0842874031
0.07	307.0878053566	-2.4580756700	0.0842885205
0.08	307.0878054656	-2.4581151594	0.0842898333
0.09	307.0878055892	-2.4581596174	0.0842913119
0.1	307.0878057273	-2.4582090395	0.0842929561

Table 8.11: Dataset 1 profile likelihood

In order to understand how much information is available within a single dataset to estimate all three of the model parameters, the information matrix is investigated. For this dataset the inverse of the information matrix was calculated to be

$$\begin{pmatrix} 0.111335 & -0.00342 & 0 \\ -0.00342 & 0.000115 & 0 \\ 0 & 0 & 406.5041 \end{pmatrix}$$

Therefore, for this dataset, the standard errors associated with the estimates of α_t and β_t are small whereas the standard error associated with the estimate of σ_v is very large.

We now consider a dataset where two different scenarios namely, $\hat{\sigma}_v = 0$ and $\hat{\sigma}_v \approx \hat{\sigma}_x$ were found. For dataset 4 in table 8.3, when the starting value of σ_v in the maximisation technique was set to 0, a maximum was found at

- $\hat{\alpha}_t$=-2.6372
- $\hat{\beta}_t$=0.0889
- -log(likelihood)=307.3798
- $\hat{\sigma}_v$=0.0001

And when the starting value of σ_v was set to 4, a maximum was also found at

- $\hat{\alpha}_t$=-100.89
- $\hat{\beta}_t$=3.3906
- -log(likelihood)=307.2205
- $\hat{\sigma}_v$=8.7992

These results suggest that there is a local maximum at $\sigma_v = 0$. If we again consider the profile likelihood for σ_v increasing in small steps away from zero, Table 8.12 shows the results of the optimisation of the log(likelihood) for the model parameters α_t and β_t. As previously shown, the value of –log(likelihood) is displayed, therefore a minimum value is required. For this dataset, there does not appear to be a maximum at $\sigma_v = 0$. As the value of σ_v increases, so the value of –log(likelihood) reduces. These results suggest that though a maximum was found by the maximisation technique at $\hat{\sigma}_v = 0$, the true maximum had not been found and the results are therefore dependent on the starting value of σ_v.

σ_v	-log(likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$
0	307.3797534804	-2.6372072001	0.0888860190
0.01	307.3797533843	-2.6372103234	0.0888861305
0.02	307.3797530962	-2.6372196300	0.0888864389
0.03	307.3797526158	-2.6372351016	0.0888869581
0.04	307.3797519434	-2.6372568150	0.0888876878
0.05	307.3797510787	-2.6372847027	0.0888886248
0.06	307.3797500220	-2.6373187236	0.0888897643
0.07	307.3797487730	-2.6373590145	0.0888911217
0.08	307.3797473319	-2.6374054776	0.0888926811
0.09	307.3797456986	-2.6374581018	0.0888944485
0.1	307.3797438731	-2.6375171791	0.0888964341

Table 8.12: Dataset 4 profile likelihood results

If we examine the associated inverse information matrix, when the starting value of σ_v was zero, the estimates of the model parameters produced the following inverse information matrix:

$$\begin{pmatrix} 0.120797 & -0.00375 & -0.00296 \\ -0.00375 & 0.000126 & 0.000023 \\ -0.01205 & 0.000329 & 422.4228 \end{pmatrix}$$

As previously observed, when the scenario is that $\hat{\sigma}_v = 0$, the standard errors associated with the estimates of α_t and β_t are small. If we consider the inverse information matrix for the scenario when the starting value of σ_v was 4, that is a scenario found at $\hat{\sigma}_v \approx \hat{\sigma}_x$ then

$$\begin{pmatrix} 1798.012 & -60.5593 & 0.043916 \\ -60.5593 & 2.043914 & -0.00757 \\ 0.043916 & -0.0139 & 0.0000708 \end{pmatrix}$$

For this scenario, the standard error associated with the estimate of σ_v is very small whereas the standard errors for the estimates of α_t and β_t have increased in size. This suggests that either there is information to estimate α_t and β_t or there is information to estimate σ_v. However, there is not enough information in the data to estimate all three of the model parameters.

8.5.5 Conclusion

In conclusion, these two datasets have shown that a maximum can be found at either $\hat{\sigma}_v = 0$ or $\hat{\sigma}_v \approx \hat{\sigma}_x$. For the dataset where the scenario of $\hat{\sigma}_v \approx \hat{\sigma}_x$ was found, the estimates of the model parameters were dependent on their starting values in the maximisation technique. The dataset showed that a maximum could not be found at $\sigma_v = 0$, even though this is the result found by the maximisation technique when the starting value for σ_v was zero. Overall, the likelihood is very flat and therefore the maximisation technique is unable to consistently find the maxima.

The calculation of the inverse of the information matrix has also shown that either the estimates of α_t and β_t have small standard errors corresponding to the scenario where $\hat{\sigma}_v = 0$, or the estimate of σ_v has a small standard error corresponding to the scenario where the estimate of σ_v tends towards the estimate of σ_x. These results suggest that there seems to be inadequate information for estimating all the model parameters from a single study data set when the explanatory variable is assumed to be Normally distributed.

8.6 Bayesian approach to estimating the model parameters from a single study dataset

8.6.1 Introduction

In chapter 7 we showed how a Bayesian approach could be used to estimate the parameters of the logistic regression measurement error problem. In that case, the method of MCMC was used to estimate the posterior distributions of the parameters α_t and β_t by using the data and prior distributions for α_t and β_t. The measurement error variance was taken to be known precisely. This approach can be extended by assuming that the measurement error variance is unknown and so is regarded as a distribution that can be estimated in the same way as α_t and β_t. Therefore, a Bayesian approach is considered in the context of estimating all the model parameters from a single data set. It is assumed that if a single data set contains enough information to estimate all three of the model parameters then a Bayesian analysis should be able to provide sensible estimates of the model parameters without using informative prior distributions.

8.6.2 Simulation Study

To test this theory a single data set was generated according to the same model that was used in chapter 7 that is,

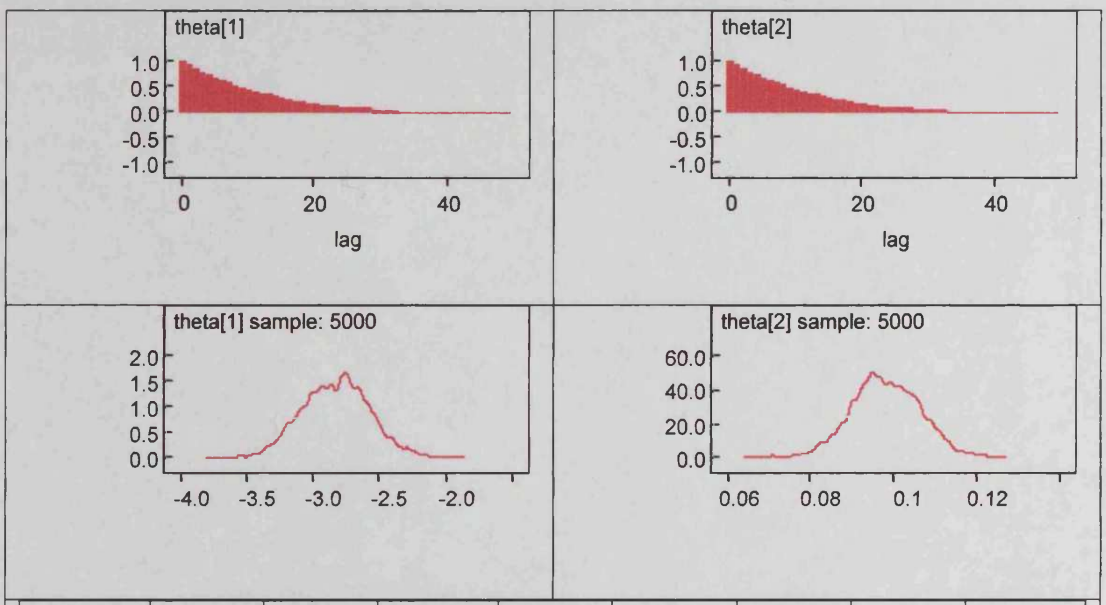
2. The sample size N, was chosen to be 1000.
6. The true X values were assumed to follow a Normal Distribution with parameters $X \sim N(30,10^2)$
7. The disease status Y was generated conditional on X for the case where $\alpha_t = -3$ and $\beta_t = 0.1$ and Y=1 with probability

$$P(Y=1|X)=\frac{\exp(\alpha_t+\beta_t X)}{1+\exp(\alpha_t+\beta_t X)}$$

8. The measurement error model is described by the Berkson measurement error model that is, X=Z+V where Z and V are independent and $V \sim N(0,\sigma_v^2)$
9. The measurement error standard deviation σ_v was taken to be 2, and given different prior distributions throughout this analysis.
10. The model parameters α_t and β_t were given vague prior distributions in the following form $\alpha_t \sim N(0,100)$ and $\beta_t \sim N(0,100)$. As before this means that no prior knowledge about these parameters is assumed.

Case 1: $\sigma_v = 2$ *assumed known*

Table 8.13 shows the BUGS summaries for the case where $\sigma_v = 2$ assumed known. In this case the chain appeared to converge quickly and the resulting estimates of the model parameters were sensible.



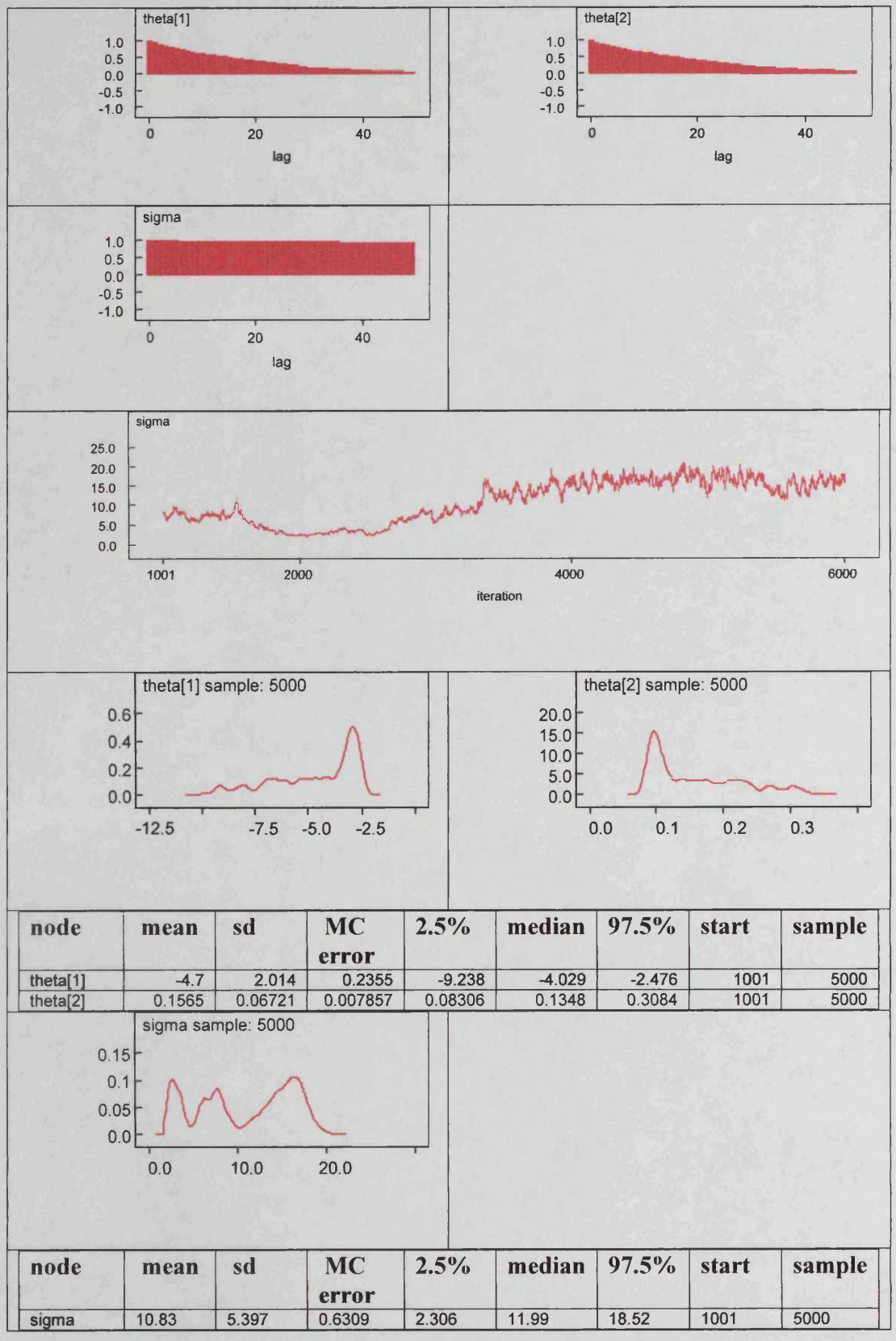
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-2.832	0.2566	0.01588	-3.326	-2.825	-2.322	1001	5000
theta[2]	0.09804	0.008302	5.14E-04	0.08197	0.09779	0.114	1001	5000

Table 8.13: Results for $\sigma_v = 2$ assumed known

Case 2: σ_v given an uninformative prior distribution

In this case σ_v was given an uninformative prior distribution. Specifically, the precision, that is the reciprocal of σ_v^2, was given a gamma distribution with a large variance. In such circumstances a gamma prior distribution is commonly chosen as it is guaranteed to be non-negative.

