Evaluating the Performance of Simulation Extrapolation and Bayesian Adjustments for Measurement Error

by

Alexandra Romann

B.Sc., Thompson Rivers University, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Statistics)

The University of British Columbia

(Vancouver)

August, 2008

Abstract

Measurement error is a frequent issue in many research areas. For instance, in health research it is often of interest to understand the relationship between an outcome and an exposure, which is often mismeasured if the study is observational or a gold standard is costly or absent. Measurement error in the explanatory variable can have serious effects, such as biased parameter estimation, loss of power, and masking of the features of the data. The structure of the measurement error is usually not known to the investigators, leading to many difficulties in finding solutions for its correction.

In this thesis, we consider problems involving a correctly measured continuous or binary response, a mismeasured continuous exposure variable, along with another correctly measured covariate. We compare our proposed Bayesian approach to the commonly used simulation extrapolation (SIMEX) method. The Bayesian model incorporates the uncertainty of the measurement error variance and the posterior distribution is generated by using the Gibbs sampler as well as the random walk Metropolis algorithm. The comparison between the Bayesian and SIMEX approaches is conducted using different cases of a simulated data including validation data, as well as the Framingham Heart Study data which provides replicates but no validation data. The Bayesian approach is more robust to changes in the measurement error variance or validation sample size, and consistently produces wider credible intervals as it incorporates more uncertainty.

The underlying theme of this thesis is the uncertainty involved in the estimation of the measurement error variance. We investigate how accurately this parameter has to be estimated and how confident one has to be about this estimate in order to produce better results by choosing the Bayesian measurement error correction over the naive analysis where measurement error is ignored.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I thank my supervisor, Professor Paul Gustafson, and my co-supervisor, Dr. Nhu Le, for their support throughout my study period in the Department of Statistics. It was an honour to work with both of them and their guidance helped me complete this thesis.

I express sincere gratitude to Professors Lang Wu, Harry Joe, Arnaud Doucet, Raphael Gottardo, Matias Salibián-Barrera, and Michael Schulzer for their excellent teaching and mentorship. I also thank Mike Marin for his friendship and encouragement, as well as the office staff for their help.

I am grateful to all the graduate students for making the department a great place to study. In particular, I thank Michael Regier, Justin Harrington, Aline Tabet, Kimberly Fernandes, and Luke Bornn for their friendship and advice.

On a personal note, I am eternally grateful to my parents, Ursula and Max Romann, who have provided me with an abundance of kind support all of my life, and to Neal Dustin for his love and patience.

Alexandra Romann

The University of British Columbia
August 2008

To My Parents

Chapter 1

Introduction

In many research areas the measurement accuracy of a variable is a frequent issue. For instance, in health research it is often of interest to understand the relationship between an outcome Y and an exposure X. When an instrument is available that measures X correctly, such a tool is called a gold standard. Suppose that X can only be measured imprecisely due to the absence or high cost of a gold standard. Examples of such a scenario are intakes of foods or drugs, and exposure to airborne pollutants or radiation. If the covariate is a categorical variable, the problem with such imprecise measurement is called misclassification, whereas it is referred to as measurement error for a continuous variable. In this thesis, we will only concern ourselves with the latter case. There exists vast literature on misclassification, for example see Gustafson (2004). Covariate measurement error problems are concerned with inference on regression coefficients where one or more of the covariates are measured imperfectly. It has been documented extensively that treating this surrogate variable as if it were the true exposure can lead to very poor results. We illustrate this with an example in Chapter 4. Carroll, Ruppert, Stefanski and Crainiceanu (2006) call the effect of ignoring imprecision in mismeasured covariates the Triple Whammy of measurement error because it causes bias in parameter estimation, leads to a loss of power, and masks the features of the data.

Many methods have been proposed for countering these issues. We will focus on the first problem of how to offset the bias caused by measurement error. It is interesting that the magnitude of a regression coefficient of the incorrectly measured variable is often biased towards zero, a phenomenon called attenuation. In other words, it will underestimate the true value. The actual measurement that is made for X is the surrogate and incorrect exposure variable X^*. Further, we denote the correctly measured response by Y and the other correctly observed covariates by Z. The set of observed covariates is thus $\{X^*, Z\}$.

Measurement error problems can be classified into nondifferential and differential problems, the first of which occurs when X^* contains no information about Y other than what is available in X and Z. In this case, the distribution of the mismeasured surrogate depends only on the true explanatory variable and not on the response. Otherwise, the measurement error is considered differential.

For a given problem it is important to know how the error is arising in the covariate. In the classical additive error model $X^* = X + U$, where U is the measurement error, and $U \perp (X, Y, Z)$, $E(U|X) = 0$. This model is used when the error-prone covariate is measured uniquely to an individual, which is especially true when the measurement can be replicated. Examples of classical additive error arise in nutrition surveys and blood pressure measurements. However, if all subjects in a small group are given the same value of the error-prone covariate and thus the independence assumption between U and X is not satisfied, then the Berkson error model is better suited. In this model $X = X^* + U$ where $U \perp X^*, Y, Z$. This applies to dust exposure data of miners, for example. Employees having worked in the mine for a fixed number of years are assigned the same exposure to dust, even though the true exposure is almost certainly unique to each individual.

The models discussed above are additive measurement error models, however, the error can also arise in a multiplicative fashion. Examples are problems where the exposure X is positive and skewed, which occurs frequently in epidemiological problems. In that case, the observed exposure might be assumed to equal the product of the true exposure and the measurement error such that $X^* = XU$. Log-exposure can be used to convert such a case to an additive measurement error model. It then follows that $\log X^* = \log X + \log U$.

1.1 Overview of Some Currently Available Methods

Many methods have been proposed to deal with measurement error. They can be broadly grouped into functional and structural models (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). In functional modelling, the Xs are regarded as fixed or random with no or minimal distribution assumptions.

In structural models the Xs are regarded as random variables, thus the potential for exposure model misspecification arises.

1.1.1 Functional Methods

In order to use functional methods, the size of the measurement error needs to be estimated. In some cases, a gold standard test is available and one is able to observe X directly for some subset of the data. This is the validation data from which the measurement error variance is estimated. In other cases replication data or measurements from another instrument may be available.

A simple and quite general method called regression calibration (Pierce and Kellerer, 2004) is appropriate when such a gold standard or replicates are available and a linear measurement error with constant variance applies. In this method $E(X|X^*, Z)$ is estimated and the unobserved Xs are replaced by these estimates. Then a standard analysis is run to obtain the parameter estimates. The resulting standard errors are adjusted to account for the estimation of the parameters, using bootstrap, for example. This method has been shown to be very useful for generalized linear models, but performs poorly for highly nonlinear models (Carroll, Ruppert, Stefanski and Crainiceanu, 2006).

Another simple method that is also potentially applicable to any regression model, but is more computationally intensive than regression calibration, is the simulation extrapolation approach (SIMEX). For SIMEX, the bias function can be expressed in closed form for linear regression, and thus having more data leads to better estimates. However, when the bias function is not known and has to be approximated, having more data does not improve this approximation. We will discuss the SIMEX method in detail in the next chapter.

Although regression calibration and SIMEX are quite general methods for eliminating or reducing measurement error bias, they result in estimators that are consistent only in important special cases such as linear regression (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). In addition, when the measurement error variance is large, both regression calibration and SIMEX may not be useful in reducing the bias induced by the mismeasurement.

Score function methods are almost as widely applicable, but result in fully consistent estimators more generally. In these methods consistency of the estimators is due to the fact that they are M-estimators whose score functions are unbiased even if there is measurement error. The conditional-score method (Stefanski and Carroll, 1987; Nakamura, 1990) and the corrected-score method (Stefanski, 1989) differ in their underlying assumptions and ease of computation. When the assumptions for both are satisfied, the conditional-score method will usually be more efficient than the corrected-score method. However, sometimes (for example in Poisson regression) the conditional-score estimator requires numerical integration, while the corrected score has a closed form expression. Both score function methods are not straightforward to implement and the problem of multiple roots is often serious (Hossain, 2007).

SIMEX and regression calibration are simple to implement because they are "add-on" packages to existing software, whereas other methods such as the score function approaches require a completely different analysis set-up.

1.1.2 Structural Methods

To use structural models, the best possible specifications of the measurement $(X^*|Y, X, Z)$, outcome $(Y|X, Z)$, and exposure $(X|Z)$ models are needed. Bayesian methods fall into this category, as they are based on a likelihood approach followed by direct analytic calculations if possible, or otherwise by estimating methods such as Markov chain Monte Carlo (MCMC) sampling techniques. We will discuss such approaches further in this thesis. Alternatively, one could use a likelihood estimation, which also requires fully specified measurement, outcome, and exposure models. In this case, the Expectation-Maximization algorithm could be applied instead of the Bayesian MCMC techniques.

Bayesian methods have the advantage that they incorporate parameter uncertainty, whereas frequentist approaches are less able to incorporate such uncertainty. However, they require fully specified exposure models, while functional approaches, such as SIMEX, do not. This eliminates the risk of misspecification in the use of functional approaches.

Chapter 2

Simulation Extrapolation

The simulation extrapolation (SIMEX) method was first proposed by Cook and Stefanski (1994) and is a simulation-based means of estimating and reducing bias due to additive measurement error (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). It can be used for inference for parametric measurement error models, when the measurement error variance is known or can be estimated (Cook and Stefanski, 1994). Additional measurement error is added to the already mismeasured covariate and the corresponding parameter estimates are calculated. The relationship between the parameter estimate and the amount of added measurement error is calculated and extrapolated back to the case of no measurement error. Cook and Stefanski (1994) show that this approach is asymptotically equivalent to the method-of-moments estimation in linear measurement error models. The authors make a very simplified comparison between SIMEX and method-of-moments estimation using Monte Carlo-derived estimating equations. SIMEX is intuitive, simple to implement, and can be applied to just about any regression model, as it creates new datasets with added measurement error. SIMEX is applicable to general estimation methods, such as least-squares, maximum likelihood, or quasilikelihood, and can also be extended to nonadditive models (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). One way to deal with multiplicative measurement error would be to convert it to additive error by taking logarithms, as discussed in the last chapter.

To implement this method we use the simex package for covariate measurement error in R, written by Lederer and Küchenhoff (2008).

We assume a simple linear regression of the form $Y = \beta_0 + \beta_1 X + \epsilon$ with additive and nondifferential measurement error $X^* = X + U$, where $U \perp (Y, X)$, $E(U) = 0$, and $Var(U) = \sigma_u^2$. Typically, although normality of the measurement error is assumed, it is not completely critical in practice. It is also assumed that the measurement error variance, σ_u^2 is known or well estimated. Now, if $\sigma_u^2 > 0$, then the ordinary least squares estimate of β_x from regressing Y on X^*, denoted by $\widehat{\beta}_{x,naive}$, is biased. Note that SIMEX is usually not used for simple linear regression, as simple method-of moments bias corrections can be used. We add further measurement error to the already noisy predictor and create a new data set in which the measurement error variance is greater. Suppose that this process is repeated until we have a total of M data sets, including the naive one. Then each of these data sets has a successively larger measurement error variance, $(1+\zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < ... < \zeta_M$ are set by the user. $\widehat{\beta}_{x,m}$, the least squares estimate of the slope of the m^{th} data set, $m = 1, 2, ..., M$, consistently estimates $\beta_x \sigma_x^2 / \{\sigma_x^2 + (1 + \zeta_m)\sigma_u^2\}$. If we regress $\widehat{\beta}_{x,m}$ on ζ_m, we get the following mean function:

$$E(\widehat{\beta}_{x,m} | \zeta) = G(\zeta) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1+\zeta)\sigma_u^2}, \quad \zeta \geq 0.$$

We can extrapolate back to $\zeta = -1$ to get to the case of no measurement error. From the equation above it is evident that this is the case because $G(-1) = \beta_x$.

