On the Bayesian estimator of interaction models with measurement error and misclassification in covariates

by

© **Mahbuba Sultana**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

November 2018

St. John's, Newfoundland and Labrador, Canada

Abstract

Measurement error and misclassification in covariates are commonly arising problems in statistical models. They have negative impacts on statistical inference about the outcome, including bias and large variability in estimators. Furthermore, in a statistical model, two or more covariates can interact, which in practice is quite challenging to deal with. One of the recent techniques is Bayesian method that incorporates the prior knowledge about parameters. In this research, Bayesian techniques are applied to the models with interaction terms, while addressing measurement error and misclassification. Moreover, through extensive simulation studies, Markov Chain Monte Carlo algorithms are used to implement the Bayesian methods.

Dedicated to my parents

Lay summary

The presence of measurement errors in variables in the statistical model are common problems in practice. They can provide incorrect statistical inference and misleading conclusions. Besides, presence of interaction terms are common in many research areas. Erroneous variable incorporated with interaction term, makes the analysis more complicated. One of the recent techniques is Bayesian methods that incorporates the prior knowledge about parameters. The primary goal of this research is to monitor the behaviour of the Bayesian estimations of the model parameters in presence of measurement error for both discrete and continuous covariates, under different frameworks including Monte Carlo iteration numbers, sample size, magnitude of measurement error, prior selection. More specifically, this study pays more attention to the behaviour of the coefficient of interaction term.

We studied the Bayesian estimates for (a) Contineous covariate without measurement error, (b) Contineous covariate with measurement error, (c) Discrete covariate without measurement error and (d) Discrete covariate without measurement error models.

We observed that, generally, lowering the measurement error as well as increasing the number of iterations and sample size improved the convergence of the MC estimates. Moreover, it was difficult to capture the behavior of the coefficient of interaction term. Therefore, it required more iterations, large sample size, less amount of measurement error to perform as well as the other coefficients.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Taraneh Abarin for her excellent support and continuous motivation. She provided me funding and taught me several concepts in statistics. Without her support I could not finish this study properly.

I am grateful to the faculty and staffs at the Department of Mathematics and Statistics for their support and help. Finally, I would like to thank Memorial University of Newfoundland to give me a chance to pursue my master degree in Statistics.

Statement of contribution

I produced the original draft of the dissertation, consolidated future drafts, performed the statistical analysis, and produced tables and figures. Dr. Abarin produced R codes, supervised research, analysis and dissertation composition, provided statistical interpretation, composed a draft of the discussion, and produced an updated draft of the dissertation.

Table of contents

List of tables

List of figures

Chapter 1

Introduction and Overview

1.1 Measurement error

A most fundamental task of any statistical model is to display the relationship between response (dependent variable) and the explanatory (independent) variables. In epidemiological studies an example is the relationship between smoking status (independent or exposure variable) and heart disease (dependent variable). Sometimes, due to some unavoidable facts, an accurate measurement of the exposure variable is hard to achieve. For instance, X is the variable we are interested in which is unobserved. Instead, we observe W, which is the substitute variable for X. We define W as the surrogate variable for X, that incorporates errors (mismeasurement) in our desired model. This mismeasurement can occur in both the discrete and continuous covariates. Mismeasured continuous variable induces measurement error and categorical variable introduces misclassification in the regression models. The existence of measurement error and misclassification has been a problem in statistical analysis for years in several sectors, for example, in biology, epidemiology, econometrics which was analyzed and discussed by different authors such as Pearson (1902), Wald (1940), Berkson (1950), Fuller (2008) and Carroll et al. (2006). Various components are indeed responsible

for the erroneous measurement such as inaccuracy in the instruments, higher price of exact measurement etc. For instance, in clinical trials, different methods may generate different measurements. Moreover, sometimes researcher may go through a method that is cheaper and more convenient for observations. Therefore, all these can lead to an independent covariate that involves measurement error.