Table 8.14 displays the results for estimating the three model parameters. In this case the interest is in how well the chain simulates the values for σ_v and the resulting posterior distribution. This can be seen in the time series of the simulated values, the auto-correlation function and the inferences that can be made from the posterior distribution. The auto-correlation function shows very poor mixing for σ_v whereas the auto-correlation functions for α_t and β_t shows the same good mixing that was observed in chapter 7. For the time series for σ_v, the simulated values have a trend of long runs up and down displaying a lack of convergence in the chain. Finally, the posterior distribution for σ_v is uninformative about the true value of σ_v. From these results it does appear that there is insufficient information in the data to estimate the measurement error standard deviation σ_v.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
theta[1]	-4.7	2.014	0.2355	-9.238	-4.029	-2.476	1001	5000
theta[2]	0.1565	0.06721	0.007857	0.08306	0.1348	0.3084	1001	5000

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
sigma	10.83	5.397	0.6309	2.306	11.99	18.52	1001	5000

Table 8.14: Case 2 results

Case 3: σ_v given an informative prior distribution

In this case, the precision $1/\sigma_v^2$ was again given a gamma distribution but this time it was centred on the correct values 0.25 with a standard deviation of 1.

Table 8.15 shows the time series plot for the 2001-4000 simulated values showing substantial auto-correlation and the posterior distribution of the precision showing considerable uncertainty.

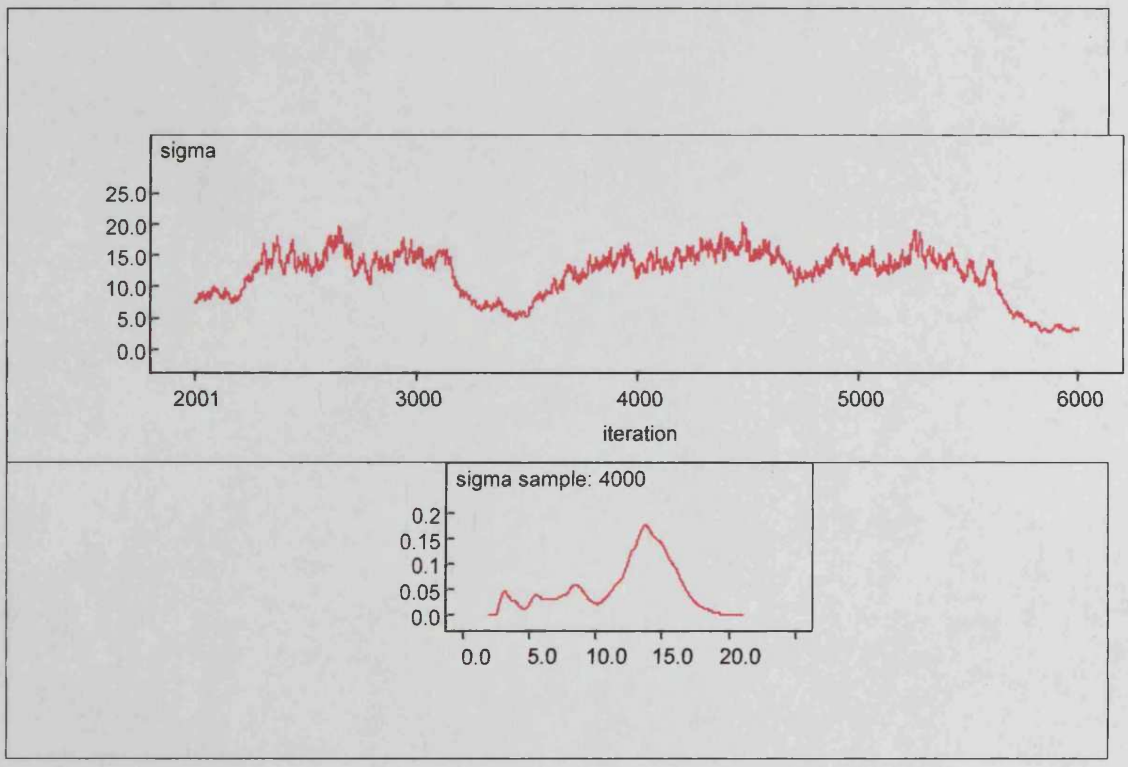
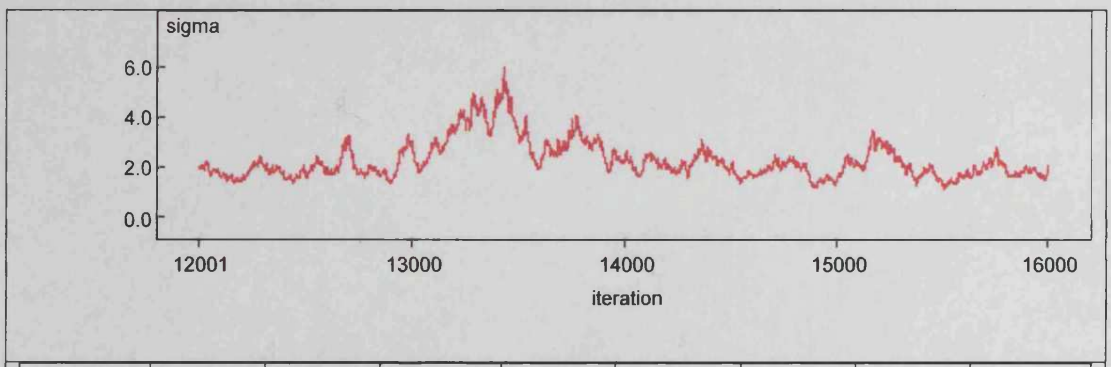


Table 8.15: Results for the 2001-4000 simulated values of the precision of σ_v^2

The above results show that to obtain sensible estimates for the precision, the chain must be run for a longer period. Table 8.16 displays these results. The time series plot shows that there is still a trend of long runs up and down, showing that the auto-correlation is still very large. On examination of the plot of the posterior distribution,

the distribution has become smoother and the estimates of the precision are much closer to their true values.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
sigma	2.37	0.7688	0.08604	1.388	2.159	4.327	11001	5000

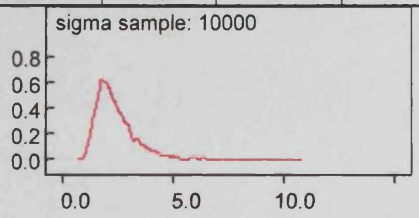
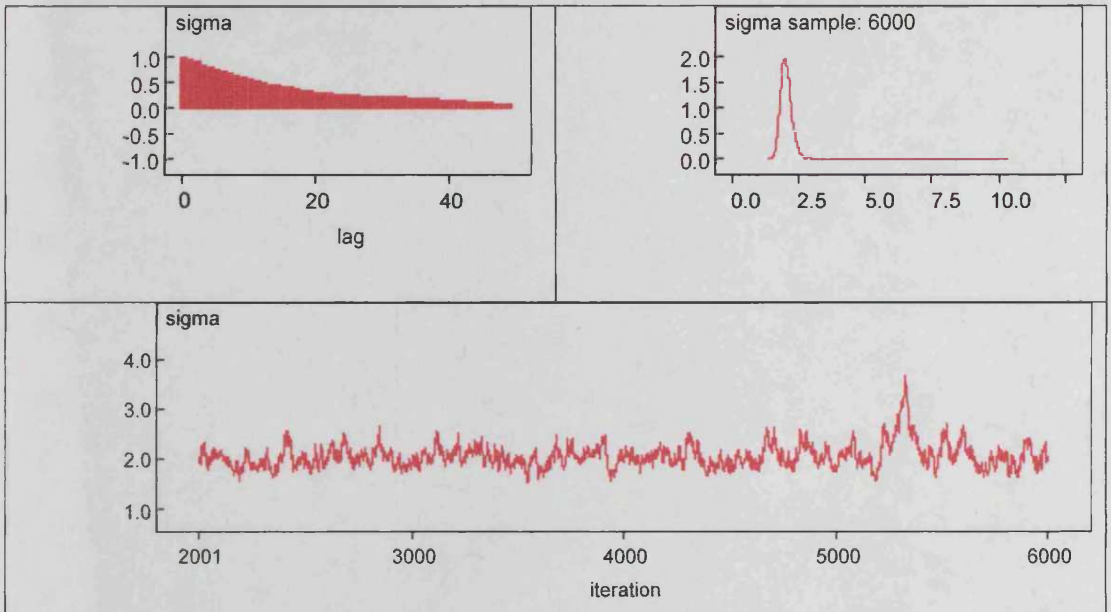


Table 8.16:Results for the precision of σ_v^2

Case 4: σ_v given a highly informative prior distribution

In this case a highly informative prior distribution was chosen for the precision $1/\sigma_v^2$. A gamma distribution was again chosen, centred on the true value of 0.25 but with a standard deviation of 0.1.

Table 8.17 displays the results for this case. As can be seen from the usual plots, by giving the precision of σ_v^2 a highly informative prior distribution the chain is mixing well and the results are comparable to the first case where σ_v^2 was assumed to be known precisely.