We now describe these steps in detail (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). First, simulated data sets with increasingly larger measurement error variance have to be created. Define

$$X^*_{m,i}(\zeta) = X^*_i + \sqrt{\zeta} U_{m,i}, \quad i = 1, ..., n, \ m = 1, ..., M,$$

where $\{U_{m,i}\}_{i=1}^n \sim N(0, \sigma^2_u)$ are mutually independent and identically distributed. They are also independent of all observed data. Because the error variance in the simulated data has been inflated by a multiplicative factor of $(1+\zeta)$, the error variance is zero, in theory, when $\zeta = -1$. Mathematically $Var\{X^*_{m,i}(\zeta)|X_i\} = (1+\zeta)\sigma^2_u = (1+\zeta)Var(X^*_i|X_i)$, which is equal to 0 when $\zeta = -1$, though we cannot actually generate $X^*_{m,i}(\zeta)$ for negative values of ζ.

Now that the data sets have been simulated, the naive estimates based on these predictors have to be computed. Let $\widehat{\beta}_m(\zeta)$ be the estimator when the m^{th} additional error is used and then the average of these estimators is

$$\widehat{\beta}(\zeta) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\beta}_m(\zeta).$$

$\widehat{\beta}(\zeta)$ is the mean from many simulations with the same measurement error. This way, we can precisely estimate the additional bias due to increasing measurement error.

In the extrapolation step $\widehat{\beta}(\zeta)$ is modeled as a function of $\zeta > 0$ and then the fitted models are extrapolated back to $\zeta = -1$. This extrapolated value is denoted by $\widehat{\beta}_{simex}$.

The choice of the extrapolation function is important, as the wrong choice could give misleading results and therefore not help in correcting the bias caused by measurement error. The simex function in R requires the user to specify an extrapolation function, otherwise the default, the quadratic function, is used. We apply the quadratic extrapolation function because it is numerically stable and seems to work best in our scenarios.

The SIMEX method is very popular because it is intuitive and simple to implement. We note, however, that this method does not completely correct the bias in practice, as the functional form of the bias is generally unknown and needs to be approximated. SIMEX works well for popular special cases, but does not always produce consistent estimators in general.

Chapter 3

Bayesian Estimation

Bayesian inference implemented with Markov chain Monte Carlo (MCMC) algorithms are mismeasurement-correcting methods that have become increasingly popular with the growing availability of MCMC algorithms.

3.1 Bayes Rule

Assume we have an independent and identically distributed (iid) sample $w = (w_1, ..., w_n)$ from a density f_θ, where $\theta \in \Theta$ is an unknown parameter. Then the likelihood function is

$$f(\theta|w) = \prod_{i=1}^{n} f_\theta(w_i).$$

Note that in the Bayesian framework θ is assumed to be random. This is the major difference between Bayesian and frequentist methods, where the parameters are assumed to be fixed and probabilities refer to limiting frequencies (Doucet, 2007; Casella and Berger, 2002). Because θ is assumed to be random, we can set a prior distribution $\pi(\theta)$ on it that expresses our belief about the parameter before having seen the data. Given these distributions and using Bayes' theorem we can construct a posterior distribution

$$\pi(\theta|w) = \frac{f(\theta|w)\pi(\theta)}{\int f(\theta|w)\pi(\theta)d\theta}.$$

Note that this implies $\pi(\theta|w) \propto f(\theta|w)\pi(\theta)$, where the proportionality constant can be obtained by normalization. The evaluation of this normalizing constant becomes an issue when θ is very high dimensional and cannot be evaluated in closed form. For example, suppose we want to compare the estimators, $\hat{\theta}$, to θ by using the quadratic loss function $\|\theta - \hat{\theta}\|^2$. In the Bayesian framework, one would calculate the expected value of θ under the posterior (Marin and Robert, 2007):

$$\hat{\theta} = \int \theta\pi(\theta|w)d\theta = \frac{\int \theta f(\theta|w)\pi(\theta)d\theta}{\int f(\theta|w)\pi(\theta)d\theta}.$$

Markov chain Monte Carlo (MCMC) techniques or other numerical methods are needed to complete such calculations. MCMC methods have, since they were introduced around 1990, revolutionized Bayesian statistics (Doucet, 2007). The posterior distribution is an updated version of the prior combined with the observed data, as represented by the likelihood. We can then consider the posterior distribution the new prior and repeat the procedure to get another posterior. MCMC generates multiple dependent samples by running a Markov chain a set number of times, whereas sequential Monte Carlo (SMC) generates multiple independent samples simultaneously.

3.2 Choice of Prior Distributions

The prior distribution $\pi(\theta)$ expresses the subject researcher's knowledge and belief about the parameter(s) of interest based on previous experience and subject area knowledge, but without having actually seen the current data. When such information is not available, the impact of the prior on the inference must be minimized. A "flat" distribution, called noninformative, is assigned to such priors, therefore not preferring any values. This process is not trivial, as there is not one agreed upon notion of "flat" (Gustafson, 2004; Marin and Robert, 2007).

3.3 Markov Chain Monte Carlo Algorithms

In Bayesian problems the posterior distributions are often very complicated, therefore it is difficult to compute probabilities. The integrals that have to be evaluated are frequently of high dimensions so that numerical integration techniques become useless. These problems have made Bayesian analysis very limited until around 1990 when MCMC algorithms were first applied to statistical analyses. The MCMC techniques are based on the idea that it is sufficient to produce a Markov chain $\mathbf{t}_n$, where n is a natural number, whose stationary distribution is f, for instance, the posterior distribution (Marin and Robert, 2007). If the marginal distribution of $\mathbf{t}_n$ is f, then f is also the marginal distribution of $\mathbf{t}_{n+1}$ (Grimmett and Stirzaker, 2001). For large n, $\mathbf{t}_n$ is approximately distributed from f.

The following three sections briefly discuss the very general Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Hastings, 1970), the random walk Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953), and the Gibbs sampler (Casella and George, 1992). Marin and Robert (2007), among many other sources, give a good overview of these and additional algorithms.

3.3.1 The Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is a general iterative method to sample from any probability distribution function $f(\theta)$ known up to a normalizing constant. The first step in the MH algorithm is to set a proposal probability distribution $p(\theta'|\theta)$. Based on the current state of the Markov chain, a new candidate for θ' is proposed as follows at the i^{th} iteration, $i \geq 1$:

1. Set $\theta^{(0)}$. This can be done randomly or deterministically.

2. Sample $\theta^* \sim p(\theta|\theta^{(i-1)})$

3. Compute

$$\alpha(\theta^{(i-1)}, \theta^*) = min \left(1, \frac{f(\theta^*)p(\theta^{(i-1)}|\theta^*)}{f(\theta^{(i-1)})p(\theta^*|\theta^{(i-1)})}\right).$$

4. With probability $\alpha(\theta^{(i-1)}, \theta^*)$, set $\theta^{(i)} = \theta^*$. Otherwise set $\theta^{(i)} = \theta^{(i-1)}$.

3.3.2 Random Walk Metropolis Algorithm

The original Metropolis algorithm incorporates a random walk proposal and in this thesis we use the following algorithm if the model in question is a logistic regression. A new candidate for θ is proposed as follows at the i^{th} iteration, $i \geq 1$:

1. Skip this step if $i \neq 1$. Compute the maximum likelihood estimate $\hat{\theta}$ and the covariance matrix $\hat{\Sigma}$ corresponding to the asymptotic (Fisher) covariance of θ. Set $\theta^{(0)} = \hat{\theta}$.

2. Generate $\theta^* \sim N_k\left(\theta^{(i-1)}, s^2\hat{\Sigma}\right)$, where s^2 is the scale factor. Typically, use $\hat{\Sigma} = I$, where I is the identity matrix.

3. Compute
$$\alpha(\theta^{(i-1)}, \theta^*) = min\left(1, \frac{f(\theta^*)}{f(\theta^{(i-1)})}\right)$$

4. With probability $\alpha(\theta^{(i-1)}, \theta^*)$, set $\theta^{(i)} = \theta^*$. Otherwise set $\theta^{(i)} = \theta^{(i-1)}$.

3.3.3 The Gibbs Sampler

The Gibbs sampler is a special case of the MH algorithm with an acceptance probability of one, when full conditional distributions are available. It is a simple and popular iterative method to sample from high dimensional probability distributions. Suppose again we want to sample from $f(\theta)$, where $\theta = (\theta_1, ..., \theta_p)$. Then the algorithm of the Gibbs sampler to generate a Markov chain at the i^{th} iteration $(\theta_1^{(i)}, ..., \theta_p^{(i)})$ is the following:

1. Set $(\theta_1^{(0)}, ..., \theta_p^{(0)})$. This can be done randomly or deterministically.

2. For $j = 1, ..., p$, sample $\theta_j^{(i)} \sim f(\theta_j | \theta_{\text{-}j}^{(i)})$,
   where $\theta_{\text{-}j}^{(i)} = (\theta_1^{(i)}, ..., \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, ..., \theta_p^{(i-1)})$

These MCMC algorithms are simple and general algorithms to sample from any target distribution, however, the technical details are usually very tedious. We will apply the Gibbs sampler and the random walk Metropolis algorithm. These will be discussed in more detail in the simulation application in the next chapter and the Framingham example in Chapter 6, respectively.

Chapter 4

Simulation Study

In this example we consider the case of a linear regression involving two covariates such that $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, where ϵ is normally distributed with mean zero. We suppose that only a mismeasured surrogate X^* is available for X, but we observe Z and Y correctly. For simplicity, we assume the variables are distributed normally and the measurement model is nondifferential. Then

$$X^*|X, Z, Y \sim N\left(X, \tau^2\right),$$

where τ^2 is an unknown parameter. The response model, parameterized by β_0, β_1, β_2, and σ^2, is

$$Y|X, Z \sim N\left(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2\right).$$

The exposure model, conditioned on the covariate Z and parameterized by α_0, α_1, and λ^2, is

$$X|Z \sim N\left(\alpha_0 + \alpha_1 Z, \lambda^2\right).$$

We consider a study with n=250 subjects, where X^*, Z, and Y are observed for all subjects. For a validation sample of m = 10 randomly chosen subjects, X is also measured. This is applicable for situations where a gold standard test does exist, but where it is not feasible to administer it to all subjects due to high costs or time constraints. Let the subscript C denote the complete data available for the validation sample, and R the incomplete or reduced data. For instance, X^*_C denotes the incorrectly measured variable for the subjects in the validation sample.