1.1.1 Impact of Measurement error

Many researchers do not consider dealing measurement error due to the lack of awareness, the unavailability of the necessary information about correcting measurement error, etc. It has long been recognized that the ignorance of the measurement error in inferential procedures may be substantial, often turning out in an unreliable conclusion with bias, large variability, incorrect inference in the estimation of parameters, reducing power of tests and inaccurate coverage probabilities of confidence intervals (Muff et al. (2015), Liao et al. (2014), Gustafson (2003)). Generally, in presence of measurement error with no additional information, the model is not identifiable. We consider a model as identifiable if the parameters in the parameter space can be estimated identically using the data. Otherwise, the non-identifiability issue arises (Gustafson (2012 and 2014)).

1.1.2 Types of Measurement error

The very initial approach of analyzing measurement error is identifying the error component properly. Several types of measurement error can be induced in the model in practice. Theoretically measurement error can be both differential and non-differential. An error whose magnitude is different for the individuals who have the outcome, for instance some disease, compared to those without the outcome can be defined as the

differential error. The non-differential measurement error is independent of the outcome variable, that is the magnitude is similar for both individuals who have and have not the outcome.

1.1.3 Measurement error models

A very preliminary step of analyzing the measurement error is to precisely identify the correct model for measurement error procedure. The two most common measurement error models are classical error model and Berkson error model.

1.1.3.1 Classical model

The classical measurement error arises in laboratory specially when an instrument is used for measurement, and the measurements vary around the true value. Consider X to be the true variable that is unobserved. Therefore, instate of X we observe W as a surrogate variable for X. That is

$$W = X + U.$$

Here, U is the measurement error which is independent of X with $E(U) = 0$ and $Var(U) = \delta^2$. In here, δ^2 is known as the measurement error. This case can be observed when measurements are disturbed by a number of uncontrollable factors and influences.

1.1.3.2 Berkson model

Another error structure is Berkson measurement error model. This arises when the average measurement value of a group of individual is assigned on each individual. When the independence assumption between U and X is often too strong, we investigate

the effect of applying the Berkson error model, which assumes

$$X = W + U.$$

Here, U is independent of W.

1.2 Methods of estimation

Measurement error in the covariate can create unavoidable issues in the inference process. Different author suggests different techniques of adjusting the measurement error in the regression model. They differ based on the assumptions on variables to be satisfied, the availability of data as well as parametric vs nonparametric. Some measurement error methods are provided in Fuller (2006), Carroll et al. (2006) and Gustafson et al. (2011). The methods that are commonly used to correct measurement error in the estimation process include - Regression calibration, simulation extrapolation (SIMEX) as well as likelihood methods including Bayesian methods.

1.2.1 Bayesian Methods

Bayesian methods recently draw the attention in statistical science, considered as an interesting alternative to the classical theory until the late 1980's (Ntzoufras et al. 2009). This method considers parameters as random variables that are characterized by a prior, and the prior is assumed to be the distribution for unknown parameters. In classical inference, data is considered as observations of random variable. This explains the main difference between Bayesian and classical inference. The Bayesian inference is based on the famous theorem called Bayes Theorem. Assume that two outcomes A

and B, then we can define the Bayes theorem as follows

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)} \propto P(B|A)P(A).$$

This equation is known as Bayes' rule. Basically in Bayesian inference, the parameter is the random variable and its distribution is known as prior distribution. We collect the sample and update our prior knowledge about the unknown parameter of interest. The updated version of prior called the posterior density, which is the conditional distribution of unobserved values (such as the parameters) given the observed data. This posterior distribution is our target and it summarizes all the information about the parameters. To calculate the posterior density, we have to find the joint density of the data and parameters and integrate out the parameters to get the marginal density of the data. We can then divide the joint density by this marginal density to get the posterior distribution. Mathematically this can be expressed as follows.

Let η is the unknown parameter which is a random variable. The distribution of η is, for instance, $p(\eta)$, which is the prior distribution. This explains the available information about the parameter we had, before observing the data, and say X . Then the posterior distribution is

$$p(\eta|x) = \frac{p(x|\eta)p(\eta)}{p(x)}.$$

As $p(\eta|x)$ does not depend on x, it can be written as

$$p(\eta|x) \propto p(x|\eta)p(\eta). \tag{1.1}$$

Here, the posterior distribution is $p(\eta|x)$ that is proportional to the multiplication of the likelihood, $p(x|\eta) = \prod_i^n p(x_i|\eta)$ and the prior distribution $p(\eta)$.