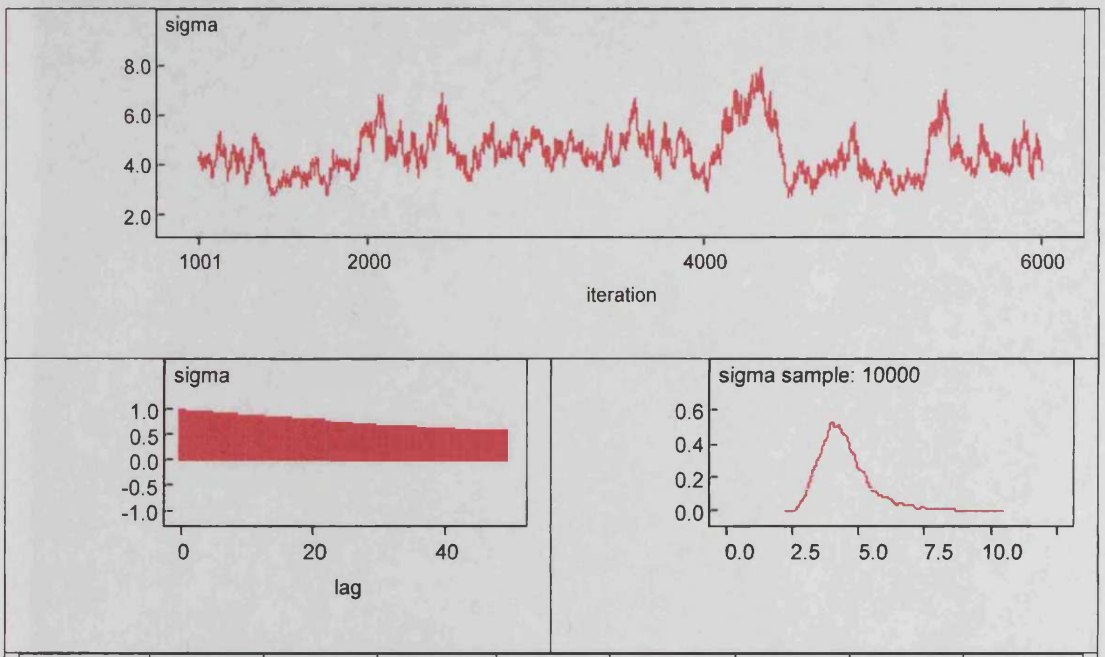
node	mean	sd	MC error	2.5%	median	97.5%	start	sample
sigma	2.057	0.4244	0.03392	1.665	2.004	2.639	1	6000

Table 8.17: Results for case 4

Case 5: σ_v given a highly informative but incorrect prior distribution

The previous case showed that if the Bayesian model is given a highly informative prior distribution for the precision $1/\sigma_v^2$ then the posterior distribution produced sensible estimates. In this case, another highly informative prior distribution is given the precision $1/\sigma_v^2$, but this time it is centred on 0.0667 with a standard deviation of 0.1. This prior distribution is equivalent to saying that $\sigma_v = 4$. This will enable us to investigate whether the Bayesian model can still estimate the true value given a single data set and an informative prior distribution centred in the wrong place.

Table 8.18 displays the results for this case. Both the time series plot and the auto-correlation function show that the chain is not mixing well. In the case of the posterior distribution, the distribution is still centred around the prior distribution. It appears that the data has not had an effect on the posterior distribution for σ_v and that the posterior distribution is largely determined by the prior distribution. This result suggests that there is little information within the dataset to estimate σ_v.



node	mean	sd	MC error	2.5%	median	97.5%	start	sample
sigma	4.433	0.8976	0.09387	3.021	4.32	6.611	1001	5000

Table 8.18: Results for case 5

Case 6: Frequentist Approach to the dataset

The dataset used to illustrate the Bayesian approach to the problem was also considered using the frequentist approach of maximizing the associated log(likelihood). As previously conducted, the log(likelihood) was maximized with respect to the three model parameters namely α_t, β_t and σ_v. For the model parameters α_t and β_t, their respective starting values were taken to be the ordinary logistic regression method estimates. For the parameter σ_v, three starting values were considered namely 0, 2 and 4. Table 8.19 displays the results of this approach.

For this dataset, the maximization technique has found the one scenario regardless of the starting value of σ_v, that is $\hat{\sigma}_v = 0$. For the same case in the Bayesian approach, that is using an uninformative prior distribution for the precision of σ_v^2 the resulting mean estimate of σ_v is the scenario where $\hat{\sigma}_v \approx \hat{\sigma}_x$.

σ_v	- log(likelihood)	$\hat{\alpha}_t$	$\hat{\beta}_t$	$\hat{\sigma}_v$
0	603.2892938885	-2.8043077707	0.0970739246	0.0000000000
2	603.2892938885	-2.8043077716	0.0970739240	0.0000000000
4	603.2892938885	-2.8043074251	0.0970739164	0.0001344377

Table 8.19: Dataset results

These differing results suggest that there is not enough information within the single dataset in order to estimate all three of the model parameters.

8.6.3 Conclusion to the Bayesian approach to estimating all three model parameters from a single data set

For the Bayesian approach to the problem, a single data set was used to illustrate the results. These results show that when the measurement error standard deviation is known and can be inputted into the model, then the Bayesian method will produce sensible estimates of the model parameters α_t and β_t. Chapter 7 showed that this was true for non-informative prior distributions for α_t and β_t and a large measurement error standard deviation. When the measurement error standard deviation was also regarded as a parameter to be estimated and was given a prior distribution, the effect of informative and non-informative prior distributions and the data were examined. Overall, the data had little effect on the posterior distribution of σ_v, which was determined largely by its prior distribution, something which was not observed for the other model parameters. This suggests that there is little information in the data to estimate the measurement error standard deviation.

8.7 Conclusion

For the linear regression measurement error problem, all of the model parameters were not identifiable from a single data set, unless at least one of the measurement error standard deviations was known. Kuchenhoff proved that for the logistic regression measurement error problem, all the model parameters were identifiable from a single study dataset.

The Kuchenhoff theorem assumed that there was information for the explanatory variable X, as $x \to \infty$. In practice, such a situation is highly unlikely. In this chapter, no practical way of implementing this theorem was found.

From studying the measurement error problem likelihood, the likelihood was found to be very flat and when trying to estimate all the model parameters, the results were very erratic and polarized around $\hat{\sigma}_v = 0$ and $\hat{\sigma}_v \approx \sigma_x$. This meant that the scenarios found were either estimating that there was no measurement error or that all the variation in the explanatory variable was measurement error. The maximization technique chosen for this problem seemed to have a sufficient accuracy level for the stopping criteria. However, the nature of the flat likelihood meant that the maximum was sometimes not found and the maximum that was found was very much dependent on the starting value of the measurement error standard deviation. These results suggested that there is not enough information in the dataset to estimate all the model parameters.

When considering a Bayesian approach to the problem, this method showed that the prior distribution for the measurement error standard deviation had much more of an

effect than the data, which again suggested that there is not enough information within a single study dataset to estimate all the model parameters.

In conclusion, Kuchenhoff proved that the logistic regression measurement error model is identifiable, though these results suggest that in practice the model is not identifiable. The practical results also fit with the theory that the probit model is not identifiable and since the two models are very similar, these results are not surprising. As the theorem requires that there is information about the explanatory variable at the limits, this work could be expanded further. However, as this did not seem to be a practical problem in a medical context, this work was not conducted. The issue of different distributions for X, other than Normal, could also be explored.

Chapter 9

9 An application to an epidemiological study

9.1 Introduction

As we have seen, several methods are available for adjusting the ordinary logistic regression estimates to allow for measurement error. In practice they have not been used a great deal in practice, especially in epidemiology. This may be because the importance of the problem has not been widely appreciated or else because they are not available in standard software packages. There is an exception to this, however, and a pragmatic method, not considered in the earlier chapters, has been used by MacMahon (1990) and Clarke (1999) in an important epidemiological context. In this chapter the methods described earlier in the thesis will be applied to that problem and compared to MacMahon's method.

9.2 Problem

In Chapters 5 and 6 a number of methods for correcting for measurement error were compared through extensive simulations. The contexts used for the comparison were not based on any practical context but were designed to consider a variety of situations, particularly different prevalences. In this chapter the methods which have been discussed earlier will be applied to an important practical problem, that of examining the role of risk factors such as blood pressure or cholesterol in the occurrence of vascular disease, in particular coronary heart disease and strokes. This is important for two reasons. Firstly it is a common but very serious condition and so it is important in analysing studies on risk factors to use the best possible methods of assessing links with the disease. Secondly, as will be seen, the extent of measurement error in commonly-used risk factors is very substantial and so the potential level of corrections is high; the use of good methods for doing this is especially important.

In Chapter 1 it was noted that blood pressure, in common with other clinical variables, was subject to two distinct sources of error. Apart from the usual measurement error, which might occur if blood pressure is measured with a sphygmomanometer, an individual's blood pressure is subject to short and medium term variation about its long term 'normal' level. A study which bases estimates of risk on a single reading of blood pressure may therefore lead to biased estimates of the association with heart disease or strokes unless account is taken of this error. In Chapter 5 an example based on the Framingham Heart Study was shown; there the effect of measurement error on values of systolic blood pressure was discussed. In this Chapter that particular context is explored further.

In this chapter we will first look at a pragmatic method which has been used in practice in such studies as MacMahon (1990), Clarke(1999) and Prospective Studies Collaboration (2002). No justification for the method was produced when its use was advocated and an analysis of the method will be given here. This shows that it is, in fact, similar to a method discussed in Chapter 4. That method will be compared to other methods for their effectiveness in this particular context, comparing both bias and standard error. This will be based on simulations and also on an example concerning heart disease and blood pressure.

An example will also be given of a multivariate situation, where more than one risk factor is subject to measurement error. Finally, implications for planning studies will be discussed.

9.3 MacMahon's method

MacMahon et al recognised that measurement error would affect the estimates of the regression coefficients and hence of the odds ratio associated with a risk factor such as diastolic blood pressure (DBP). The method they proposed for dealing with this 'regression dilution' effect was the following. It relied on having a second reading of DBP made some time later. They were actually conducting a systematic review and found that in some studies, such as the well-known Framingham study, values of DBP had been measured at on at least one occasion after the baseline reading on a proportion of the study participants. We first describe their method and then show that it is equivalent to one described earlier.

Suppose the main study contains n subjects, each of whom have a value of the risk factor x measured at baseline. Suppose that a subset has further values measured at later time points. For simplicity suppose initially there is a single further measurement on m subjects at a later time point. Let z refer to the first, or baseline, value of one of these subjects and w refer to the second.

Suppose the sample of m baseline values is divided into a number of equal groups, say quintiles. The means of the baseline values for the subjects in each quintile are calculated. Using the same division of subjects into quintiles the means of the follow-up values in each quintile are also calculated.

Let r_1 denote the difference between the means of the top and bottom quintiles at baseline and let r_2 denote the corresponding quantity at follow-up. Then MacMahon et al argue that the difference will have been shrunk, so that $r_2 < r_1$. They suggest that the usual maximum likelihood estimates be derived and that the estimated coefficient $\hat{\beta}$ be inflated by a factor $\frac{r_1}{r_2}$. No justification was given for this empirical approach. We shall first investigate its properties and see that it is closely related to a method considered already. We will then compare its properties to those of other methods.

The first point is to note that the reason for the shrinkage of the difference in means between the quintiles is due to regression towards the mean. Given that individuals do not have a fixed level, which is subject to measurement error, an individual who is selected as being in the highest quintile at baseline may have been placed there because their normal level was high but it may have occurred because their value was unusually high above a relatively average normal value. If the latter was the case, then

we would expect the random fluctuation to be smaller in a repeated measurement and therefore would expect the mean of such values to be lower than at baseline. In a similar way we would expect the mean of the values in the lowest quintile to be higher at follow-up and hence the difference in means to be smaller.

We assume there are two observations Z and W on each member of the sample. We assume that Z and W have a bivariate Normal distribution and that their means are μ_z and μ_w and SDs σ_z and σ_w. Suppose the correlation between them is ρ.

We need to calculate the distribution of W conditional on a given value of Z.