Following Gustafson (2004), we simulate a data set with (X, Z) having a bivariate normal distribution, each marginal distribution having a standard normal distribution, and a correlation coefficient of 0.75. This set-up yields $(\alpha_0, \alpha_1) = (0, 0.75)$ and $\lambda^2 = 0.4375$. We initially set $\tau = 0.5$ (Scenario 1), and increase it to $\tau = 1.5$ (Scenario 2) later. We set $(\beta_0, \beta_1, \beta_2) = (0, 0.5, 0.25)$ and $\sigma = 1$. Firstly, we analyze this hypothetical study using the frequentist SIMEX approach, secondly from a Bayesian perspective, and finally compare the results from both methods.

4.1 Simulation Extrapolation Approach

We use the simex package in R (Lederer and Küchenhoff, 2008) for this analysis. The simex() function in this package takes an estimate of τ as input, therefore it needs to be estimated from the validation data:

$$\hat{\tau} = \sqrt{\frac{\sum_{i=1}^{m} (x_{ic} - x^*_{ic})^2}{m}}.$$

We assume this value to be correct for its use in the SIMEX algorithm. This may be too strong an assumption as our validation sample size is small and we will investigate the validity of this further later. We use the quadratic extrapolation function because it seems to perform best in the case of this set-up. This function is usually used when the relationship is not known, but we use it because of its numerical stability even though we know the relationship.

The additional data sets with successively larger measurement error are computed as described in Chapter 2. The extrapolation for this example is illustrated in Figure 4.1 for Scenario 1 with moderate measurement error variance, and in Figure 4.2 for Scenario 2 with large measurement error variance. We will later elaborate on these two scenarios, and the performance of SIMEX in each case. Figures 4.3 and 4.4 show how well the SIMEX approach works in Scenario 1 and Scenario 2, respectively. SIMEX seems to perform very well in the case of moderate measurement error variance, but has its shortcomings when the measurement error variance is increased to 1.5 from 0.5.

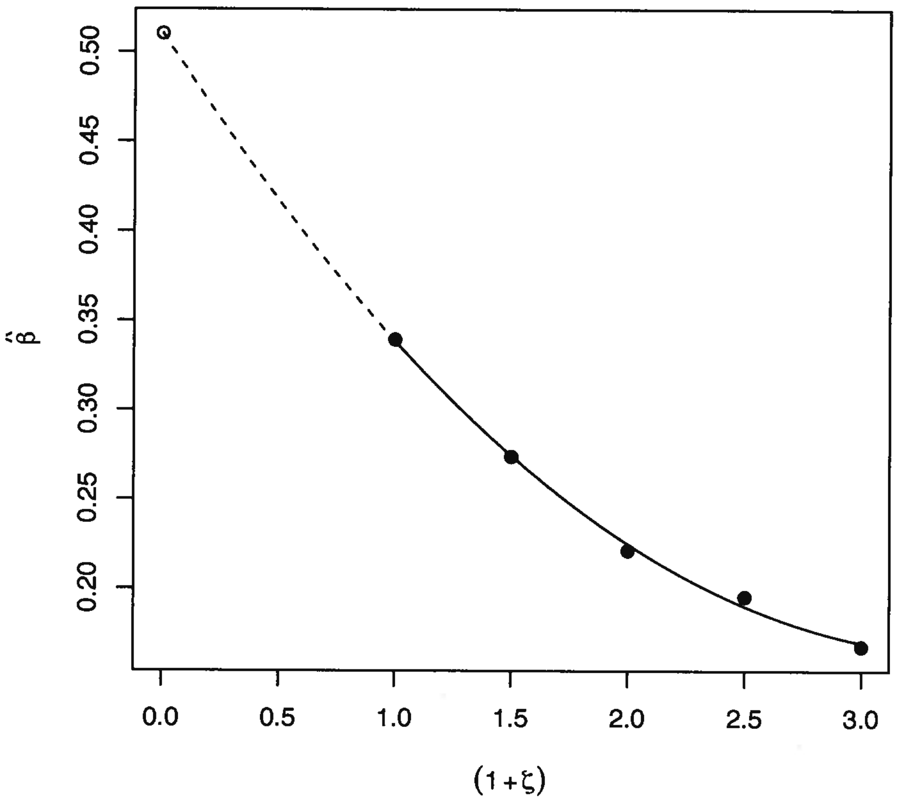

Figure 4.1: *SIMEX extrapolation plot for Scenario 1 where* $\tau = 0.5$, *illustrating the effect of adding consecutively more measurement error to the original naive estimate at* $\zeta = 0$. *The SIMEX estimate is the extrapolation to* $\zeta = -1$.

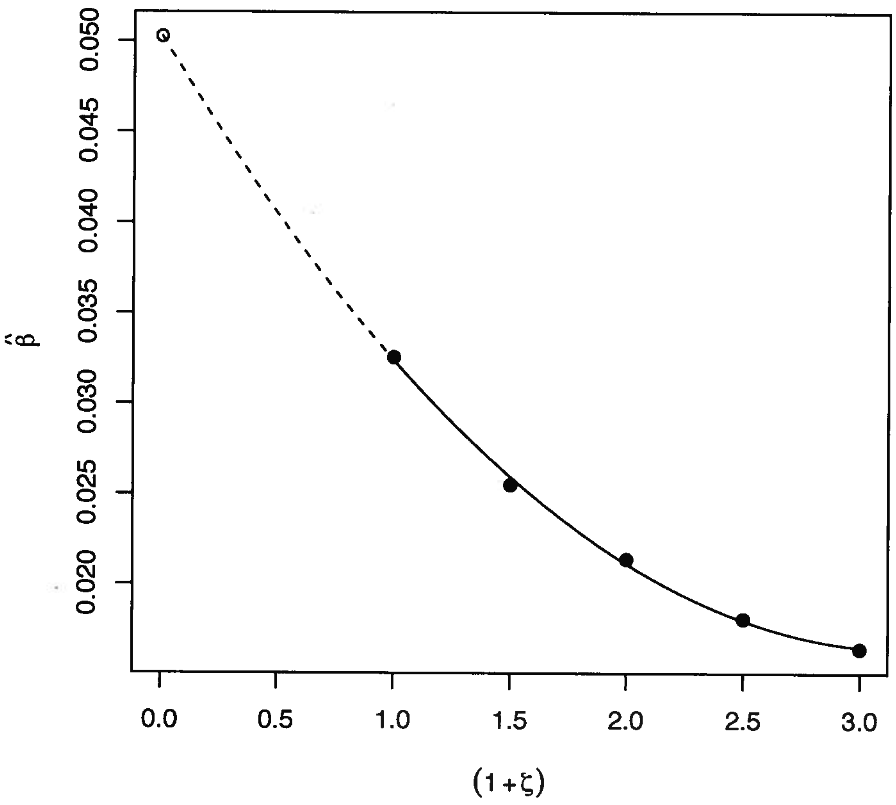

Figure 4.2: *SIMEX extrapolation plot for Scenario 2 where τ = 1.5, illustrating the effect of adding consecutively more measurement error to the original naive estimate at $\zeta = 0$. The SIMEX estimate is the extrapolation to $\zeta = -1$.*

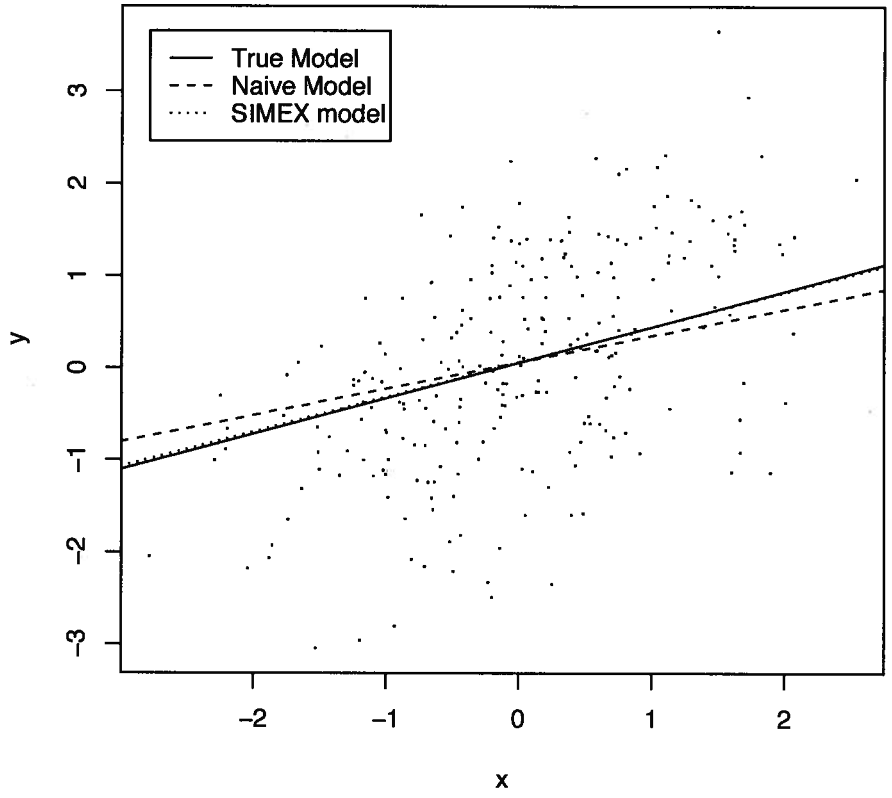

Figure 4.3: *In the case of Scenario 1 (τ = 0.5), the solid line shows the least-squares line had the data been measured correctly, while the dashed line illustrates the case of the naive analysis where measurement error is not taken into account. The dotted line illustrates the correction the SIMEX algorithm provides for this particular example.*

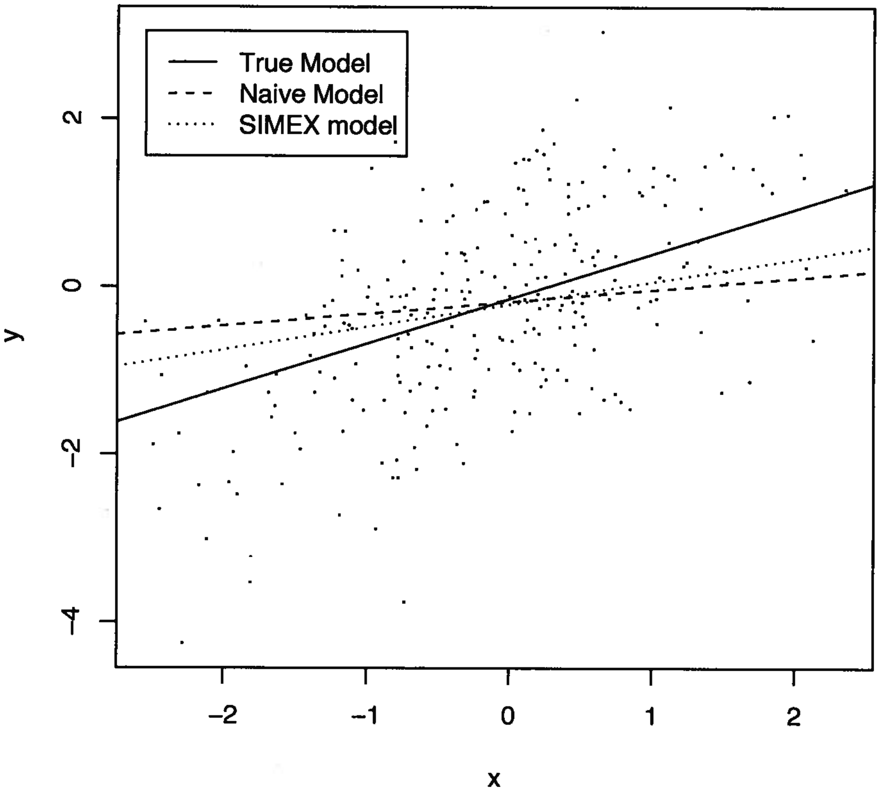

Figure 4.4: *In the case of Scenario 2 (τ = 1.5), the solid line shows the least-squares line had the data been measured correctly, while the dashed line illustrates the case of the naive analysis where measurement error is not taken into account. The dotted line illustrates the correction the SIMEX algorithm provides for this particular example.*