Since generally, measurement error creates the problem of non-identifiability of the parameters, which requires extra information to deal with, some researchers recommend

using more informative prior distributions (Gustafson et al. (2005) and Gustafson and McCandless (2014)). This additional information from the prior distribution is expecting to handle the non-identifiability problem. Some other researchers put some controversy with this concept as several types of priors are accessible in practice and selecting an inappropriate prior can assemble a misleading conclusion. Therefore, using validation subsample and replication that acquires extra information for solving the issue of non-identifiability is also proposed.

1.2.1.1 Prior selection

Developing prior distributions is undoubtedly the most controversial aspect of any Bayesian analysis (Lindley (1983), Walters and Ludwig (1994)). It becomes more crucial when the method has to deal with measurement error. An inappropriate choice for priors will undoubtedly distort the inference procedure (Gustafson and McCandless (2014)). Therefore, considerable care should be taken when selecting priors.
In general there are three different types of priors.

1. Non-informative prior

A prior distribution is non-informative if this does not provide any useful information about the parameter of interest. It can be defined as vague, diffuse, and uniform prior as well. Many statisticians favor non-informative priors because they appear to be more objective. However, it is unrealistic to expect that non-informative priors represent total ignorance about the parameter of interest.

2. Informative prior

An informative prior is the one which is not dominated by the likelihood and it has a significant impact on the posterior distribution. Therefore, when a prior distribution dominates the likelihood, it is clearly an informative prior. These types of distributions

must be specified with care in practice.

3. Improper prior

A prior is said to be improper if it is not a legitimate probability function. For example, a uniform prior distribution on the real line, $p(\eta) \propto 1$, for $-\infty < \eta < \infty$, is an improper prior.

1.2.2 Adjustment for measurement error

1.2.2.1 Study design

Generally, obtaining consistent estimators from a non-identifiable model in presence of measurement error is not possible Dealing with this problem demands additional information that can be acquired from the validation subsample and replicated data set (Carroll and Li (1992), Cook and Stefanski (1994) and Carroll et al. (2006)).

1. Validation subsample

In some cases, one can effectively observe the true variable of interest from a subset of the data. This is called the validation subsample. That is, this study design allows one to measures the true exposure variable.

2. Replication data

Another study design for estimating the parameters and dealing with non-identifiability is replicated data. Here the researcher can obtain independent measurements of the error-prone variable.

1.2.3 Markov Chain Monte Carlo (MCMC)

In this part, we will introduce the Markov Chain Monte Carlo (MCMC), that allows us to approximately build the posterior distribution as calculated by Bayes' Theorem. A Markov chain is a stochastic process that describes a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. More precisely, future states are independent of past states given the present state. Consider a draw of $\eta^{(t)}$ to be a state at iteration t. The next draw $\eta^{(t+1)}$ is dependent only on the current draw $\eta^{(t)}$, and not on any past draws. This satisfies the Markov property

$$p(\eta^{(t+1)}|\eta^{(1)}, \eta^{(2)}, ...\eta^{(t)}) = p(\eta^{(t+1)}|\eta^{(t)}).$$

Markov chain is a number of draws of η that are dependent on the previous one. Monte Carlo is a solution to the difficult problem of sampling from a high dimensional distribution for the purpose of numerical integration. This uses repeated random sampling to generate simulated data, similar to the experimental data to use with a mathematical model.

In context of Bayesian, our goal is producing independent draws from the posterior distribution through simulation and making summary by using those draws. The posterior distribution presented in formula (1.1), is proportional to the multiplication of likelihood and prior. As in some cases, the normalizing constant is not known, the draws from the multiplication are dependent on each other and may be treated as Markov Chain samples. If our chain satisfies some regularity conditions, then the chain will eventually converge to the stationary distribution (in our case the posterior) and we have approximate draws from the desired distribution which is the posterior distribution in here. Once we have a Markov chain that has converged to the stationary distribution, then the draws in our chain appear to be like draws from posterior and we should be able to use Monte Carlo Integration methods to find quantities of interest.