The joint probability distribution function of Z and W is $f(Z = z \ \& \ W = w) =$

$$\frac{1}{2\pi\sigma_z\sigma_w\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{z-\mu_z}{\sigma_z}\right)^2+\left(\frac{w-\mu_w}{\sigma_w}\right)^2-2\rho\left(\frac{z-\mu_z}{\sigma_z}\right)\left(\frac{w-\mu_w}{\sigma_w}\right)\right]\right)$$

The marginal of Z is

$$f_Z(z)=\frac{1}{\sigma_z\sqrt{2\pi}}\exp\left(-0.5\left[\frac{z-\mu_z}{\sigma_z}\right]^2\right)$$

So the conditional probability distribution function of W given $Z = z$ is

$$f(w\,|\,z)=\frac{f(z,w)}{f_Z(z)}$$

and we find the standard result, namely that, given $Z = z$, W is Normally distributed with mean

$$\mu_w+\frac{\rho\sigma_w}{\sigma_z}(z-\mu_z)$$

and variance

$$\sigma_w^2(1-\rho^2)$$

Now consider the mean of those selected by being above a threshold C. The pdf of the values is

$$\frac{f_Z(z)}{P(Z \geq C)}$$

and has mean

$$\frac{1}{P(Z \geq C)\sigma_z\sqrt{2\pi}} \int_C^\infty z \exp\left(-0.5\left[\frac{z-\mu_z}{\sigma_z}\right]^2\right) dz$$

Let $u = (z - \mu_z)/\sigma_z$. Then this expression becomes

$$\text{Mean of selected } Z = \frac{1}{P(Z \geq C)\sqrt{2\pi}} \int_{C^*}^\infty (\sigma_z u + \mu_z) \exp\left(-0.5u^2\right) du z$$

where $C^* = (C - \mu_z)/\sigma_z$

So Mean of selected Z =

$$\frac{1}{P(Z \geq C)\sqrt{2\pi}} \left\{ \sigma_z \left[-\exp(-0.5u^2) \right]_{C^*}^\infty + \mu_z \int_{C^*}^\infty \exp\left(-0.5u^2\right) du \right\}$$

$$= \frac{1}{P(Z \geq C)\sqrt{2\pi}} \left\{ \sigma_z \exp\left(-0.5C^{*2}\right) + \mu_z \sqrt{2\pi} P(Z \geq C) \right\}$$

$$= \mu_z + \sigma_z \frac{\phi(C^*)}{P(U \geq C^*)}$$

$$= \mu_z + K\sigma_z$$

where $K = \dfrac{\phi(C^*)}{P(U \geq C^*)}$ and U is a standard normal variable.

Now consider the regression to the mean effect. We want to calculate the expected value of W for those selected as satisfying $Z \geq C$.

Mathematically, we want to find

Expected value $(W \mid Z \geq C)$

We will assume that there has been no systematic change, so that $\mu_Z = \mu_W = \mu$, say.

From above

$$\text{Expected value } (W \mid Z = z) = \mu + \frac{\rho\sigma_w}{\sigma_z}(z - \mu)$$

We need to integrate this over all possible values of z ($\geq C$).

So the required value is

$$\mathrm{E}(W \mid Z \geq C) = \int_C^{\infty} \left(\mu + \frac{\rho\sigma_w}{\sigma_z}(z-\mu) \right) . \frac{1}{\sigma_z\sqrt{2\pi}} \exp\left(-0.5 \left[\left(\frac{z-\mu}{\sigma_z} \right)^2 \right] \right) dz \Big/ P(Z \geq C)$$

$$= \frac{\mu P(Z \geq C) + \frac{\rho\sigma_w}{\sqrt{2\pi}} \left[-\exp\left\{ -0.5 \left(\frac{z-\mu}{\sigma_z} \right)^2 \right\} \right]_C^{\infty}}{P(Z \geq C)}$$

$$= \mu + K\rho\sigma_W$$

Suppose that the SD of W is equal to that of Z; denote both by σ.

Suppose C is chosen to define the top quintile. Then K= 0.28/0.2 = 1.4. So the mean of the values of Z in the top quintile is $\mu + 1.4\sigma_z$ and the mean of those of W is $\mu + 1.4\sigma_w\rho$. By symmetry the mean of the values of Z in the bottom quintile is $\mu - 1.4\sigma_z$ and the mean of those of W is $\mu_w - 1.4\sigma_w \rho$. Hence the difference of means in the two quintiles is $2.8\sigma_z$ for Z and $2.8\sigma_w\rho$ for W. Thus the shrinkage factor is the correlation coefficient $\rho\sigma_w/\sigma_z$.

Consider now the situation we are concerned with in this chapter.

Let μ denote the population mean of, say, diastolic blood pressure. Suppose that subject i has a subject effect X_i , so that $\mu + X_i$ represents the normal level for that person. Let V denote a random error, which may be made up of measurement error or random fluctuations over time.

Dropping the suffix i for simplicity, denote the baseline value by

$$Z = \mu + X + V_1 \qquad (9.1)$$

and the follow-up value by

$$W = \mu + X + V_2 \qquad (9.2)$$

So we can think of X as being the true value and Z as the observed one.

Assuming the random errors are independent of the subject effect, and that the variance does not change so that $\sigma_w = \sigma_z$ then we have

$$\text{Var}(Z) = \text{Var}(W) = \sigma_x^2 + \sigma_v^2$$

and

$$\text{Cov}(Z,W) = \text{E}(X^2) = \sigma_x^2$$

in the obvious notation.

Hence the correlation between these values is $R = \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$

Therefore the MacMahon method is equivalent to simply scaling the ordinary regression coefficient by the reciprocal of this factor. This is, of course, the standard result for maximum likelihood estimates for linear regression and corresponds to Rosner's (simple) method for logistic regression.

So the empirical method, based on purely pragmatic approach, has a theoretical basis.

In their paper MacMahon et al suggested that there might sometimes be several follow-up values available, citing the example of the Framingham study. In that case, the variable W may be the mean of, say, k follow up values. Therefore a similar analysis would apply, except that the random error would now have a variance of σ_v^2 / k.

Note that we can write the shrinkage formula as $\rho \dfrac{\sigma_w}{\sigma_z} = \dfrac{Cov(W,Z)}{\sigma_z^2}$

Whatever the value of k then $\text{Cov}(W,Z) = \sigma_x^2$

Hence the shrinkage factor is always given by $R = \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$

MacMahon suggests that to compute the standard error of the modified estimator, that of the uncorrected one should be divided by the estimate of R. This does not, of course, take account of sampling error in the estimation of R.

9.4 Comparison of estimators

In this section we will compare four different estimators for the regression coefficient β, namely the ordinary logistic regression estimator, and those of MacMahon, Rosner and Reeves. Those of MacMahon and Rosner both involve the use of a correction factor based on estimates of the shrinkage coefficient R; they differ in their estimation of this. The Reeves method also involves the shrinkage factor where,

$$\hat{\beta}_t = \frac{\hat{\beta}_s}{R\left(1 - k^2 \hat{\beta}_s^2 \hat{\sigma}_v^2 / R^2\right)^{\frac{1}{2}}}$$

and so gives a greater correction than the simple correction factor method.

We can write the revised estimate as

$$\hat{\beta}_t = \hat{\beta}_s \left(\frac{\hat{\sigma}_x^2 + \hat{\sigma}_v^2}{\hat{\sigma}_x^2} \right)$$

$$= \hat{\beta}_s \left(1 + \frac{\hat{\sigma}_v^2}{\hat{\sigma}_x^2} \right)$$

To find the variance of this estimator we can use the result that if X and Y are independent then

$$\text{Var}(XY) = \sigma_x^2 \mu_y^2 + \sigma_y^2 \mu_x^2 + \sigma_x^2 \sigma_y^2$$

To use this result we assume that $\hat{\beta}_s$ and $1+\frac{\hat{\sigma}_v^2}{\hat{\sigma}_x^2}=\theta$ are approximately independent.

The estimate of $\hat{\sigma}_v^2$ is based only on the number of subjects m on whom there are 2 measurements, while that of $\hat{\sigma}_x^2$ is based on all n subjects; usually $n>>m$ and so

$Var\left(1+\frac{\hat{\sigma}_v^2}{\hat{\sigma}_x^2}\right)\approx\frac{1}{\hat{\sigma}_x^4}Var\left(\hat{\sigma}_v^2\right)=\frac{2(m-1)\sigma_v^4}{\sigma_x^4 m^2}$ if the error terms are normally distributed.

As there is no closed expression for the variance of the ordinary logistic regression estimator, we cannot obtain one for the variance of an adjusted estimator.

Hence

$$Var(\hat{\beta}_t)\approx\theta^2 Var(\hat{\beta}_s)+\hat{\beta}_s^2\frac{2(\theta-1)^2}{m}+\frac{2(\theta-1)^2}{m}Var(\hat{\beta}_s) \quad (9.3)$$

where $\theta = 1+\frac{\hat{\sigma}_v^2}{\hat{\sigma}_x^2}$

Because the variance of the ordinary estimator for β_t is based on a sample of size n, we would expect that asymptotically its variance will tend to zero and the variance will be dominated by the second term, dependent on the size of the validation study.

We cannot find a simple expression for that of other estimators, however; the variance of the estimator of R by MacMahon's method is based on only $0.4m$ observations and would be expected to be larger than that shown above, while the variance of the correction factor for the Reeves method is difficult to calculate or even approximate. Thus one of the primary aims of the simulation was to obtain expressions for the standard errors of the estimators.

Before embarking on the simulation, we consider a specific example designed to show the importance of correcting for measurement error and to motivate the particular context used in the simulation. This example concerns the risk of heart disease associated with diastolic blood pressure. The diastolic blood pressure of 650 adults was measured twice, 5 years apart, a study by Health Promotion Wales in 1985 and 1990. Applying MacMahon's method they were divided into approximate quintiles. The mean DBP in the lowest quintile was 64.71 and in the highest was 93.09, giving a difference of 28.38. The same subjects after 5 years had respective means of 70.28 and 88.42, giving a difference of 18.14. The shrinkage factor was therefore 0.64.

The standard deviation of the measured values of DBP at baseline was 10.56. The variance of the error terms was estimated from the variance of the changes in DBP, as that variance is twice the required value. The estimate was 50.56. Hence $\hat{\sigma}_z^2 = 111.59$, $\hat{\sigma}_v^2 = 50.56$ and so $\hat{\sigma}_x^2 = 111.59 - 50.56 = 60.93$. The estimate of the shrinkage factor is therefore $60.93/111.59 = 0.546$.

Two points are striking. The first is that, not surprisingly, the methods give different estimates of the shrinkage factor; the second is that, since the ordinary logistic regression estimate will be divided by this factor, the extent of the correction is very substantial, with the coefficient nearly doubling.

To investigate the differences between the methods more closely a simulation was carried out to compare the distributions of the four different estimators, specifically estimating the mean and standard deviation of the sampling distribution of each. It

was designed to resemble this case of heart disease and diastolic blood pressure (DBP). A data set of n values of 'true' DBP was generated and random error superimposed as in (9.1). In a subset of m subjects follow-up values were generated according to model (9.2).

The logistic regression model was given by

$$\text{logit}\left(\frac{p}{1-p}\right)=\alpha+\beta x$$

where p denotes the probability of heart disease in an individual with true DBP x.

In order to implement the methods, estimates of σ_x^2 and σ_v^2 are required. Two sources of data were used for these purposes. The first uses the data collected by Health Promotion Wales described above. The second set was on a set of 3114 men with angina, aged between 29 and 72, who took part in a randomised controlled trial (DART) investigating the effects of different diets; all were followed up after 6 months. This second data set is on very different individuals, all undergoing treatment, and the estimates for σ_v and σ_x were rather different, at 8.46 and 9.27 respectively. Remarkably the value of R was again 0.546, leading to a correction factor of 1.83 in both cases.

As they formed a more representative population the first data set formed the basis of the simulation, taking the observed values of DBP to have a mean of 78.7 and variance based on the estimates of 7.11 for σ_v and 7.81 for σ_X. The true values of the model parameters were $\alpha_t = -4$ and $\beta_t = 0.04$. These were estimated from the

meta-analyses reported by MacMahon (1990) and the Prospective Studies Collaboration (2002).