To make the SIMEX approach more comparable to the Bayes-MCMC method, it seems sensible to add uncertainty to the estimate of $Var(X^*|X) = \tau^2$. We achieve this by using the bootstrap method where the pairs (X^*, X)

are sampled and the estimate of τ^2 is obtained using this bootstrap sample. This process is repeated 1000 times and 95% confidence limits are obtained for τ. We then perform the SIMEX analysis three times; using the estimated τ, as well as the lower and upper bounds of the 95% CI. For the β estimates, it is possible to get "added-variance confidence intervals" in addition to the regular confidence intervals. We obtain the former as follows: We take the 2.5^{th} and the 97.5^{th} percentiles of the simex β estimates using the lower and upper bound of τ, respectively. These "added-variance confidence intervals" are wider than the regular confidence intervals on the estimated coefficients.

4.2 Bayes-MCMC Approach

We assume prior independence of all unknown parameters, so

$$f(\alpha, \beta, \lambda^2, \sigma^2, \tau^2) = f(\alpha)f(\beta)f(\lambda^2)f(\sigma^2)f(\tau^2).$$

We assign improper locally uniform priors to the α's and β's such that $f(\alpha) \sim 1$ and $f(\beta) \sim 1$, although alternatively we could use proper, but very flat, priors. We use such uninformative priors because they do not favour any one value for the coefficients. Partly due to their conjugate property, Inverse Gamma distributions seem sensible to use as priors for the variances of normal distributions (Gustafson, 2004; Doucet, 2007). So we assign $IG(0.5, 0.5)$ priors to λ^2, σ^2, and τ^2.

The posterior distribution is

$$\begin{aligned} f(x_R, \beta, \alpha, \tau^2, \sigma^2, \lambda^2 | x^*, x_C, y, z) \propto & \prod_{i=1}^{n} f(x_i^* | x_i, \tau^2) \times \\ & \prod_{i=1}^{n} f(y_i | x_i, z_i, \beta, \sigma^2) \times \\ & \prod_{i=1}^{n} f(x_i | z_i, \alpha, \lambda^2) \times \\ & f(\tau^2, \sigma^2, \lambda^2), \end{aligned}$$

which expands to

$$\begin{aligned} f(x_R, \beta, \alpha, \tau^2, \sigma^2, \lambda^2 | x^*, x_C, y, z) \propto & \left(\frac{1}{\tau^2}\right)^{n/2} \exp\left(\frac{-\sum_{i=1}^{n}(x_i^* - x_i)^2}{2\tau^2}\right) \times \\ & \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(\frac{-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i)^2}{2\sigma^2}\right) \times \\ & \left(\frac{1}{\lambda^2}\right)^{n/2} \exp\left(\frac{-\sum_{i=1}^{n}(x_i - \alpha_0 - \alpha_1 z_i)^2}{2\lambda^2}\right) \times \\ & \left(\frac{1}{\tau^2}\right)^{(0.5+1)} \exp\left(\frac{-0.5}{\tau^2}\right) \times \\ & \left(\frac{1}{\sigma^2}\right)^{(0.5+1)} \exp\left(\frac{-0.5}{\sigma^2}\right) \times \\ & \left(\frac{1}{\lambda^2}\right)^{(0.5+1)} \exp\left(\frac{-0.5}{\lambda^2}\right). \end{aligned}$$

From this posterior we can get the full conditional distributions as follows. Note that the superscript C denotes the complement.

$$\alpha | \alpha^C \sim N\left((A'A)^{-1} A' x, \lambda^2 (A'A)^{-1}\right),$$

$$\beta|\beta^C \sim N\left((B'B)^{-1}B'y, \sigma^2(B'B)^{-1}\right),$$

where $A = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix}$ and $B = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{pmatrix}$.

We also have:

$$f\left(\tau^2|\tau^{2C}\right) \propto \left(\frac{1}{\tau^2}\right)^{\frac{(n+1)}{2}+1} \exp\left(-\frac{1}{\tau^2}\left(\frac{\sum_{i=1}^{n}(x_i^* - x_i)^2 + 1}{2}\right)\right),$$

which implies that

$$\tau^2|\tau^{2C} \sim IG\left(\frac{n+1}{2}, \frac{\|x^* - x\|^2 + 1}{2}\right).$$

The full conditional distributions for the other two variance components can be obtained in a very similar fashion:

$$\lambda^2|\lambda^{2C} \sim IG\left(\frac{n+1}{2}, \frac{\|x - A\alpha\|^2 + 1}{2}\right),$$

$$\sigma^2|\sigma^{2C} \sim IG\left(\frac{n+1}{2}, \frac{\|y - B\beta\|^2 + 1}{2}\right).$$

Only considering the cases where no gold standard measurement is made, the full conditional distribution for x_R follows a normal distribution with the following mean and variance:

$$E(x_i|x_i^C) = \frac{(1/\tau^2)x_i^* + (\beta_1^2/\sigma^2)\left((y - \beta_0 - \beta_2 z_i)/\beta_1\right) + (1/\lambda^2)\left(\alpha_0 + \alpha_1 z_i\right)}{(1/\tau^2) + (\beta_1^2/\sigma^2) + (1/\lambda^2)},$$

$$Var(x_i|x_i^C) = \frac{1}{(1/\tau^2) + (\beta_1^2/\sigma^2) + (1/\lambda^2)}.$$

We are now ready to use the Gibbs sampler and to do so, we need to continually sample from the full conditional distributions above. We use 10,000 iterations after 1000 burn-in iterations. The traceplots and posterior densities of β and τ in the case of Scenarios 1 and 2 are shown in Figures 4.5 and 4.6, respectively. From this we see that there are no mixing problems.

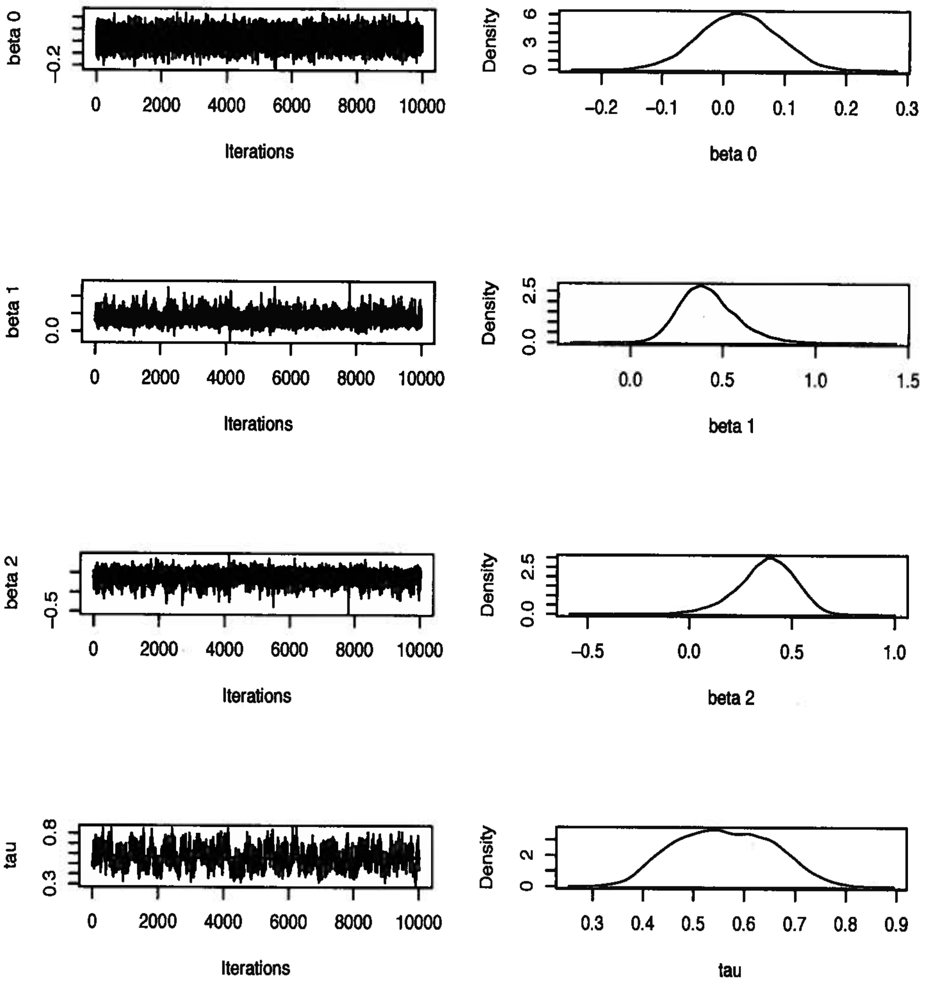

Figure 4.5: *Traceplots and posterior densities of β and τ from the Gibbs sampler for Scenario 1, where the measurement error variance is moderate. The traceplots show the 10,000 iterations after the 1000 burn-in iterations.*