To ensure that the MC draws are less dependent, it is common in practice to omit a part of the draws by either thinning or burn-in method.

The advantage of the MCMC algorithm is that this method can deal with complicated integration in the posterior distribution. Another most important aspect is that, integrals often do not have closed-form solution and MCMC method can help to obtain the closed-form solution of the estimates (Brooks (1998), Richardson et al. (2013)).

Two MCMC algorithms have commonly been used in Bayesian for collecting draws from the posterior - the Gibbs Sampler and the Metropolis-Hastings algorithm.

1.2.3.1 Gibbs sampler

Gibbs sampling generates a Markov chain of samples. The idea is to sample from a full conditional distribution with the remaining variables fixed to their current values when we have a p.d.f. or p.m.f. that is difficult to sample from directly. We set some starting value and obtain a sequence of random values of the parameters. These satisfy the property of being a Markov chain.

Suppose we have a joint distribution that we want to sample from (for example, a posterior distribution). We can use the Gibbs sampler to sample from the joint distribution if we knew the full conditional[1] distributions for each parameter.

Let us consider taking sample from the full conditional distribution, where say we have only three parameters, $\eta = (\eta_1, \eta_2, \eta_3)$. The followings are some steps to collect Gibbs sampler:

[1]Full conditional distribution is the distribution of the parameter (η_i) conditional on the known information (x) and all the other parameters: $p(\eta_i|\eta_{-i}, x)$

1. Provide an initial value for the unknown parameters $\eta^{(0)}$.

2. Start with any η value and draw a value η_1 from the full conditional distribution $p(\eta_1|\eta_2^{(0)}, \eta_3^{(0)}, x)$.

3. Draw a value $\eta_2^{(1)}$ from the full conditional $p(\eta_2|\eta_1^{(1)}, \eta_3^{(0)}, x)$. $\eta_1^{(1)}$ is the updated value from the first iteration.

4. Draw a value $\eta_3^{(1)}$ from the full conditional $p(\eta_3|\eta_1^{(1)}, \eta_2^{(1)}, x)$ using both updated values.

5. Draw $\eta^{(2)}$ using $\eta^{(1)}$ and continue this process using the most updated values.

6. Repeat until we get M draws, with each draw being a vector.

If the draws are large enough and satisfy some regularity conditions then according to Ergodic theorem, these draws converge to the stationary distribution which is the posterior distribution in here. Gibbs sampling is applied where the full conditional distributions are obtained. Besides, when it is difficult to obtain the conditional distributions, we use Metropolis-Hastings Algorithm as the solution.

1.2.3.2 Metropolis-Hastings Algorithm

When the full conditional distribution for the unknown parameters are not available then this algorithm can be used. It can approximate the desired distribution comprised of any combination of prior probabilities and sampling models.

The Metropolis-Hastings Algorithm follows the following steps:

1. Choose a starting value $\eta^{(0)}$.

2. At iteration t, draw a candidate $\eta^{(*)}$ from a jumping distribution[2] $J_t(\eta^*|\eta^{(t-1)}, x)$.

[2]transition probability matrix

3. Compute an acceptance ratio r

$$r = \frac{p(\eta^*|x)/J_t(\eta^*|\eta^{(t-1)})}{p(\eta^{(t-1)}|x)/J_t(\eta^{(t-1)}|\eta^*)}.$$

4. Accept $\eta^{(t)}$ as $\eta^{(*)}$ if it has higher probability $min(r, 1)$. If $\eta^{(*)}$ is not accepted, then $\eta^{(t)} = \eta^{(t-1)}$. Basically, if our candidate draw has higher probability than our current draw, then our candidate is better, and we definitely accept it.

5. Repeat steps 2-4 M times to get M draws from the stationary distribution, with optional burn-in.

In this study, the Gibbs sampler algorithm has been used for collecting draws from the posterior.

1.2.4 Test statistics and diagnostic plots

Making valid inference based on all the outcomes and hypothesis testing is an essential part in any statistical analysis.