The size of the main data set was taken to be 1000 first and then 5000, and follow-up studies of sizes 100, 250 and 500 were considered. In each case 10,000 data sets were simulated. The results shown in tables 9.1 to 9.6 display the mean $\hat{\beta}$ values of the 10,000 simulations.

n=1000	Ordinary logistic regression		MacMahon		Rosner		Reeves	
Method	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$
m = 100	0.0217	0.0063	0.0436	0.9861	0.0404	0.0130	0.0408	0.0134
m = 250	0.0219	0.0065	0.0424	0.0183	0.0403	0.0123	0.0407	0.0126
m = 500	0.0217	0.0065	0.0407	0.0140	0.0399	0.0121	0.0403	0.0124

Table 9.1: Method results for n=1000

The simulations confirmed that (9.3) approximates well to the standard error of the revised estimator, with the correlation between the terms $\hat{\beta}_S$ and θ close to 0. However, it does not take into account the error in using $\hat{\sigma}_x^2$ in place of σ_x^2 and therefore was not found to be sufficiently accurate to be included in the simulation.

n=5000	Ordinary logistic regression		MacMahon		Rosner		Reeve	
Method	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$
m = 100	0.0217	0.0028	0.0519	0.3811	0.0402	0.0073	0.0405	0.075
m = 250	0.0217	0.0028	0.0423	0.0131	0.0400	0.0061	0.0403	0.0063
m = 500	0.0217	0.0029	0.0406	0.0087	0.0398	0.0056	0.0401	0.0058

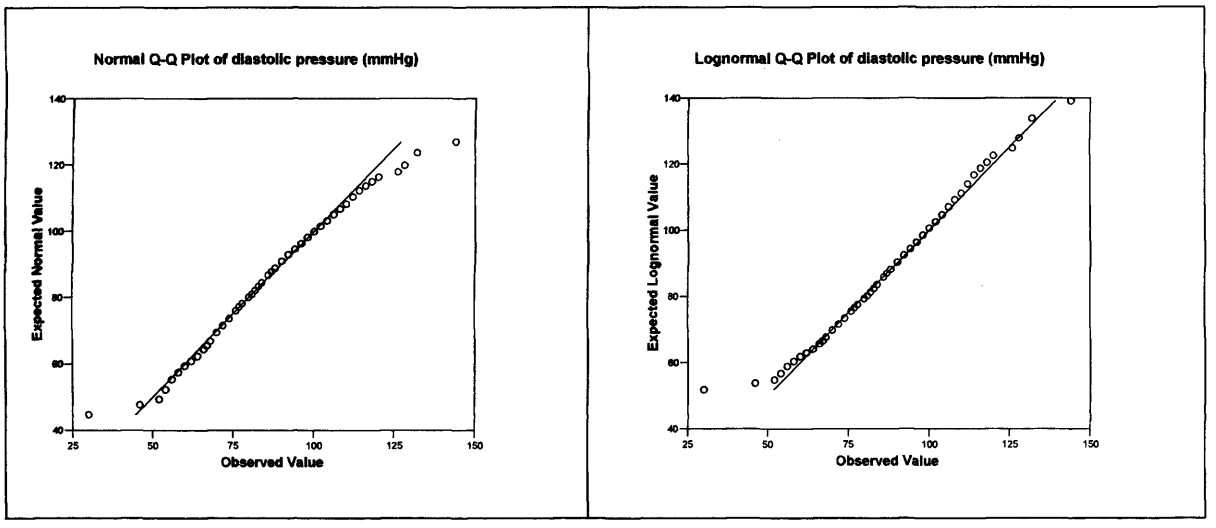
Table 9.2: Method results for n=5000

The results in Table 9.1 and Table 9.2 confirm the situation presented in Chapter 5, namely that the uncorrected logistic regression estimates are highly biased; the extent of the correction here is considerably more than in the examples considered there. They confirm that Rosner's and Reeves' method are very similar, with the Reeves method always giving estimates slightly further than zero. MacMahon's method is

biased for small follow-up studies with a much larger variance, and so cannot be recommended over the others, provided the model is known to be appropriate.

Another relevant aspect in any comparison is the robustness of the methods to departures from assumptions. MacMahon's method is based on the assumption of normality; how does it fare in comparison to others if data are not normally distributed?

Two situations were considered. The first is if the underlying distribution is lognormal with the same mean and variance. For example, from the Caerphilly Cohort study (McCarron, 2001), the following Q-Q plots suggest that a lognormal distribution may be more appropriate than a normal distribution.



n=1000	Ordinary logistic regression		MacMahon		Rosner		Reeve	
Method	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$
m = 100	0.0220	0.0064	0.0521	0.4385	0.0407	0.0130	0.0411	0.0134
m = 250	0.0220	0.0063	0.0448	0.0224	0.0404	0.0120	0.0408	0.0124
m = 500	0.0219	0.0064	0.0428	0.0147	0.0401	0.0119	0.0405	0.0123

Table 9.3: method results for n=1000

n=5000	Ordinary logistic regression		MacMahon		Rosner		Reeve	
Method	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$
m = 100	0.0219	0.0028	0.0260	2.3043	0.0406	0.0074	0.0409	0.0076
m = 250	0.0219	0.0028	0.0440	0.0423	0.0401	0.0060	0.0404	0.0062
m = 500	0.0219	0.0029	0.0428	0.0094	0.0400	0.0056	0.0404	0.0057

Table 9.4: Method results for n=5000

The other situation considered was that of a markedly skewed distribution. Here a chi square distribution with 5 degrees of freedom was taken, but scaled to have the same mean and variance, 85 and 7.8, as in Chapter 5.

n=1000	Ordinary logistic regression		MacMahon		Rosner		Reeve	
Method	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$	$\hat{\beta}_t$	$SE(\hat{\beta}_t)$
M = 100	0.0222	0.0064	0.0592	0.5604	0.0413	0.0133	0.0417	0.0138
m = 250	0.0221	0.0065	0.0507	0.8656	0.0409	0.0122	0.0414	0.0125
m = 500	0.0221	0.0064	0.0545	0.0210	0.0408	0.0120	0.0412	0.0123

Table 9.5: Method results for n=1000

N=5000	Ordinary logistic regression		MacMahon		Rosner		Reeve	
Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
m = 100	0.0222	0.0028	0.1450	6.0082	0.0413	0.0075	0.0416	0.0077
m = 250	0.0221	0.0029	0.0567	0.1475	0.0409	0.0061	0.0412	0.0062
m = 500	0.0222	0.0029	0.0547	0.0162	0.0408	0.0056	0.0412	0.0057

Table 9.6: method results for N=5000

The Rosner and Reeves methods have been affected only slightly by the different distributions of the X values but the MacMahon method gives much poorer results, particularly in the case of the heavily skewed distribution. This is not surprising as that will have a major impact on the extreme quintiles. It shows that the method has little to commend it over other approaches. However, if there is evidence that the logistic regression model is inappropriate, the fact that this is not an assumption made by the MacMahon method will give it a degree of flexibility over the other methods. In that case transformations to normality are important if the method is to produce reasonable results.

9.5 Bivariate example

Another factor linked to the occurrence of heart disease is blood cholesterol. Many studies in recent years have routinely measured the total cholesterol in the blood and have modelled the risk of heart disease as a function of this, although more recent evidence suggests that it is the level of low density lipoprotein, or possibly the ratio of low to high density lipoprotein, that is a better predictor of risk. Cholesterol levels within a subject are not constant and undergo variation in the same manner as

diastolic blood pressure, though the level of variation is probably less, relative to the between subject variation, than with DBP.

The correction factor method extends to this situation as discussed in Chapter 4. The correction factor is now defined by a matrix

$$\Gamma=\Sigma_x(\Sigma_x+\Sigma_v)^{-1}$$

where x now refers to the pair of underlying true values and v to the possibly correlated random departures from these true values.

$$\hat{\beta}_t = \hat{\beta}_s\hat{\Gamma}^{-1} = \hat{\beta}_s(I+\hat{\Sigma}_v\hat{\Sigma}_x^{-1})$$

where $\hat{\beta}_t$ now represents a row vector of the coefficients of the model parameters. If both the true values and the measurement errors are uncorrelated then the matrix $I+\hat{\Sigma}_v\hat{\Sigma}_x^{-1}$ will be diagonal and the estimates can be corrected individually, as with Rosner's method, for example. If this is not true, then errors in one variable will have an impact on the revised estimate for the effect of the other variable.

The data from Health Promotion Wales referred to above also contained follow up values of cholesterol and so the joint effects were considered. Both DBP and cholesterol measurements were available on 637 subjects .

Denote the DBP values by

$$Z_i = \mu_D + X + V_{1i} \qquad (9.4)$$

for i=1 and 2, and the cholesterol values by

$$W_i = \mu_C + U + V_{2i} \qquad (9.5)$$

where X and U are the 'usual' levels of DBP and cholesterol and μ_D and μ_C represent the underlying true mean values of diastolic blood pressure and cholesterol that are assumed not to change between the two recordings.

Clearly, under the usual independence assumptions,

$$\text{cov}(Z, W) = \text{cov}(X, U) + \text{cov}(V_1, V_2)$$

Further, $\text{cov}(V_{11} - V_{12}, V_{21} - V_{22}) = 2\text{cov}(V_{11}, V_{12})$ assuming the covariance matrix of the error terms can be estimated using the within subject differences since, for example, $V_{11} - V_{12} = Z_1 - Z_2$.

The observed correlation between DBP and cholesterol was 0.294. The above methods were used to estimate the covariance, and hence the correlation, of the error terms and hence of the true values. The resulting correlations were 0.426 for the true values and 0.127 for the errors.

If correcting factors were applied independently, then the coefficient for DBP would be multiplied by 1.83 and that for cholesterol by 1.34.

The estimated covariance matrix of the error terms was $\begin{pmatrix} 50.56 & 0.525 \\ 0.525 & 0.338 \end{pmatrix}$ and that of the true values was $\begin{pmatrix} 60.93 & 3.406 \\ 3.406 & 1.013 \end{pmatrix}$. The matrix by which the ordinary logistic regression estimates are multiplied is therefore $\begin{pmatrix} 1.986 & -2.800 \\ -0.012 & 1.375 \end{pmatrix}$.

This then gives as revised estimates

$$(\hat{\beta}_{t1}\ \ \hat{\beta}_{t2})=(\hat{\beta}_{s1}\ \ \hat{\beta}_{s2})\begin{pmatrix}1.986 & -2.800\\ -0.012 & 1.375\end{pmatrix}$$

$$= 1.986\hat{\beta}_{s1}-0.012\hat{\beta}_{s2}\,,\ -2.8\hat{\beta}_{s1}+1.375\ \hat{\beta}_{s2}$$

Plausible values for β_1 and β_2 are 0.04 and 0.3 based on evidence from the literature. Revised estimates are then 0.0758 and 0.301. So the effect of the correlation is to reduce the correction for the cholesterol coefficient. It is clearly important to allow for this correlation and it is unclear how MacMahon's method would do so.

9.6 Implications for study planning

This example has shown that to ignore the issue of measurement error can lead to grossly biased estimates of regression coefficients. In the above bivariate example, the odds ratio associated with an increase of 10 mm Hg in DBP is 1.49 if measurement error is ignored, but 2.13 when adjustments are made. In order to make these adjustments, knowledge is required of the covariance matrix of any measurement errors involved. This information may already be available from other studies, but if not then it is important in planning a study such as this to include in the design an element for estimating this covariance matrix.

Equation (9.3) is only approximate but it shows the dependence on the size of the validation study. The criterion for choosing the study sample size will usually be based on the precision required of the estimates of the regression coefficient. The implications for sample size will depend on the values of the parameters involved. The sizes of the main and validation studies, n and m respectively, are clearly important but so are the values of β and θ. It was noted earlier that usually the

second term in (9.3) will dominate other terms but for some values of β and θ that will not be the case. For the particular values in this study, with a large value of θ arising from the large measurement error variance, the first term in (9.3) is the dominant term unless m is very small. The results suggest that for the values of n considered here, a fairly small validation study would probably suffice. This may not always be the case and therefore there could be implications of sampling error in the estimation of σ_v.

Chapter 10

In this thesis we have introduced the logistic regression errors-in-variables problem. Throughout the chapters we have provided a framework for understanding the background to the problem, the methods that can be implemented for both ordinary logistic regression and the measurement error model, and an investigation into the best methods for estimating the measurement error model parameters. The issue of whether the measurement error standard deviation is known, either estimated from a validation study or internal to the main study has also been investigated. The thesis has been concluded with a view as to how this work can be included in studies being conducted today.