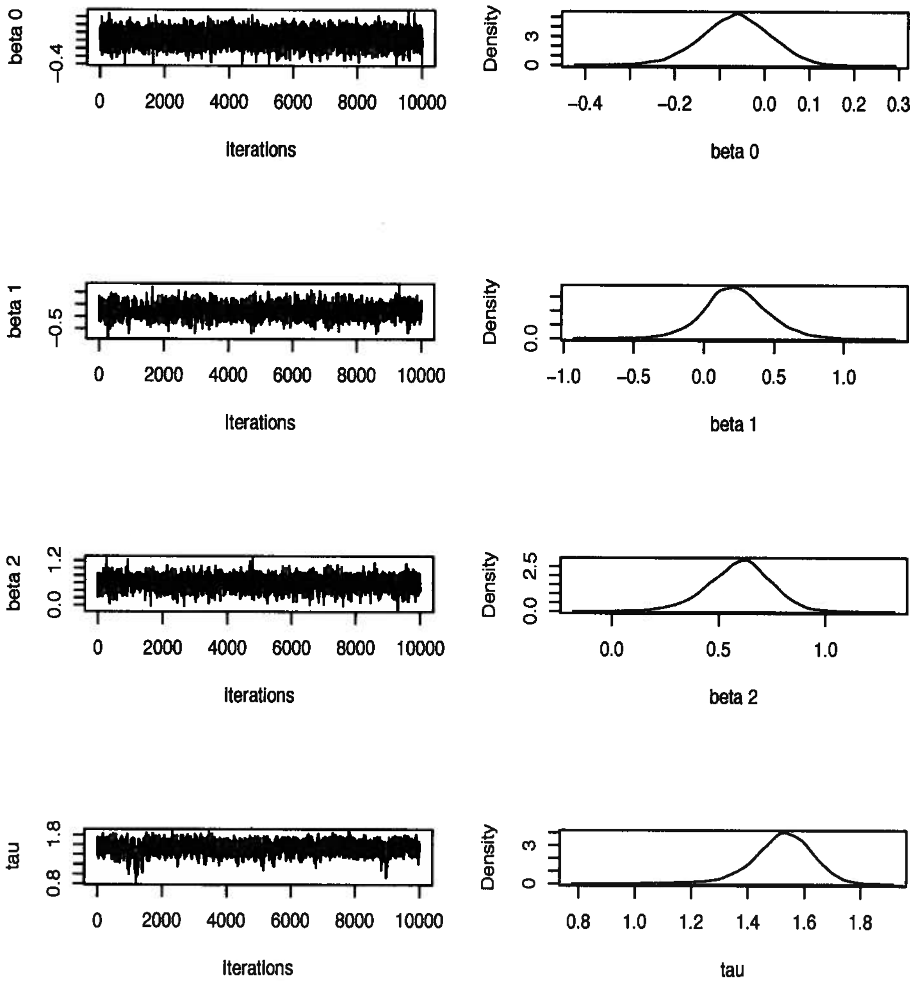

Figure 4.6: *Traceplots and posterior densities of β and τ from the Gibbs sampler for Scenario 2, where the measurement error variance is large. The traceplots show the 10,000 iterations after the 1000 burn-in iterations.*

4.3 Comparison: SIMEX versus Bayes-MCMC

It makes sense to compare the performances of the Bayes-MCMC and the SIMEX approaches by looking at the mean squared errors of the relevant parameters using both methods. For the Bayesian analysis, both the mean and the median of the 10,000 MCMC β_1 estimates were used for comparison. Table 4.1 shows the MSEs of $\hat{\tau}$ and $\hat{\beta}_1$ based on one hundred simulated data sets.

Table 4.1: *Mean Squared Errors of SIMEX and Bayesian estimates of τ and β_1 with varying validation sample size (m) and true τ. We use both, the mean and the median, in the Bayesian case. The MSEs are based on 100 comparisons between the SIMEX and Bayesian methods.*

	MSE $\hat{\tau}$			MSE $\hat{\beta}_1$		
	simex	Bayes median	Bayes mean	simex	Bayes median	Bayes mean
$m = 10$ $\tau_{true} = 0.5$	0.0089	0.0054	0.0050	0.0175	0.0369	0.0436
$m = 10$ $\tau_{true} = 1.5$	0.0802	0.0083	0.0090	0.1311	0.0718	0.0700
$m = 50$ $\tau_{true} = 0.5$	0.0025	0.0020	0.0020	0.0159	0.0200	0.0203
$m = 50$ $\tau_{true} = 1.5$	0.0223	0.0067	0.0067	0.1280	0.0303	0.0298

When the measurement error and the validation sample are relatively small, SIMEX performs better than the Bayesian method in estimating the β_1 coefficient. However, the Bayes-MCMC method outperforms the SIMEX algorithm in estimating τ in all performed scenarios. As soon as the measurement error variance is increased, the Bayes-MCMC method works better than SIMEX in the estimation of β_1. SIMEX seems to deal badly with large measurement error variance ($\tau^2 = 1.5^2$), producing a MSE over three times larger than the Bayes method in the τ estimation and four times the MSE of the Bayes method in the β_1 estimation, even when the validation sample size is large ($m = 50$). When the validation sample size, m, is ten and the measurement error variance is 1.5^2, SIMEX produces a MSE of nine times that of the Bayes-MCMC correction for τ and almost twice the MSE of the Bayesian approach for β_1. This confirms the intuition that SIMEX should be vulnerable in the large measurement error case. In summary, SIMEX seems to perform well in simple scenarios with reasonably low measurement error variance and the Bayes-MCMC method is more robust to changes in validation size and measurement error variance.

The average widths, across the one hundred runs, of the 95% confidence (SIMEX) and credible (Bayes-MCMC) intervals of $\hat{\tau}$ and $\hat{\beta}_1$ are shown in Table 4.2. The added-variance (bootstrap) confidence bounds are also provided. The Bayes-MCMC CIs for $\hat{\tau}$ are more narrow than the SIMEX ones obtained via bootstrapping. Since is expected that the Bayesian approach creates wider CIs than the SIMEX method for $\hat{\beta}_1$, the values we get from the simulation study are reassuring. It is notable that the added-variance SIMEX confidence bounds are, although wider than the regular SIMEX CIs, constantly more narrow than the Bayesian ones.

Table 4.2: *Average widths of Confidence and Credible Intervals (CI) using SIMEX and Bayesian CIs of $\hat{\tau}$ and $\hat{\beta}_1$ with varying validation sample size (m) and true τ. The averages are based on 100 simulated data sets.*

	$\hat{\tau}$ Average CI Width		$\hat{\beta}_1$ Average CI Width		
	simex boot	Bayes	simex	simex boot	Bayes
$m = 10$ $\tau_{true} = 0.5$	0.3802	0.2922	0.4010	0.5322	0.7742
$m = 10$ $\tau_{true} = 1.5$	1.1406	0.4189	0.2362	0.2823	1.0734
$m = 50$ $\tau_{true} = 0.5$	0.1846	0.1755	0.3978	0.4735	0.4921
$m = 50$ $\tau_{true} = 1.5$	0.5537	0.3143	0.2368	0.2649	0.6669

Another way to compare the SIMEX and the Bayes-MCMC approaches is to count how many times the confidence intervals and the credible intervals of $\hat{\tau}$ and $\hat{\beta}_1$ contain the true values. We use 95% CIs mentioned above for this measure of coverage. Table 4.3 gives these values out of 100 runs for the SIMEX, the added-variance (bootstrap) SIMEX, and the Bayes-MCMC corrections. The SIMEX (bootstrap) confidence bands included the true τ 84 times when the validation sample size was 10 and 94 times when the validation sample size increased to 50, regardless of the true τ. When the validation sample size is large, the SIMEX and the Bayesian approaches work about equally well in estimating τ, while the Bayesian approach seems to be the better choice when the validation sample size is small. However, when it comes to the coverage of the true β_1 by the CIs, it becomes evident that the Bayes-MCMC method performs much better, especially when the measurement error variance is large. If the measurement error variance is small, it may be more economical to use SIMEX and perhaps account for more variance by bootstrapping, than the time and cost consuming Bayesian approach.

Table 4.3: *Confidence and Credible Interval (CI) coverage of the true values using SIMEX and Bayesian CIs of $\hat{\tau}$ and $\hat{\beta}_1$ with varying validation sample size (m) and true τ. The values denote how many times, out of the 100 runs, the CIs contained the true values. We use the SIMEX and Bayesian methods, as well as the SIMEX-bootstrap method.*

	$\hat{\tau}$ CI Coverage (100)		$\hat{\beta}_1$ CI Coverage (100)		
	simex boot	Bayes	simex	simex boot	Bayes
$m = 10$ $\tau_{true} = 0.5$	84	97	83	96	96
$m = 10$ $\tau_{true} = 1.5$	84	99	0	1	97
$m = 50$ $\tau_{true} = 0.5$	94	93	89	96	90
$m = 50$ $\tau_{true} = 1.5$	94	95	0	0	95

From this study, it seems evident that the Bayes-MCMC method generally produces more accurate results than the SIMEX approach. When the measurement error variance is large, SIMEX does a poor job in correcting for the mismeasurement. Bayesian approaches are more expensive than conventional methods, due to computational difficulties as well as the scarcity of pre-packaged functions, and may not be worth using if the measurement error variance is small with a simple study set-up.

Chapter 5

Exploring the Bayes-MCMC Correction

It is well known that the more one knows about the size of the measurement error, the better the results of the techniques mentioned in this thesis. The estimate of the measurement error can stem from a validation sample, another instrument, or subject area knowledge. But how correctly do we have to estimate this measurement error in order to improve our analysis? Moreover, how certain do we have to be that this estimate is correct? In a Bayesian framework the first question refers to the location of a chosen prior, and the second to the prior's width. It seems intuitive that there is a trade-off on both scales. If our estimate is correct, a very narrow prior should be best suited. What if our estimate is wrong and we choose a narrow prior distribution, or our estimate is correct but we have little faith in it and assign it a flat prior? Is there a threshold where we end up with better results by not correcting for measurement error at all? We investigate these points further using the simulation example set-up from Chapter 4, but now assume that we have no validation data and thus $n = 250$ incomplete observations.

The question about how much one needs to know about the measurement error in a given problem could be approached from many directions. We choose to use five different inverse gamma prior distributions, with scale parameter α and shape β, for τ^2. All these priors have different widths but the same mode $= \beta/(\alpha + 1) = 0.25$, and are shown in Figure 5.1 below.

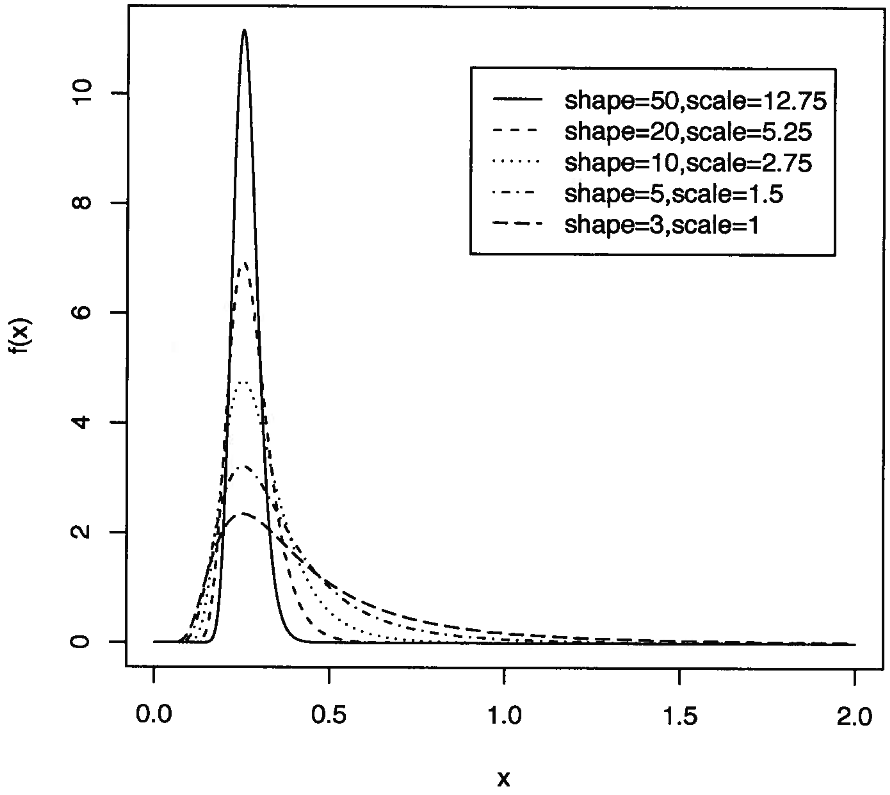

Figure 5.1: *Inverse Gamma ($f(x)=IG(\alpha,\beta)$) plots with mode=0.25, where α is the shape and β is the scale of the distribution.*

We place the, realistically unknown, true value of τ^2 at several different percentiles of these prior distributions. The idea is to investigate what effect arises from placing these priors on τ^2 when the true value of the measurement error variance is located towards the center or either tail of each distribution. We examine what happens when we estimate τ^2 to be around

0.25, at the prior's mode, but the true value is at a particular percentile of the distribution. The width of the prior indicates how certain we are about the estimate.