For this research purpose and inspection, the Geweke and Heidelberg-Welch diagnostic test has been performed where we check the hypothesis

$$H_0 : \text{Markov Chain is from stationary distribution.}$$

After generating a chain and calculating the test we decide whether to accept or reject the null hypothesis based on some steps of all the tests.

1.2.4.1 Geweke Test

Geweke diagnostic takes two non overlapping parts, usually the first 10% and last 50% proportions, of the Markov chain and compares the means of both parts, using the difference of means this test try to see if the two parts of the chain are from the same distribution (null hypothesis).

1.2.4.2 Heidelberg-Welch Test

The test is successively applied, firstly to the whole chain, then after discarding the first 10%, 20% of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded. If the outcome constitutes the failure of the stationarity test, it indicates that a longer MCMC run is needed.

Besides conducting the hypothesis testing, diagnosing the MCMC requires some graphical representation that assess the behaviour of the samplers with respect to the fitted parameters. Therefore, the cross correlation plots, autocorrelation plots and trace plots has generated accordingly.

Trace plots for parameters in context of Bayesian needs to perform to make sure that the prior distribution is well calibrated. It shows precisely where the chain has been exploring. If the chain is stationary it should not show long-term trend and the average value of the chain should be roughly flat and more dense around the true value. Long-term trends or drifts in the plot indicate slower convergence.

Auto correlation plot measures the auto correlations between the samples returned by our MCMC. High auto correlation is an indication of slow convergence and for this case

reparameterization can help.

The cross correlation plot calculates the cross correlation between the monitored variables for each of chains. High correlation among parameters indicates low convergence. Therefore, this may need reparameterization.

1.3 Interaction in the model

Let us consider an example, the impact of smoking and age on the weight. Both the smoking and age has separate influence on the weight. However, an adult smoker's weight differs from the young smoker's weight. That is, the impact of smoking on weight depends on the level of age, that leads the term interaction.

The presence of interactions can have important implication for the interpretation of statistical models. However, in some cases, due to the complex nature of capturing information from the interaction term, some researchers ignore this which causes the issue of model misspecification. Avoiding the interaction term in the model can question the efficiency of the inference of model. Most importantly, in presence of both the error-prone and accurately measured variables, the interaction terms can possibly become erroneous as well that is quite challenging to detect and deal with, according to Carroll et al. (2006), Gustafson et al. (2011), Richardson et al. (2002) and Buzas et al. (2014).

In order to make the inference reliable and precise, proper care should be taken for the erroneous variables. One of the possible techniques can be Bayesian method which has mentioned and discussed before, that provides prior information about the unknown parameters and can be used to handle measurement error in the independent variables and interaction terms.

1.4 Organization of the Dissertation

In most of the practical cases, information remains in some regressors (independent variables) are not completely accurate and this inaccuracy guides to the terms measurement error and misclassification. This thesis defines the general concepts of measurement error, outlines the impact of ignoring measurement error and finally addresses Bayesian method as a well known fixing procedure for model with measurement error and misclassification. In chapter 2, we study the interaction model without any error in the variables where the main objectives are to illustrate how the Bayesian method is applied and how statistical tools are used to diagnosis the convergency of the Markov Chains. In chapter 3, an error-prone continuous variable has been added to the model that interact with an accurately measured variable where we asses the performance of the MCMC estimates from validation subsamples and replication and compare their performances with the naive estimates that ignores the error in the variables. Analyzing the discrete variable without misclassification is the topic of chapter 4 where an extensive simulation study has been conducted under different scenarios. The following chapter includes a misclassified covariate in our regression model where sensitivity analysis is performed to exhibit a transparent vision about the impact of error in covariates. Finally, chapter 6 outlines the conclusion about all the analysis scheme that has done on the former chapters.

Chapter 2

Interaction model with continuous covariate without measurement error

The effects of measurement error on the estimated parameters often studied through simulation studies. In this chapter we demonstrate the Bayesian approach to estimate the parameters in the linear model with interaction. Moreover, we analyze the MCMC estimates using graphical and statistical tests introduced in Chapter 1.