There are many examples in medical studies of mis-measured risk factors being used in relation to a disease status. In chapter 1 we introduced a few of these examples, where a correction method for measurement error would be required, but are not necessarily routinely applied. The fact that risk factors are known to be subject to

measurement error is well documented. There are also a number of methods in the statistical and epidemiological literature available to estimate the model parameters when the risk factor is known to be subject to measurement error. However, in practice these methods are not widely used. The aim of this thesis was to provide understanding of which methods could be used to estimate the model parameters and when the methods should be used, in-order to provide a practical guide to correction methods so that they can be more widely used in practice.

This thesis has provided a summary of some of the correction methods that are currently available, as well as a practical comparison of these methods, which has not been conducted elsewhere. This work compared a number of methods chosen from a range of techniques in different situations. By comparing the bias associated with estimating the model parameters, as well as likelihood and empirical standard errors, confidence intervals and 95% coverage terms two methods proved to provide the best estimates of the model parameters: a simple correction factor method, and a complex modelling technique. These methods were also compared with respect to their robustness to certain modelling assumptions. In conclusion, both when the model assumptions held and when they did not, the Reeves method provided comparable results to the conditional score method. As this method is simpler to implement and therefore more likely to be used in practice, this method is strongly recommended in all cases examined within this thesis.

A further technique that can be used in measurement error problems is the Bayesian approach. With the advent of specialized software, Bayesian techniques are more widely used than in previous years. They have been previously advocated for

measurement error problems but have never been practically compared to frequentist techniques. Though only a limited number of simulations could be conducted for the Bayesian approach, this technique proved to be a viable alternative to the Reeves method. By nature, the Reeves method is simpler to implement and therefore would probably be more widely used in practice. However, the Bayesian approach has a number of advantages, including its flexibility in modelling the situation and including prior information, so that this method should not be discounted when choosing a measurement error method.

In the linear regression measurement error problem, no analytical estimates of the model parameters could be identified when both the measurement errors were unknown. For the logistic regression measurement error problem, Kuchenhoff (1994) proved that all the model parameters are identifiable from a single study data set. This was true for both the classical and Berkson measurement error models but not for the probit regression measurement error problem. As Kuchenhoff reported, it does not seem possible to find a way of translating this theoretical result into a practical method for estimating all model parameters. The close links with the non-identifiable probit model suggest that the information in the data is not really sufficient to estimate all parameters in practice. The experiences with the Bayesian approach confirm that this appears to be the case and that in practice information on the parameters of the measurement error model is necessary.

Though there are a number of sophisticated techniques that can be used to estimate the model parameters in the logistic regression measurement error problem, the most commonly used method that appears to be used in epidemiology is a very ad-hoc

method. This is the method of MacMahon, suggested for use in relating blood pressure to coronary heart disease. Our investigation showed that this particular method was not as efficient and robust as the Reeves method. Therefore, studies using this particular method could mis-interpret the true relationship between blood pressure and disease. This proved that without some understanding of the comparison between correction methods and which are the best methods to use, studies could still be conducted using inappropriate correction methods.

This thesis only practically compared a few different methods. As we have shown, there are many more correction methods that have been designed for particular uses. The Reeves method proved, in the cases that we tested, to produce sensible estimates of the model parameters for all values of the measurement error standard deviation. To understand whether this method should be used in all cases, further practical work must be conducted in order to compare the Reeves method with other more specifically designed methods. Further to this, this thesis only practically compared methods where there was a single explanatory variable in the model, a practical comparison of the methods available for more than one variable is also required.

The results from the investigation using a Bayesian approach to the measurement error problem, suggested that the Bayesian method could be comparable to the simpler Reeves method. A larger simulation study is required, in order to examine the estimates of the model parameters in various cases where the prevalence of the disease changes, as well as the measurement error model and distribution of the measurement error. As this method is more involved than a simple correction method,

a practical comparison is required to frequentist methods to see whether there is any gain in using the more flexible Bayesian approach.

Limited work was conducted in this thesis on the Kuchenhoff theory of being able to estimate all three of the model parameters from a single dataset and so far, it seems that this work has theoretical interest only. However, the issue of identifiability remains open. From the current results it appears that there is not sufficient information to estimate all of the parameters efficiently. Further examination of the probit regression model and why the theory does not hold for this case could provide information as to why this is the case. The theoretical side of the problem could also be taken further by examining whether there are other measurement error distributions where this theory also holds or perhaps where the X values are not normally distributed. Further theoretical work could perhaps lead to a practical use.

Over the past few years a number of authors have considered the errors-in-variables problem from different perspectives. These include both Bayesian (Berry et al (2002), Gossl et al (2001) and Roddam (2004)) and Classical (Carroll et al (2004), Cui et al (2004), Freedman et al (2004), Li (2002), Schennach (2004), Vajk et al (2003), Wang (2002, 2003) and Wang et al (2001) approaches. Further work has also considered more practical applications for example, sample size calculations when covariates are known to be subject to measurement error (Devine (2003) and Tosteson (2003)) as well as practical applications (Pilote et al (2000) and Rindskopf (2004)). The continued work in this area confirms the requirements to continue to understand when particular techniques should be used and therefore become standard procedure when covariates are known to be subject to measurement error.

In conclusion, measurement error associated with a risk factor can have a major effect when estimating the relationship between the risk factor and a disease status. Therefore, medical studies should be designed and implemented taking into account the effect of the measurement error. Studies should be designed with a large a sample size as possible, with a moderately sized subsidiary or validation study in order to determine the nature and size of the measurement error. Lastly, instead of the ordinary logistic regression method, the most appropriate correction method should be used that fits the problem and the model assumptions. Overall, it is important that such methods are made widely available, such as being a part of standard statistical packages, and that the use of such methods becomes standard.

Appendix A

A.1 Overview

This chapter contains the proofs of the results that were quoted within the study that had no associated reference.

A.2 Logistic Model Variance

The logistic probability distribution function is

$$\frac{e^{-x}}{\left(1+e^{-x}\right)^2} \qquad -\infty < x < \infty$$

By symmetry $E(X)=0$. Hence,

$$E(X^2) = 2\int_0^{\infty} x^2 \frac{e^{-x}}{\left(1+e^{-x}\right)^2} dx \qquad \text{by symmetry}$$

$$= 2\int_0^{\infty} x^2 e^{-x}\left(\sum_0^{\infty}(k+1)(-1)^k e^{-kx}\right)dx$$

$$= 2\int_0^{\infty}\sum_0^{\infty} x^2 (k+1)(-1)^k e^{-(k+1)x} dx$$

$$= 2\sum_0^{\infty}(-1)^k \int_0^{\infty}(k+1)x^2 e^{-(k+1)x} dx$$

$$= 4\sum_{0}^{\infty}\frac{(-1)^k}{(k+1)^2}$$

$$= 4\left(1-\frac{1}{2^2}+\frac{1}{3^2}-\frac{1}{4^2}+\ldots\right)$$

$$= 4\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\right)-8\left(\frac{1}{2^2}+\frac{1}{4^2}+\frac{1}{6^2}+\ldots\right)$$

$$= 4\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\right)-2\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\right)$$

$$= 2\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\right)$$

$$= 2\times\frac{\pi^2}{6}=\frac{\pi^2}{3}=Var(X)$$

A.3 Delta method

To find the variance of the variable Z when

$$Z = f(X,Y)$$

using the delta method.

$$Z = f(X,Y)$$

$$= f(X-\mu_x+\mu_x, Y-\mu_y+\mu_y)$$

$$\cong f(\mu_x,\mu_y)+(X-\mu_x)\frac{\partial f}{\partial X}\Big|_{\underline{\mu}}+(Y-\mu_y)\frac{\partial f}{\partial Y}\Big|_{\underline{\mu}}+\frac{(X-\mu_x)^2}{2}\frac{\partial^2 f}{\partial X^2}\Big|_{\underline{\mu}}$$

$$+\frac{(Y-\mu_y)^2}{2}\frac{\partial^2 f}{\partial Y^2}\Big|_{\underline{\mu}}+(X-\mu_x)(Y-\mu_y)\frac{\partial^2 f}{\partial X\partial Y}\Big|_{\underline{\mu}}$$

$$E(Z)\cong f(\mu_x,\mu_y)+\frac{\sigma_x^2}{2}\frac{\partial^2 f}{\partial X^2}\Big|_{\underline{\mu}}+\frac{\sigma_y^2}{2}\frac{\partial^2 f}{\partial Y^2}\Big|_{\underline{\mu}}+\rho\sigma_x\sigma_y\frac{\partial^2 f}{\partial X\partial Y}\Big|_{\underline{\mu}}$$

$$Z-E(Z)\cong(X-\mu_x)\frac{\partial f}{\partial X}\Big|_{\underline{\mu}}+(X-\mu_x)\frac{\partial f}{\partial X}\Big|_{\underline{\mu}}$$

$$Var(Z)=E(Z-E(Z))^2$$

$$=\sigma_x^2\left(\frac{\partial f}{\partial X}\Big|_{\underline{\mu}}\right)^2+\sigma_y^2\left(\frac{\partial f}{\partial Y}\Big|_{\underline{\mu}}\right)^2+2\rho\sigma_x\sigma_y\left(\frac{\partial f}{\partial X}\Big|_{\underline{\mu}}\right)\left(\frac{\partial f}{\partial Y}\Big|_{\underline{\mu}}\right)$$

If (X,Y) are independent then $\rho = 0$.

A.4 Proof of statement in section 4.3.3

Carroll, Spiegelman, Lan, Bailey & Abbott (1984) proved that the integral

$$P(Y=1|Z)=\int P(Y=1|X)f(X|Z)dX \qquad (4.4)$$

can be analytically evaluated if the probit model is substituted for the logistic model through the equation

$$\int_{-\infty}^{\infty}\Phi(a+bx)\phi(x)dx=\Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

The following is a proof of this statement.

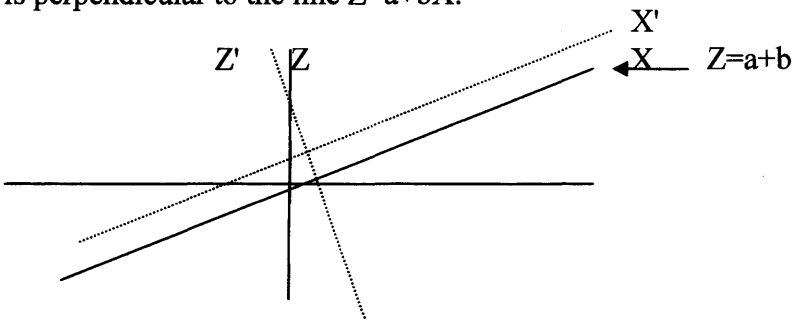
$$\int_{-\infty}^{\infty}\Phi(a+bx)\phi(x)dx$$

$$=\int_{-\infty}^{\infty}\left(\int_{-\infty}^{a+bx}\phi(y)dy\right)\phi(x)dx$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{a+bx}\phi(x)\phi(z)dxdz$$

Where $\phi(x)\phi(z)$ is the joint pdf of xz which is a bivariate normal with $\rho=0$, $\sigma_X=\sigma_Z=1$ and $\mu_X=\mu_Z=0$.

The two densities are rotationally invariant therefore, we can rotate the axis such that the new Z-axis is perpendicular to the line Z=a+bX.



Therefore, the perpendicular distance from the line Z=a+bX to the origin is $\frac{a}{\sqrt{1+b^2}}$.