For instance, if we estimate the measurement error variance to be around 0.25 and are certain, due to prior knowledge, that it is larger than 0.15 but no larger than 0.45, say, we would choose the $IG(50, 12.75)$ prior out of the five distributions mentioned. This example is illustrated in Figure 5.2.

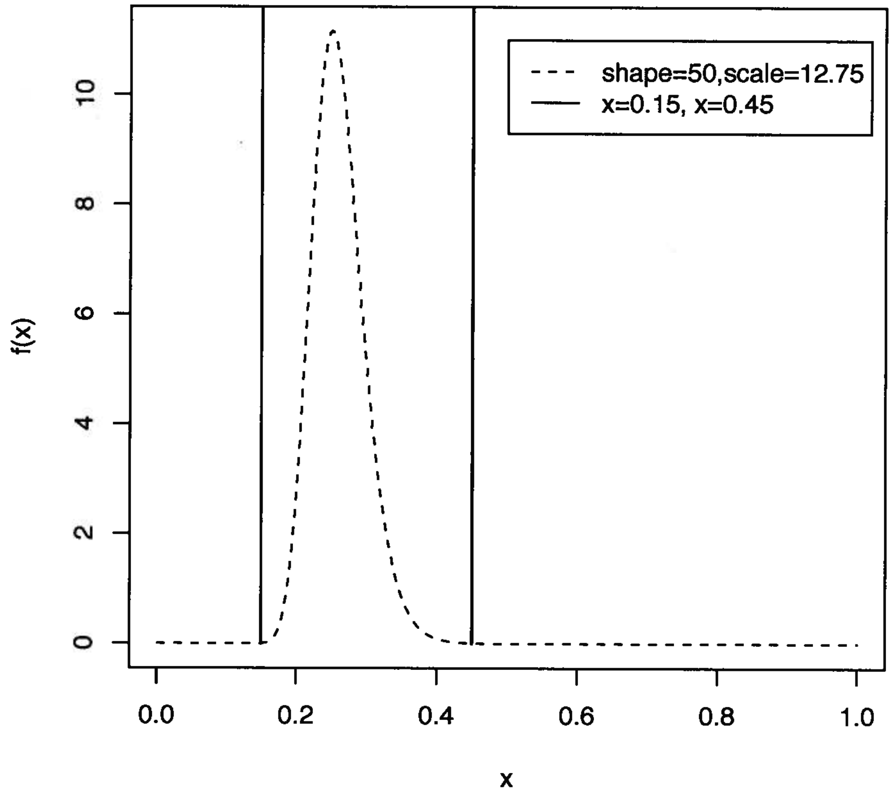

Figure 5.2: *Inverse Gamma (f(x)=IG(*50, 12.75*)) plot with mode=0.25, and a rough approximation of where its support is greater than zero.*

Table 5.1 below shows the mean square errors for the β_1 parameter, calculated by performing the analysis with different inverse gamma priors on τ^2 with the true τ^2 values being located at several different percentiles of these prior distributions. For each prior, the mean square errors are calculated in the case of the naive analysis, where no measurement error correction is performed, and in the case of the Bayesian measurement error correction described above. It follows that when the naive mean square error is larger than the corrected one, then the measurement error should be taken into account. However, if the naive mean square error is smaller than the corrected one, it would be counter-productive to try to correct for measurement error, as we get a similar or better result by performing the cheaper and by far simpler naive analysis. To see this relationship better, the ratio=naiveMSE/correctedMSE is given for each case. When this ratio is greater than one, a correction for measurement error is needed. Of course, this is not a matter of absolute certainty, as different problems will have different thresholds, and there is always some "gray area" around the threshold.

Table 5.1: *For the five prior distributions below, the mode is 0.25. There are eight cases for each model: τ^2_{true} is at the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, or 95^{th} percentile; or at the mode of the distribution. Mean squared errors of the β_1 estimates are based on 50 runs. For each model, the naive MSE estimates (ignoring measurement error), the corrected MSE estimates, and their ratios are given.*

Prior		5^{th}	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}	95^{th}	mode
		MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE
$IG(50, 12.75)$	naive	0.0338	0.0362	0.0361	0.0416	0.0447	0.0495	0.0521	0.0401
	corrected	0.0333	0.0285	0.0228	0.0201	0.0160	0.0173	0.0181	0.0204
	ratio	1.0163	1.2728	1.5870	2.0690	2.7917	2.8572	2.8787	1.9697
$IG(20, 5.25)$	naive	0.0270	0.0294	0.0335	0.0395	0.0453	0.0529	0.0598	0.0379
	corrected	0.0741	0.0542	0.0480	0.0260	0.0205	0.0168	0.0207	0.0319
	ratio	0.3638	0.5430	0.6966	1.5215	2.2106	3.1408	2.8914	1.1886
$IG(10, 2.75)$	naive	0.0257	0.0280	0.0363	0.0446	0.0540	0.0664	0.0760	0.0378
	corrected	0.1397	0.1197	0.0803	0.0456	0.0243	0.0184	0.0246	0.0633
	ratio	0.1838	0.2341	0.4521	0.9779	2.2194	3.6096	3.0899	0.5965
$IG(5, 1.5)$	naive	0.0246	0.0296	0.0375	0.0512	0.0697	0.0907	0.1056	0.0390
	corrected	0.2074	0.1783	0.1302	0.0630	0.0324	0.0269	0.0346	0.1097
	ratio	0.1186	0.1658	0.2884	0.8117	2.1494	3.3761	3.3761	0.3551
$IG(3, 1)$	naive	0.0240	0.0284	0.0399	0.0857	0.0857	0.1147	0.1386	0.0367
	corrected	0.2556	0.2384	0.1641	0.0859	0.0393	0.0303	0.0425	0.1739
	ratio	0.0938	0.1191	0.2430	0.9983	2.1786	3.7915	3.2613	0.2112

Figure 5.2 shows the ratios from Table 5.1 graphically. The true τ^2 values and the ratios are represented on the horizontal and vertical axes, respectively. Each line illustrates the behaviour of the ratios of five prior distributions across the different scenarios. When the line of the chosen prior distributions in the figure are well above one, a measurement error correction is needed. However, if it is below one, it may be useless to perform the measurement error correction because of lack of prior knowledge. We note that this "rule" is not black and white. When the ratio is close to one, one may want to investigate measurement error correction further, taking into account subject area knowledge, cost of a potential correction, and value. Priors much wider than the ones shown lead to greater biases in the β parameter estimation than the naive analysis.

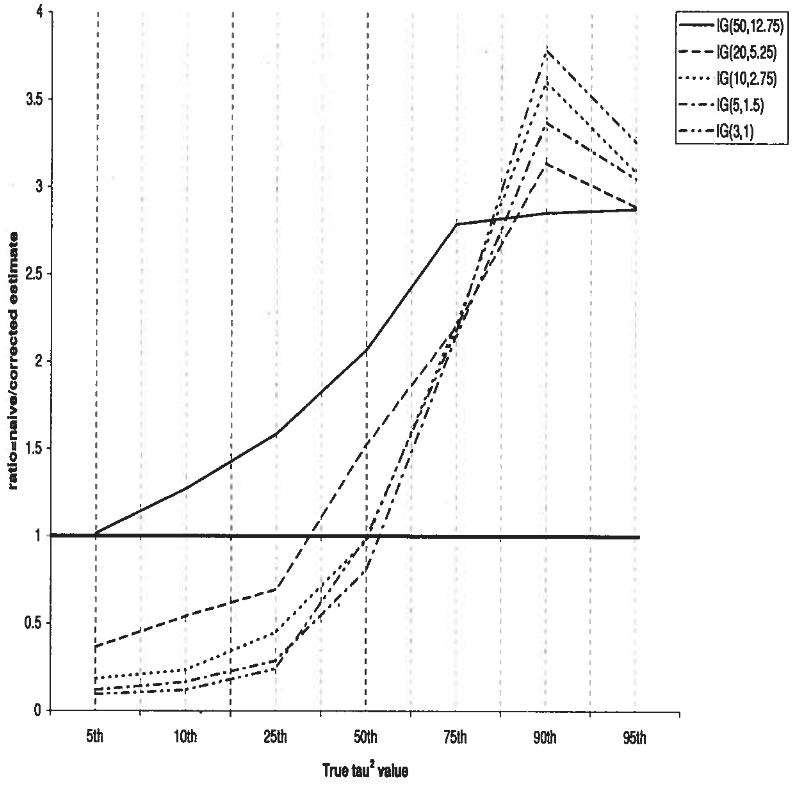

Figure 5.3: *The ratios of the naive over the corrected mean square errors from Table 5.1 are shown graphically.*

Chapter 6

Framingham Study Example

To illustrate the use of SIMEX and the Bayes-MCMC methods in a real-world application, we use a subset of the Framingham Heart Study data. This still ongoing, large cohort study was started in 1948 under the direction of the National Heart, Lung, and Blood Institute in the United States with the objective to identify some of the characteristics that contribute to cardiovascular disease. We use only complete data of male adults aged between 31 and 65 at the first exam, which yields a subset of $n = 1615$ subjects (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). The study consists of a series of medical exams, about two years apart, where a number of variables were recorded. The response is the indicator variable Y, taking the value 1 if the individual has developed a coronary heart disease within eight years of the third exam, and 0 otherwise. It has become well known, primarily through this study, that high blood pressure is one of the leading causes of cardiovascular diseases. We assume that the systolic blood pressure (SBP) measurements may be mismeasured and will try to account for the bias caused by this. We adopt a transformation of SBP introduced by Carroll, Ruppert, Stefanski and Crainiceanu (2006), setting $X^* = \log(SBP - 50)$. We do not have any validation data available, but SBP is measured at the second and third exams. We treat these measurements as replicates for our practical purposes, even though technically they are not. Therefore, we use the mean, for simplicity denoted by X_i^*, for each individual i. The nondifferential measurement error assumption seems plausible because we would like to be able to measure long-term systolic blood pressure (X_i), but what we observe in reality is blood pressure on a single day (X_i^*). It is possible that this actual measurement indicates little information about the long-term systolic blood pressure (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). The model is $X_i^* = X_i + U_i$, where $E(U_i) = 0$ and $Var(U_i) = \tau^2$. We also have the correctly recorded age, Z, at the second exam for each subject.