Let us consider the following model, where, Y is the response variable, X is accurately measured and z is a non-random continues covariates.

$$Y|X \sim N(\beta_0 + \beta_1 X + \beta_2 z + \beta_3 X z, \delta^2).$$

Here,

$$X \sim N(\phi_x, \gamma^2).$$

Consider we have n independent subjects in our study. Using Bayesian approach we can calculate the joint density of unobserved quantities given the observed one as follows.

$$f(\eta|y,x) \propto \prod_{i=1}^{n}\{f(y_i|x_i,\tilde{\beta},\delta^2)f(x_i|\phi_x,\gamma)\}f(\tilde{\beta})f(\phi_x)f(\gamma^2)f(\delta^2), \tag{2.1}$$

where, $\tilde{\beta} = (\beta_0,\, \beta_1,\, \beta_2,\, \beta_3)$ and $\eta = (\beta_0,\, \beta_1,\, \beta_2,\, \beta_3, \phi_x, \delta^2, \gamma^2)$ is the vector of unknown parameters in the model 2.1. Consider applying improper priors for the regression coefficients β values and for ϕ_x and proper priors of Inverse Gamma distribution (IG(a,b)) where, a is shape parameter and b is scale parameter and IG(a,b) is centred at b/a, for the variance components δ^2 and γ^2.

$$f(\beta) \propto 1, \qquad f(\phi_x) \propto 1, \qquad \gamma^2 \sim IG(0.5, 0.5), \qquad \delta^2 \sim IG(0.5, 0.5).$$

Therefore, the posterior can be written as

$$\begin{aligned} f(\eta|y,x) \propto & (\frac{1}{\delta^2})^{n/2} e^{-\Sigma_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i-\beta_2 z_i-\beta_3 x_i z_i)^2 \big/ 2\delta^2} \\ & \times (\frac{1}{\gamma^2})^{n/2} e^{-\Sigma_{i=1}^{n}(x_i-\phi_x)^2 \big/ 2\gamma^2} \\ & \times (\frac{1}{\gamma^2})^{0.5+1} e^{-(0.5) \big/ 2\gamma^2} \\ & \times (\frac{1}{\delta^2})^{0.5+1} e^{-(0.5) \big/ 2\delta^2}. \end{aligned}$$

The expression of full conditional distributions of the parameters is given as a function of one individual unobserved quantity at a time given the observed amounts. Following Gustafson (2004) the notation, say m^c denotes all the unobserved quantities other than m. Let B be a matrix of order $n \times 4$ where, the i^{th} row is $(1,\, x_i, z_i, x_i z_i)$. Therefore, similar to Chapter 4 Mathematical Details in Gustafson (2004), the full conditional distributions are as follows:

$$\tilde{\beta}|\tilde{\beta}^c \sim N\{(B'B)^{-1}B'y, \delta^2(B'B)^{-1}\}$$
$$\phi_x|\phi_x^c \sim N(\bar{x}, \gamma^2/n)$$
$$\delta^2|\delta^{2c} \sim IG\{(n+1)/2, (\Sigma_{i=i}^n(y_i - B\beta)^2 + 1)/2\}$$
$$\gamma^2|\gamma^{2c} \sim IG\{(n+1)/2, (\Sigma_{i=i}^n(x_i - \phi_x)^2 + 1)/2\}.$$

To estimate the unknown parameters in the model and to investigate the limiting behaviour of the posterior distributions, extensive simulation studies have been performed under different scenarios for the impact of (a) Monte Carlo iteration number and (b) sample size which has been presented as follows.

2.1 Simulation Studies

2.1.1 Monte Carlo iteration number (M, number of draws)

In this section, a sample of 1000 subjects were iterated 50000, 100000 and 300000 times to observe the pattern of convergence to the true value of parameters with probability one. The true value for model parameters considered as $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0.5$, $\phi_x = 0, \gamma^2 = 1.0$ and $\delta^2 = 1.0$.

The following tables and graphs show the summary of the simulation results. The Estimate is the sample mean of the Monte Carlo iterations. The MSE and S.D. represents the Mean Squared Error and Standard Deviation of the samples, respectively. The empirical 95% coverage probability is calculated as the proportion of the times $\pm$ 2S.D. contains the true value of interest.