By changing the limits of integration in line with the above transformation we have

$$\frac{1}{2\pi}\int_{X'=-\infty}^{\infty}\int_{Z'=-\infty}^{\frac{a}{\sqrt{1+b^2}}}\phi(Z')\phi(X')dZ'dX'$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\frac{a}{\sqrt{1+b^2}}}\frac{e^{-\frac{1}{2}Z'^2}}{\sqrt{2\pi}}dZ'\right)e^{-\frac{1}{2}X'^2}dX'$$

$$= \int_{-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi(Z')dZ' = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

Hence,

$$P(Y=1|Z) = \Phi\left\{\frac{\alpha_t + \beta_t\left(\frac{\sigma_x^2}{\sigma_z^2}z + \frac{\sigma_v^2}{\sigma_z^2}\mu_x\right)}{\left(1+\beta_t^2\frac{\sigma_v^2\sigma_x^2}{\sigma_z^2}\right)^{\frac{1}{2}}}\right\}$$

A.5 Proof of variance statement 2.19

To prove that

$$Var(\hat{\beta}_t) = \frac{\sum p_i(1-p_i)}{\sum x_i^2 p_i(1-p_i)\sum p_i(1-p_i) - \left(\sum x_i p_i(1-p_i)\right)^2} \qquad (2.19)$$

requires the use of Fisher's Information matrix defined by

$$I_\theta = E_\theta\left[\left\{\frac{\partial \log p(x,\theta)}{\partial\theta}\right\}^2\right]$$

It can be shown that maximum likelihood estimators have a covariance matrix which, for large samples, can be approximated by the inverse of this matrix.

The logistic log likelihood is defined by

$$\ell = LogL = \sum_{i=1}^{n} Y_i \log p_i + \sum_{i=1}^{n}(1-Y_i)\log(1-p_i)$$

The information matrix is calculated from the matrix of the second derivatives of the logistic log likelihood, so from (2.14)-(2.17)

$$I(\alpha_t, \beta_t) = \begin{pmatrix} \sum_{i=1}^{n} p_i(1-p_i) & \sum_{i=1}^{n} x_i p_i(1-p_i) \\ \sum_{i=1}^{n} x_i p_i(1-p_i) & \sum_{i=1}^{n} x_i^2 p_i(1-p_i) \end{pmatrix}$$

On taking the inverse

$$I(\alpha_t,\beta_t)^{-1}=\frac{1}{\sum_{i=1}^{n}x_i^2p_i(1-p_i)\sum_{i=1}^{n}p_i(1-p_i)-\left(\sum_{i=1}^{n}x_ip_i(1-p_i)\right)^2}\begin{pmatrix}\sum_{i=1}^{n}x_i^2p_i(1-p_i) & \sum_{i=1}^{n}x_ip_i(1-p_i)\\ \sum_{i=1}^{n}x_ip_i(1-p_i) & \sum_{i=1}^{n}p_i(1-p_i)\end{pmatrix}$$

Therefore,

$$Var(\hat{\beta}_t)=\frac{\sum p_i(1-p_i)}{\sum x_i^2p_i(1-p_i)\sum p_i(1-p_i)-\left(\sum x_ip_i(1-p_i)\right)^2}$$

References

Armstrong B. 1985. Measurement Error in the Generalised Linear Model; *Commun. Statist* **14**(3):529-544.

Armstrong BG, Whittemore AS, Howe GR. 1989. Analysis of Case-control Data with Covariate Measurement Error: Application to Diet and Colon Cancer; *Statistics in Medicine* **8:** 1151-1163

Armstrong BG. 1990. The Effects of Measurement Errors on Relative Risk Regressions; *American Journal of Epidemiology* **132:**1176-1184

Barron BA. 1977. The Effects of Misclassification on the Estimation of Relative Risk; *Biometrics* **6:**414-418

Bashir SA, Duffy SW. 1995. Correction of Risk Estimates for Measurement Error in Epidemiology; *Methods of Information in Medicine* **34:**503-510

Berry SM, Carroll RJ, Ruppert D. 2002. Bayesian smoothing and regression splines for measurement error problems; *Journal of the American Statistical Association* **457**:160-169.

Blackwood L. 1988. Latent Variable Models for the Analysis of Medical Data with Repeated Measures of Binary Variables; *Statistics in Medicine* **7:**975-981

Bowman, A.W. and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, *Oxford Science Publications.*

Bross I. 1954. Misclassification in 2 x 2 Tables; *Biometrics* **12:**478-486

Buzas JS. 1997. Instrumental variable estimation in nonlinear measurement error models; *Commun. Statistit. –theory meth* **26**(12):2861-2877

Caroll RJ. 1989. Covariance Analysis in Generalized Linear Measurement Error Models; *Statistics in Medicine* **8:** 1075-1093

Carroll RJ, Spiegelman C, Lan KK, Bailey KT, Abbott RD. 1984. On Errors-in-Variables for Binary Regression Models; *Biometrics* **71**(1):19-25

Carroll RJ, Freedman LS, Kipmis V, Li Li. 1998. A new Class of Measurement-Error Models, with Applications to Dietary Data; *The Canadian Journal of Statistics* **26**(3):467-477.

Carroll RJ, Hall P. 2004. Low order approximations in deconvolution and regression with errors in variables; *Journal of the Royal Statistical Society Series B – Statistical Methodology* **66**:31-46.

Carroll RJ, Ruppert D, Stefanski LA. 1995. Measurement error in nonlinear models; *Monographs on Statistics and Applied Probability 63: Chapman & Hall/CRC*

Carroll RJ. 1997. Surprising Effects of Measurement Error on an Aggregate Data Estimator; *Biometrika* **84**(1):231-234

Chavance M, Dellatolas G, Lellouch J. 1992. Correlated Nondifferential Misclassifications of Disease and Exposure: Application to a Cross-Sectional Study of the Relation between Handedness and Immune Disorders; *International Journal of Epidemiology* **21**(3):537-546.

Cheng C & Van Ness JW. 1999. Statistical Regression with Measurement Error. Kendall's Library of Statistics 6; *Arnold*

Chen MH, Ibrahim JG, Yiannoutsos C. 1999. Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models; *B-Statistical Methodology* **61**(1):223-242

Chen TT. 1989.A Review of Methods for Misclassified Categorical Data in Epidemiology; *Statistics in Medicine,* **8:**1095-1106

Clarke R, Shipley M, Lewington S, Youngman L, Collins R, Marmot M, Peto R. 1999. Underestimation of Risk Associations Due to Regression Dilution in Long-Term Follow-up of Prospective Studies, *American Journal of Epidemiology* 150(4):341-353

Cleveland WS. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots; *Journal of the American Statistical Association* **74**(368):829-836

Collett D. 1991. Modelling Binary Data. *Chapman & Hall*

Copas JB. 1988. Binary Regression Models for Contaminated Data; *J R Statist. Soc* **50**(2):225-265.

Copeland KY, Checkoway H, McMichael AJ, Holbrook RH. 1977. Bias Due to Misclassification in the Estimation of Relative Risk; *American Journal of Epidemiology* **105**(5):488-495.

Cox B, Elwood M. 1991. The Effect on the Stratum-specific Odds Ratios of Nondifferential Misclassification of a Dichotomous Covariate; *American Journal of Epidemiology* **133**(2):202-207.

Cox DR, Hinkely DV. 1974. Theoretical Statistics; *Chapman & Hall: Appendix A*

Crouch EAC, Spiegelman D. 1990. The evaluation of integrals of the form $\int dx$: Application to logistic-normal models; *Journal of the American Statistical Association* **85**(410):464-469

Cui HJ, Ng KW, Zhu LX. 2004. Estimation in mixed effects model with errors in variables; *Journal of Multivariate Analysis* **91**:53-73.

Dawber TR. 1980. The Framingham Study. The epidemiology of atherosclerotic disease. *Cambridge: Harvard University Press*

Dellaportas P, Stephens DA. 1995. Bayesian-Analysis of errors-in-variables regression-models; *Biometrics* **51**(3):1085-1095

Demidenko E, Spiegelman D. 1997. A Paradox: More Measurement Error can lead to more Efficient Estimate; *Commun. Statist. – Theory Meth* **26**(7):1649-1675

Devanarayan V, Stefanski LA. 2002. Empirical simulation extrapolation for measurement error models with replicate measurements; *Statistics & Probability Letters* **59**:219-225.

Devine OJ, Smith JM. 1998. Estimating Sample Size for Epidemiologic Studies: The Impact of Ignoring Exposure Measurement Uncertainty; *Statistics in Medicine* **17**(12):1375-1389

Devine O. 2003. The impact of ignoring measurement error when estimating sample size in epidemiological studies; *Evaluation & the Health Professions* **26**:315-339.

Diamond EL, Lilienfeld AM. 1962. Effects of Errors in Classification and Diagnosis in Various Types of Epidemiological Studies; *AJPH* **52**(7):1137-1144.

Diamond EL, Lilienfeld AM. 1962. Misclassification Errors in 2 x 2 Tables with one Margin Fixed: Some further comments; *AJPH* **52**(12):2106-2110.

Dosemici M, Wacholder S, Lubin JH. 1990. Does Nondifferential Misclassification of Exposure always Bias a True Effect Toward the Null Value; *American Journal of Epidemiology* **1990:**746-748.

Duffy SW, Maximovitch DM, Day NE. 1992. External validation, repeat determination, and precision of risk estimation in misclassified exposure data in epidemiology; *Journal of Epidemiology and Community Health* **46:**620-624

Espeland MA, Hui SL. 1987. A General Approach to Analyzing Epidemiologic Data That Contain Misclassification Errors; *Biometrics* **43:**1001-1012.

Evans RC, Fear S, Ashby D, Hackett A, Williams E, Van Der Vliet M, Dunstan FD, Rhodes JM. 2002. Diet and colorectal cancer: an investigation of the lectin/galactose hypothesis.*Gastroenterology*. **122**:1784-92.

Flegal KM, Browne C, Haas JD. 1986. The Effects of Exposure Misclassification on Estimates of Relative Risk; *American Journal of Epidemiology* **123**(4):736-751.

Fortran Library Mark 1 http://www.nag.co.uk/numeric/FLOLCH/MK17.html

Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. 2004. A new method for dealing with measurement error in explanatory variables of regression models; *Biometrics* **60**:172-181.

Fuller WA. 1988. Measurement Error Models; *Wiley, New York*

Gamerman D. 1997. Markov Chain Monte Carlo Stochastic simulation for Bayesian Inference, *Chapman & Hall*

Gelman A, Rubin R. (1992a). A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statisics 4* 625-631.

Gelman A, Rubin R. (1992b). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**:457-511.

Geweke J. 1989. Bayesian inference in econometric models using Monte Carlo integration, *Econometrica* **57**:1317-1339.

Geweke J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion), *Bayesian Statistics 4*: 169-193

Gossl C, Kuchenhoff H. 2001. Bayesian analysis of logistic regression with an unknown change point and covariate measurement error; *Statistics in Medicine* **20**:3109-3121.

Greenland S. 1980. The Effect of Misclassification in the Presence of Covariates; *American Journal of Epidemiology* **112**(4):564-418

Greenland S. 1982. The Effect of Misclassification in Matched-Pair Case-control Studies; *American Journal of Epidemiology* **116** (2):402-406

Guildea ZES , Fone DL, Dunstan FD, Sibert JR, Cartlidge PHT. 2001. Social deprivation and the causes of stillbirth and infant mortality, *Archives of Disease in Childhood* **84**: 307-310

Hanfelt JJ, Liang KY. 1997. Approximate Likelihoods for Generalized Linear Errors-in-Variables Models; *Journal of the Royal Statistical Society Series B-Methodological* **59**(3):627-637

Haukka JK. 1995. Correction for Covariate Measurement Error in Generalized Linear Models – A Bootstrap Approach; *Biometrics* **51:** 1127-1132

Hill K, Iles TC, Nix ABJ. 1999. Asymptotic Information and Variance-Covariance Matrices for the Linear Structural Model; *Journal of the Royal Statistical Society Series D* **48**: 477-493

Hosmer DW, Lemeshow. 1980. Goodness of fit tests for the multiple logistic regression model; *Commun. Statist. –theory Meth.* **A9**(10):1043-1069

Huber PJ. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkley Symposium* **1**:221-233

Kendall MG, Stuart A. 1979. The Advanced Theory of Statistics, Vol 2. (4th edn). *Griffin, London*

Kim MY, Zeleniuch Jacquotte A. 1997. Correcting for Measurement Error in the Analysis of Case-Control Data with Repeated Measurements of Exposure; *American Journal of Epidemiology* **145**(11):1003-1010

Kosinski AS, Flanders WD. 1999. Evaluating the Exposure and Disease Relationship with Adjustment for Different Types of Exposure Misclassification: A Regression Approach; *Statistics in Medicine* **18:**2795-2808.