Since we are dealing with a binary response, we choose to use the logit link on Y to make sure its domain covers the real numbers. Thus the response model is

$$\text{logit}(P(Y=1|X,Z)) = \log\left(\frac{P(Y=1|X,Z)}{1-P(Y=1|X,Z)}\right) = \beta_0 + \beta_1 X + \beta_2 Z.$$

The measurement and exposure models are defined as in the simulation study above. Under the nondifferentiality assumption, the measurement model is

$$X^*|X, Z, Y \sim N(X, \tau^2).$$

The exposure model is

$$X|Z \sim N(\alpha_0 + \alpha_1 Z, \lambda^2).$$

We will use this set-up to conduct a SIMEX and a Bayes-MCMC analysis.

6.1 SIMEX Analysis

As before with the simulated dataset, we use the simex package in R and apply the quadratic fitting method because of its numerical stability. The replicate measurements allow us to estimate the components of variance estimator V as follows:

$$\hat{V} = \sum_{i=1}^{n} \sum_{j=1}^{2} (X_{ij}^{*} - \bar{X}_{i.}^{*})^2 = 0.01278,$$

where $X_{i.}^{*}$ is the mean of the replicates (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). We thus estimate the measurement error variance $\hat{V}/2 = \hat{\tau}^2 = 0.00639$ from the data as well as $Var(X^{*}) = 0.04543$. We use the glm function in R for the logistic regression model. The results from the SIMEX analysis are shown in section 6.2 and the extrapolation plot appears in Figure 6.1.

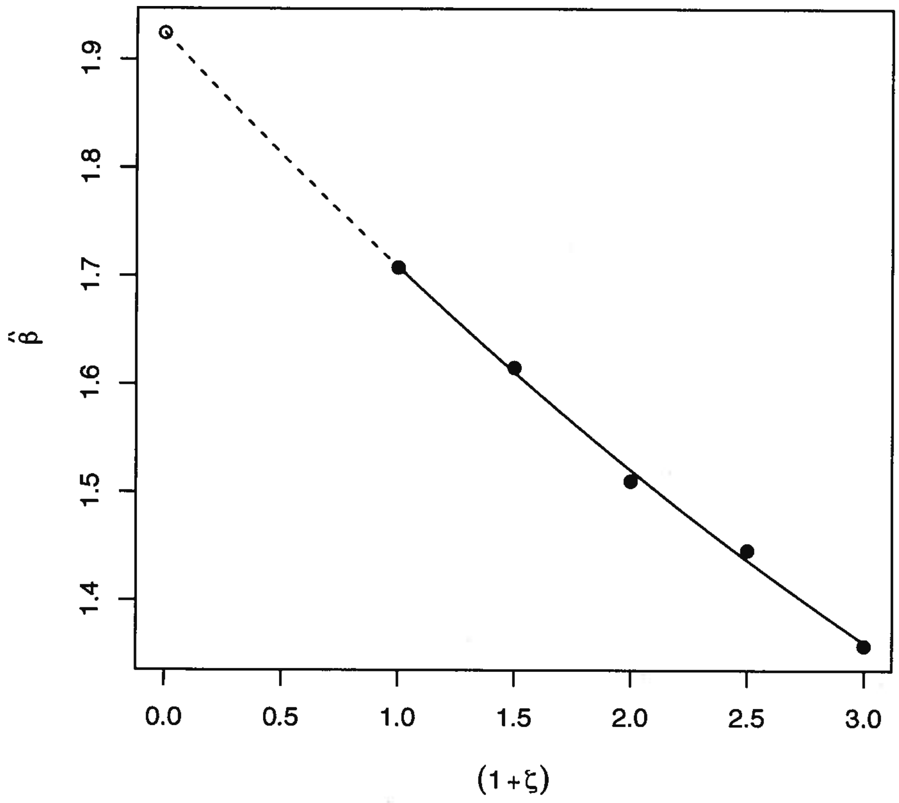

Figure 6.1: *The extrapolation performed by the simex function is shown for the mismeasured systolic blood pressure in the Framingham data.*

6.2 Bayes-MCMC Analysis

Unlike in the SIMEX analysis, we can use the replicate measurements to incorporate the uncertainty about τ^2 in the Bayesian framework. As earlier in the simulation example, we assign improper locally uniform priors to the αs and βs such that $f(\alpha) \sim 1$ and $f(\beta) \sim 1$, and $IG(0.5, 0.5)$ priors to λ^2 and τ^2. We first perform the analysis by coding the complete MCMC algorithm in R, and secondly using WinBUGS software as part of the R program.

6.2.1 Analysis Using R

The posterior distribution is

$$\begin{aligned} f(x, \alpha, \beta, \lambda^2, \tau^2 | x^*, y, z) \propto & \prod_{i=1}^{n} f(x_i^* | x_i, \tau^2) \times \\ & \prod_{i=1}^{n} f(y_i | x_i, z_i, \beta) \times \\ & \prod_{i=1}^{n} f(x_i | z_i, \alpha, \lambda^2) \times \\ & f(\alpha, \beta, \lambda^2, \tau^2), \end{aligned}$$

which expands to

$$
\begin{aligned}
f(x,\alpha,\beta,\lambda^2,\tau^2|x^*,y,z) \propto {} & \left(\frac{1}{\tau^2}\right)^n \prod_{i=1}^{n} \exp\left(\frac{-1}{2\tau^2}\left((x^*_{i1}-x_i)^2+(x^*_{i2}-x_i)^2\right)\right) \times \\
& \prod_{i=1}^{n} \frac{\exp\left(y_i(\beta_0+\beta_1 x_i+\beta_2 z_i)\right)}{1+\exp\left(\beta_0+\beta_1 x_i+\beta_2 z_i\right)} \times \\
& \left(\frac{1}{\lambda^2}\right)^{n/2} \prod_{i=1}^{n} \exp\left(\frac{-1}{2}\frac{(x_i-\alpha_0-\alpha_1 z_i)^2}{\lambda^2}\right) \times \\
& \left(\frac{1}{\lambda^2}\right)^{(0.5+1)} \exp\left(\frac{-0.5}{\lambda^2}\right) \times \\
& \left(\frac{1}{\tau^2}\right)^{(0.5+1)} \exp\left(\frac{-0.5}{\tau^2}\right).
\end{aligned}
$$

We cannot obtain full conditional distributions for x and β, and thus require the use of the random-walk MH algorithm to update β. We use the algorithm for logistic regression given in 3.3.2 with the scale $s = 0.5$. We use the glm function in R, such that `model=summary(glm(y ~ x* + z, family=binomial))`. Then `model$coef` provides the maximum likelihood estimates $\hat{\beta}$, while `model$cov.unscaled` gives $\hat{\Sigma}$. The other parameters are updated via the Gibbs sampler as in the simulation study in Chapter 4.

Figure 6.2 shows that there are no apparent mixing problems using 300,000 iterations after 5,000 burn-in iterations. The results are presented and discussed in the next section.

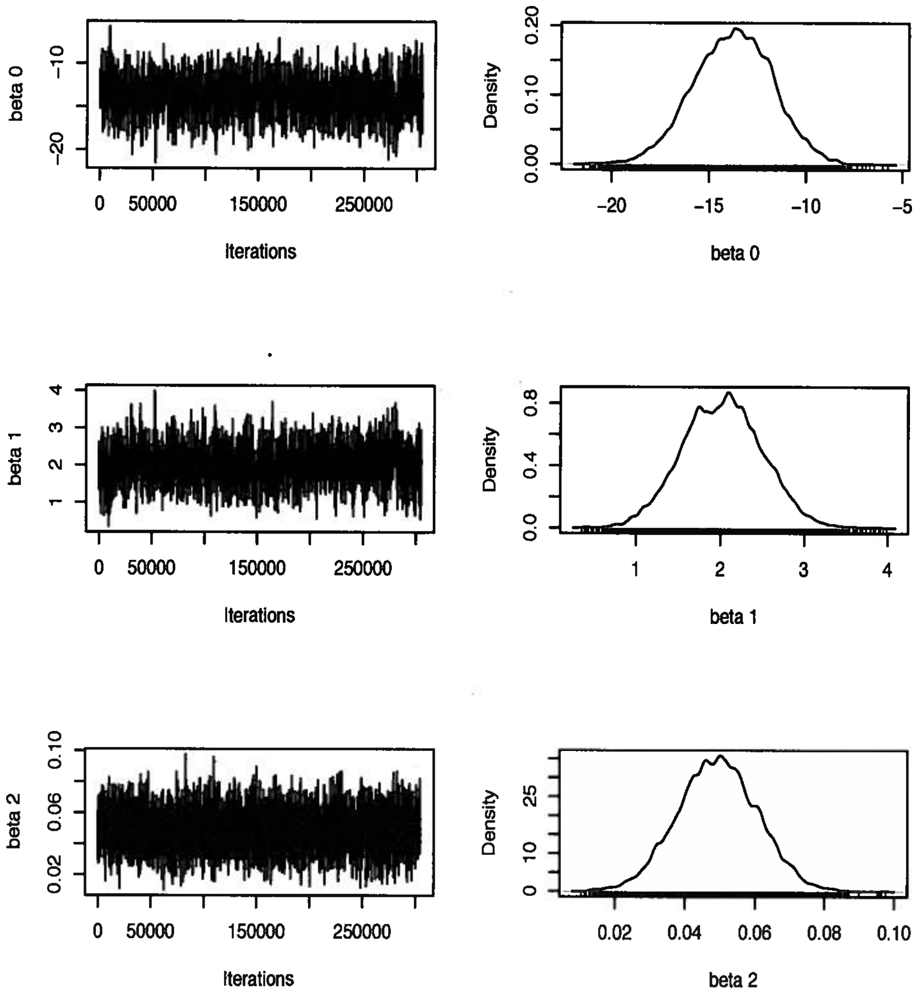

Figure 6.2: *Traceplots and posterior density plots of 300,000 iterations after 5,000 burn-in iterations of the random-walk Metropolis sampler of the three β parameters using R.*

6.2.2 Analysis Using WinBUGS

In this second analysis we use WinBUGS (Bayesian Inference Using Gibbs Sampling for Windows) for the execution of the MCMC steps. WinBUGS is a computational tool for MCMC that elimiates much of the hard work that is involved in the coding of MCMC algorithms for the user (Haneuse, 2008). This way, we need not worry about the random walk Metropolis Hastings algorithm and thus need not specify a tuning parameter, and such. Full conditional distributions for the Gibbs sampler updates are also not required. We only provide the likelihood, the priors, and initial values. If the user does not provide initial values, WinBUGS will assign them.

Although WinBUGS can be used by itself, we set up the problem in R and run WinBUGS using the R2WinBUGS package. The bugs function in this package allows us to call WinBUGS from R and then do any further analyses of the posterior in R. Traceplots and posterior densities are shown in Figure 6.3, and it is evident that there is no need for concern.