Küchenhoff H, Carroll RJ. 1997. Segmented Regression with Errors in Predictors: Semi-Parametric and Parametric Methods; *Statistics in Medicine* **16:**169-188

Kuchenhoff H. 1995. The identification of logistic regression model with errors in the variables; *Statistical Papers* **36**:41-48

Kuha J. 1994. Corrections for Exposure Measurement Error in Logistic Regression Models with an Application to Nutritional Data; *Statistics in Medicine* **13:**1135-1148

Kulathinal SB, Kuulasmaa K, Gasbarra D. 2002. Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations; *Statistics in Medicine* **21**:1089-1101.

Kupper LL. 1984. Effects of the Use of Unreliable Surrogate Variables on the Validity of Epidemiological Research Studies; *American Journal of Epidemiology* **120**(4):643-648

Lagakos SW. 1988. Effects of Mismodelling and Mismeasuring Explanatory Variables on Test of their Association with a Response Variable; *Statistics in Medicine* **7:**257-274

Li T. 2002. Robust and consistent estimation of nonlinear errors-in-variables models; *Journal of Econometrics* **110**:1-26.

Liu Xinhua, Liang Kung-Yee. 1991. Adjustment for non-differential misclassification error in the Generalized Linear Model; *Statistics in Medicine,* **10:**1197-1211

Lyles RH, Kupper LL.1999. A Note on Confidence Interval Estimation in Measurement Error Adjustment; *American Statistician* **53**(3):247-253

Macmahon S, Peto R, Cutler J *et al.* 1990. Blood Pressure, Stroke and Coronary Heart Disease – Prolonged Differences in Blood Pressure: Prospective Observational Studies Corrected for the Regression Dilution Bias; *The Lancet* **335:**765-74

Mallick BK, Gelfand AE. 1996. Semiparametric Error-in-Variables Models – A Bayesian Approach; *Journal of Statistical Planning and Inference* **53**(3):307-321

Marshall JR. 1989. The Use of Dual or Multiple Reports in Epidemiologic Studies; *Statistics in Medicine* **8:**1041-1049.

McCarron P, Greenwood R, Elwood P, Shlomo YB, Bayer A, Baker I, Frankel S, Ebrahim S, Murray L, & Smith GD. (2001) The incidence and aetiology of stroke in the Caerphilly and Speedwell Collaborative Studies II: risk factors for ischaemic stroke, *Public Health* **115**, 12-20

McCullagh P, Nelder JA. 1989. Generalized Linear Models, 2nd edn; *Chapman & Hall, London*

Mertens TE. 1995. Estimating the effects of Misclassification; *The Lancet* **342:**418-421.

Michalek JE, Tripathi RC. 1980. The Effect of Errors in Diagnosis and Measurement on the Estimation of the Probability of an Event; *Journal of the American Statistical Association* **75**(371):713-721

Miettinen O. 1974. Confounding and Effect-Modification; *American Journal of Epidemiology* **100**(3):350-353

Mote VL, Anderson RL. 1965. An Investigation of the Effect of Misclassification on the Properties of x^2-tests in the Analysis of Categorical Data; *Biometrika* **52**(1&2):95-109.

Muller P, Roeder K. 1997.A Bayesian Semiparametric Model for Case-Control Studies with Errors in Variables;*Biometrika***84**(3):523-537

Neuhaus JM. 1999. Bias and Efficiency Loss due to Misclassified Responses in Binary Regression; *Biometrika* **86**(4):843-855.

Newell D. 1962. Errors in the Interpretation of Errors in Epidemiology; *AJPH* **52**(11):1925-1928.

Ollerton RL, Dunstan FD, Playle R, Luzio SD, Ahmen K, Owens DR. 1999. Day to Day Variability of Fasting Plasma Glucose in Newly Diagnosed Type 1 Diabetic Subjects; *Diabetes Care* **22**(3).

Oppenheimer L, Kher U. 1999. The Impact of Measurement Error on the Comparison of Two Treatments using a Responder Analysis; *Statistics in Medicine* **18:**2177-2188

Palmgren J, Ekholm A. 1987. Exponential Family Non-linear Models for Categorical Data with Errors of Observation; *Applied Stochastic Models and Data Analysis* **3:**111-124

Palta M, Lin CY. 1999. Latent Variables, Measurement Error and Methods for Analysing Longitudinal Binary and Ordinal Data; *Statistics in Medicine* **18:**385-396.

Pilote I, Joseph L, Belisle P, Robinson K, Van Lente F, Tager IB. 2000. Iron stotes and coronary artery disease: A clinical application of a method to incorporate measurement error of the exposure in a logistic regression model; *Journal of Clinical Epidemiology* **53**:809-816.

Prospective Studies Collaboration. 2002. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies, *Lancet* **360**:1903-1913

Reeves GK, Cox DR, Darby SC, Whitley E. 1998. Some Aspects of Measurement Error in Explanatory Variables for Continuous and Binary Regression Models; *Statistics in Medicine* **17:**2157-2177

Richardson S, Leblond L. 1997. Some comments on misspecification of priors in Bayesian Modelling of measurement error problems; *Statistics in Medicine* **16**:203-213

Richardson S, Leblond L, Jaussent I, Green PJ. 2002. Mixture models in measurement error problems, with reference to epidemiological studies; *Journal of the Royal Statistical Society Series A – Statistics in Society* **165**:549-566.

Richardson S, Gilks W. 1993. Conditional Independence models for epidemiological studies with covariate measurement error; *Statistics in Medicine* **12**:1703-1722

Rindskopf D, Strauss S. 2004. Determining predictors of true HIV status using an errors-in-variables model with missing data; *Structural Equation Modelling – A Multidisciplinary Journal* **11**:51-59.

Robins JM, Rotnitzky A, Zhao LP. 1994. Estimation of Regression Coefficients when some Regressors are not always observed; *Journal of the American Statistical Association* **89**(427)

Roddam AW. 2004. Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments; *Journal of the Royal Statistical Society Series A – Statistics in Society* **167**:570-571.

Rosner B, Spiegelman D, Willett WC. 1990. Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Measurement Error: The Case of Multiple Covariates Measured with Error; *American Journal of Epidemiology* **132**(4):734-745

Rosner B, Spiegelman D, Willett WC. 1992. Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Random Within-Person Measurement Error; *American Journal of Epidemiology* **136**(11):1400-1413

Rosner B, Willett WC, Spiegelman D. 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error; *Statistics in Medicine* **8**:1051-1069

Savitz DA, Baron AE. 1989. Estimating and Correcting for Confounder Misclassification; *American Journal of Epidemiology* **129**(5):1062-1071.

Schafer DW. 1987. Covariate Measurement Error in Generalized Linear Models; *Biometrika* **74** (2):385-91

Schennach SM. 2004. Estimation of nonlinear models with measurement error; *Econometrica* **72**:33-75.

Schmid C, Rosner B. 1993. A Bayesian approach to logistic regression models having measurement error following a mixture distribution; *Statistics in Medicine* **12**:1141-1153

Selen J. 1986. Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data; *Journal of the American Statistical Association* **81**(393).

Spiegelman D, Casella M. 1997. Fully Parametric and Semi-Parametric Regression Models for Common Events with Covariate Measurement Error in Main Study/ Validation Study Designs; ***Biometrics*** **53**:395-409.

Spiegelman D, Casella M. 1997. Fully Parametric and Semi-Parametric Regression Models for Common Events with Covariate Measurement Error in Main Study Validation Study Designs; *Biometrics* **53**(2):395-409

Spiegelman D, Valanis B. 1998. Correcting for Bias in Relative Risk Estimates Due to Exposure Measurement Error: A Case Study of Occupational Exposure to Antineoplastics in Pharmacists; *American Journal of Public Health* **88**(3):406-412

Srivastava AK, Shalabh. 1997. A New Property of Stein Procedure in Measurement Error Model; *Statistics & Probability Letters* **32**:231-234.

Stefanski LA, Carroll RJ. 1987. Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models; *Biometrika* **74**(4):703-716

Stefanski LA, Carroll RJ. 1990. Structural Logistic Regression Measurement Error Model; *Contemporary Mathematics* **112**:115-127

Thomas D, Stram D, Dwyer J. 1993. Exposure Measurement Error: Influence on Exposure-Disease Relationships and Methods of Correction; *Ann Rev. Public Health* **14**:69-93

Thoresen M, Laake P. 1999. Instrumental Variable Estimation in Logistic Measurement Error Models by Means of Factor Scores; *Commun. Statist.- Theory Meth.* **28**(2):297-313.

Tosteson TD, Buzas JS, Demidenko E, Karagas M. 2003. Power and sample size calculations for generalized regression models with covariate measurement error; *Statistics in Medicine* **22**:1069-1082.

Tosteson TD, Stefanski LA, Schafer DW. 1989. A Measurement-Error model for Binary and Ordinal Regression; *Statistics in Medicine* **8:**1139-1147

Townsend P, Phillimore P, Beattie A 1988. Health and Deprivation, *Croom Helm, New York*

Thurigen D, Spiegelman D, Blettner M, Heuer C, Brenner H. 2000. Measurement error correction using validation data: a review of methods and their applicability in case-control studies; *Statistical Methods in Medical research* **9**: 447-474

Vajk I, Hetthessy J. 2003. Identification of non-linear errors-in-variables models; *Automatica* **39**:2099-2107.

Wald JW, Kennard E, Hackshaw A, McGuire A. 1998. Anti-natal screening for Down's Syndrome; *Health Technology Assessment* **2**(1)

Walker AM, Blettner M. 1985. Comparing Imperfect Measures of Exposure; *American Journal of Epidemiology* **121**(6)

Walker AM, Lanes SF. 1991. Misclassification of Covariates; *Statistics in Medicine* **10:**1181-1196

Walter SD, Irwig LM. 1988. Estimation of Test Error Rates, Disease Prevalence and Relative Risk from Misclassified Data: A Review; *J Clin Epidemiology* **41**(9):923-937.

Wang CY, Wang S. 1997. Semiparametric Methods in Logistic Regression with Measurement Error; *Statistica Sinica* **7:**1103-1120

Wang LQ. 2002. A simple adjustment for measurement errors in some limited dependent variable models; *Statistics & Probability Letters* **58**:427-433.

Wang LQ. 2003. Estimation of nonlinear Berkson-type measurement error models; *Statistica Sinica* **13**:1201-1210.

Wang N, Lin X, Gutierrez RG, Carroll RJ. 1998. Bias Analysis and SIMEX Approach in Generalised Linear Mixed Measurement Error Models; *Journal of the American Statistical Association* **93**(441):249-261.

Wang SJ, Wang CY. 2001. A note on kernel assisted estimators in missing covariate regression; *Statistics & Probability Letters* **55**:439-449.

Wedderburn RWM. 1974. Quasi-likelihood functions, Generalised Linear Models, and the Gauss-Newton method; *Biometrika* **61**(3):43-447.

Willett W. 1989. An Overview of Issues related to the Correction of Non-differential Exposure Measurement Error in Epidemiologic Studies; *Statistics in Medicine* **8:**1031-1040.

White I, Frost C, Tokunaga S. 2000. Correcting for measurement error in binary and continuous variables using replicates.

Yanex III ND, Kronmal RA, Shemanski LR. 1998. The Effects of Measurement Error in Response Variables and Tests of Association of Explanatory Variables in Change Models; *Statistics in Medicine* **17**:2597-2606.

Zhao YJ, Lee AH, Vanhui Y. 1994. Influence Diagnostics for Generalized Linear Measurement Error Models; *Biometrics* **50**(4):1117-1128