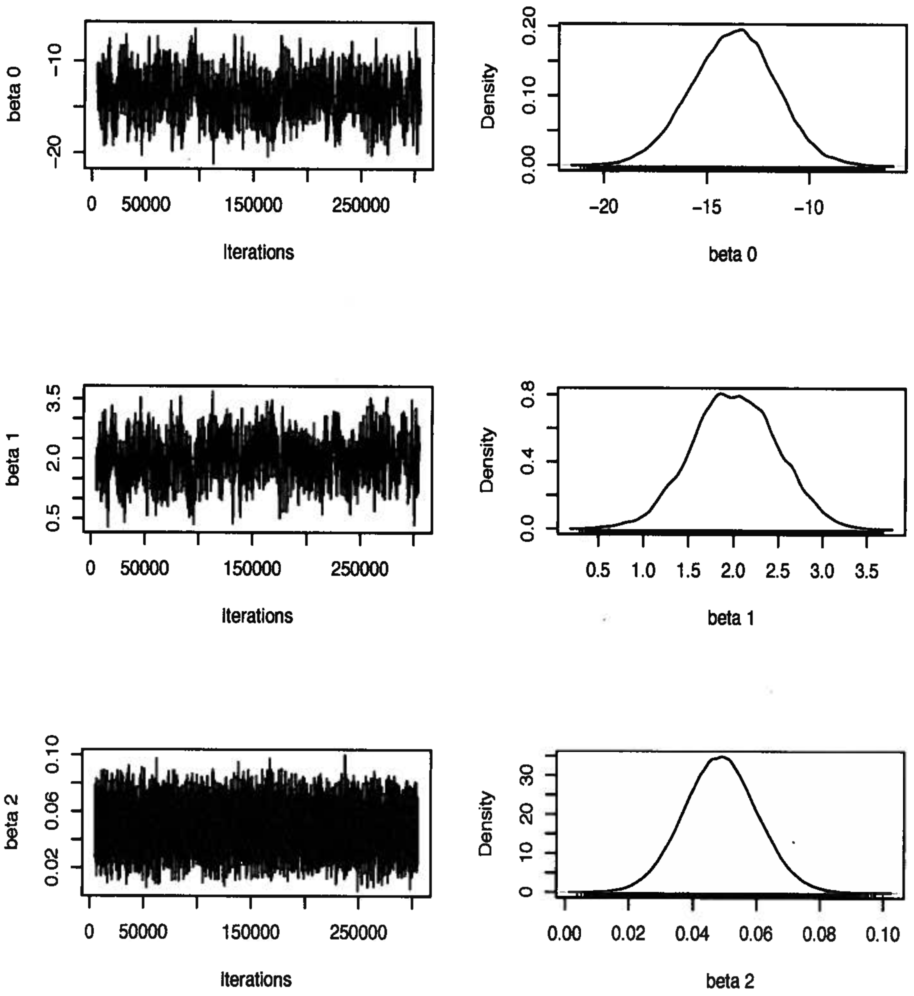

Figure 6.3: *Traceplots and posterior density plots of 300,000 iterations after 5,000 burn-in iterations of the WinBUGS output of the three β parameters.*

6.3 Results

The results for the β estimates from the SIMEX and Bayesian analyses are recorded in Table 6.1 below. (Note that the mean of the 300,000 β realizations is shown in the Bayesian results.)

Table 6.1: *β parameter estimates with 95% confidence, or credible, intervals using the naive, SIMEX, and Bayesian analyses.*

Parameter		Estimate	95% CI	CI width
β_0	naive	-12.4066	(-15.8394, -8.9737)	6.8657
	SIMEX	-13.1492	(-16.9464, -9.3520)	7.5944
	Bayes (R)	-13.7247	(-17.8030, -9.8861)	7.9169
	WinBUGS	-13.6909	(-17.7200, -9.7190)	8.001
β_1	naive	1.7080	(0.9079, 2.5080)	1.6001
	SIMEX	1.8926	(0.9961, 2.7890)	1.7929
	Bayes (R)	2.0295	(1.1083, 2.9485)	1.8402
	WinBUGS	2.0184	(1.0930, 2.9510)	1.8580
β_2	naive	0.0507	(0.0282, 0.0733)	0.0451
	SIMEX	0.0493	(0.0264, 0.0721)	0.0457
	Bayes (R)	0.0490	(0.0263, 0.0720)	0.0457
	WinBUGS	0.0488	(0.0262, 0.0720)	0.0458

If we choose the Bayesian analysis using R, the resulting model for explaining the probability of whether or not an individual has developed coronary heart disease within eight years of the third exam is:

$$\log\left(\frac{P(Y=1|X^*,Z)}{1-P(Y=1|X^*,Z)}\right) = -13.72 + 2.03X + 0.05Z$$

$$P(Y=1|X,Z) = \frac{\exp(-13.72 + 2.03X + 0.05Z)}{1+\exp(-13.72 + 2.03X + 0.05Z)}.$$

In light of the simulation study in Chapter 4, the results for this subset of the Framingham data behave as we would expect. The Bayesian analyses produce wider credible intervals than the SIMEX or the naive analyses. Recall that we estimated a value for τ^2 and then used it as if it were known in the SIMEX analysis, whereas we treated it as an unknown parameter in both Bayesian analyses, thus accounting for more variability. The naive analysis produces the most narrow confidence intervals, thus being too confident. It is comforting that the results of both Bayesian analyses are very close. SIMEX seems to give improved results over the naive analysis, with its parameter estimates moving towards the corresponding Bayesian estimates.

Chapter 7

Conclusion and Future Work

Measurement error, when ignored, can have serious effects on statistical analyses, such as biased parameter estimation, loss of power, and masking of the features of the data. Nevertheless, the possible presence of measurement error often fails to be investigated outside of academic research. Many methods have been proposed to deal with measurement error, and they can be broadly grouped into functional and structural models (Carroll, Ruppert, Stefanski and Crainiceanu, 2006). Simple "black-box" functional methods, such as simulation extrapolation (SIMEX) and regression calibration, have been introduced to make the solving of such problems more accessible. However, such approaches are rarely appropriate when the problem is complicated or the measurement error is large. We show that SIMEX, a simulation-based method for which the measurement error variance has to be known or well estimated, fails when the measurement error variance is large.

Bayesian measurement error adjustment, a structural method, allows for more accurate bias correction more generally, while integrating more variability. The risk with structural methods is that of exposure model misspecification. Ignoring any philosophical issues associated with Bayesian methods for the sake of this discussion, the only other drawback concerning the Bayesian framework is that the posterior distributions are often very complicated and numerical integration techniques become useless. Computationally intensive sampling methods such as Markov chain Monte Carlo algorithms often take a long time to code as well as implement (Robert and Casella, 2004).

In Chapter 4 we show that when the measurement error variance and the validation sample are relatively low, SIMEX performs better than the Bayes-MCMC method in estimating the coefficient of X^*. If either the measurement error variance, the validation sample, or both are increased, the Bayes-MCMC correction outperforms SIMEX. The Bayesian method is robust to changes in validation sample or measurement error variance sizes. SIMEX incorporates less variance in its results than the Bayesian approach, and thus produces more narrow confidence intervals even when forced to take into account more variance by using the bootstrap method. The Bayesian credible intervals contain the true values more often than the SIMEX confidence intervals. However, when the measurement error variance is small, it may be more economical to use SIMEX and account for more variance by bootstrapping.

To implement Bayesian methods, thorough understanding of statistics as well as statistical software is needed and thus non-statisticians often shy away from using Bayesian methods due to time or financial constraints. WinBUGS is a useful tool to use for the computational part of a Bayesian analysis. Knowledge of the Bayesian framework is still needed to use it, however, WinBUGS takes care of many complicated details. As shown in the Framingham Heart study example, the results it produces are very similar to the "custom" results. This example also illustrates another source of information when validation data is not available: replication data. The replicates are used to estimate and update the measurement error variance. Again, the Bayesian measurement error correction method produces wider credible intervals than the SIMEX analysis. However, the SIMEX analysis does seem to improve the results over the naive method as well.

The Bayesian adjustment presented in this thesis can be applied and extended to a variety of problems in many different research areas. As an obvious extension, one could consider adding more correctly measured covariates, which would be fairly straightforward when using WinBUGS. Otherwise, one could extend the current model quite simply as well.

One crucial part of Bayesian statistics is the choice of priors for the unknown parameters. In this thesis, we investigate the effects of different choices of priors for the measurement error variance, τ^2. As shown in Chapter 5, it may so happen that correcting for measurement error is of no value due to lack of knowledge (or wrong assumption) about the measurement error variance. We investigate when measurement error correction is worth the trouble. Our findings indicate that the accuracy of the measurement error estimation is less important than the width of the prior we assign to it.

Our investigation concerning these issues only considers mean squared errors, so an immediate extension would be to consider the width of credible intervals as well. Another relevant extension would be to investigate the Bayes-MCMC correction further. It would be useful to develop "rules" for when to correct for measurement error in the cases of more general models. If validation or replication data are available, incorporating such information as part of the decision of whether or not it is worth correcting for measurement error would be very useful for subject area researchers.

Bibliography

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models, a Modern Perspective*, Vol. 105 of *Monographs on Statistics and Applied Probability*, second edn, Chapman & Hall/CRC, Boca Raton.

Casella, G. and Berger, R.L. (2002). *Statistical Inference*, Vol. 6 of *Duxbury Advanced Series*, second edn, Duxbury.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler, *The American Statistician* **46**: 167-174.

Cook, J.R. and Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models, *Journal of the American Statistical Association* **89**: 1314-1328.

Doucet, A. (2007). Statistical computing and Monte Carlo methods course notes, The University of British Columbia.

Grimmett, G.R., and Stirzaker, D.A. (2001). *Probability and Random Processes*, third edn, Oxford University Press Inc., New York.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Vol. 13 of *Interdisciplinary Statistics*, Chapman & Hall/CRC, Boca Raton.

Haneuse, S. (2008). An introduction to WinBUGS, Summer school on Bayesian Modeling and Computation at the University of British Columbia. PIMS collaborative research group.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**: 97-109.

Hossain, S. (2007). *Dealing with Measurement Error in Covariates with Special Reference to Logistic Regression Model: A Flexible Parametric Approach*, Doctor of Philosophy Thesis, The University of British Columbia.

Lederer, W. and Küchenhoff, H. (2008). The simex Package. The R Project for Statistical Computing.

Marin, J-M., and Robert, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer Science+Business Media, LLC, New York.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1992). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**: 1087-1092.

Nakmura, T. (1990). Corrected score functions for errors-in-variables models: methodology and application to generalized linear models, *Biometrika* **77**: 127-137.

Pierce, D.A. and Kellerer, A.M. (2004). Adjusting for covariate errors with nonparametric assessment of the true covariate distribution, *Biometrika* **91**: 863-876.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second edn, Springer Science+Business Media, LLC, New York.

Stefanski, L.A. (1987). Unbiased estimation of a nonlinear function of a normal mean with application to measuremnt error models, *Communications in Statistics, Series A* **18**: 4335-4358.

Stefanski, L.A. and Carroll, R.J. (1987). Conditional scores and aptimal scores in generalized linear measurement error models, *Biometrika* **74**: 703-716.