Table 2.1: MCMC summary for 50000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.547	0.105	0.004	0.745
β_1	0.5	0.612	0.108	0.007	0.721
β_2	0.5	0.417	0.056	0.003	0.712
β_3	0.5	0.013	0.050	0.002	0.632
ϕ_x	0.0	-0.032	0.042	0.001	0.504
γ^2	1.0	1.201	0.063	0.005	0.704
δ^2	1.0	0.974	0.057	0.003	0.643

Table 2.2: MCMC summary for 100000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.504	0.103	0.004	0.768
β_1	0.5	0.520	0.107	0.005	0.752
β_2	0.5	0.492	0.050	0.003	0.788
β_3	0.5	0.501	0.041	0.001	0.832
ϕ_x	0.0	-0.013	0.035	0.001	0.613
γ^2	1.0	1.042	0.063	0.005	0.902
δ^2	1.0	0.976	0.055	0.002	0.810

Table 2.3: MCMC summary for 300000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.501	0.101	0.001	0.821
β_1	0.5	0.545	0.107	0.002	0.882
β_2	0.5	0.560	0.047	0.001	0.932
β_3	0.5	0.482	0.031	0.001	0.913
ϕ_x	0.0	-0.003	0.045	0.001	0.807
γ^2	1.0	1.040	0.023	0.002	0.887
δ^2	1.0	0.985	0.041	0.003	0.802

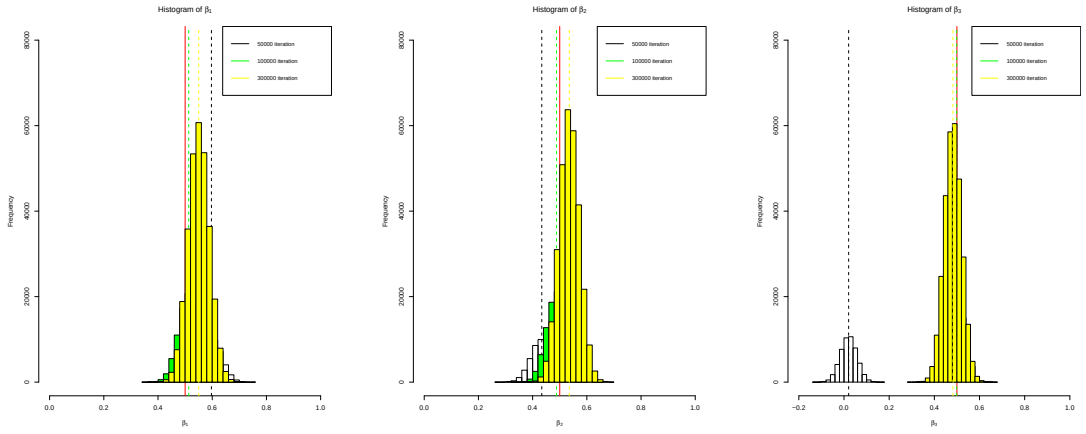

Figure 2.1: Histogram of the MCMC estimates of β_1, β_2 and β_3 with respect to different replication numbers. The vertical red solid line represents the true parameter value and the dashed black, green and yellow line indicates 50000, 100000 and 300000 iterations, respectively.

Tables 2.1, 2.2, 2.3 and Figure 2.1 represent the histograms of the estimated regression

coefficients. In the tables, we considered up to three digits after decimal point for the tabulated values. In the graphs, the x-axis indicates the parameter values and y-axis shows the frequency.

It is expected that with the increment of iterations number, the distance between the true parameters and MCMC estimated values will decrease. Considering both the mean and variability, the Bayesian estimates meet the expectations. More specifically, the MSE decreases as the number of iterations increases. Moreover, the empirical 95% coverage probability improves with the increase in iteration numbers. The results also confirms the fact that for the coefficient of the interaction term, larger number of iterations are required, in order to have a better coverage probability. It is due to the fact that the interaction term adds an extra parameter to the model, without adding extra data. That is why, generally, capturing the true value for the parameter of interaction term is more challenging.

2.1.1.1 Diagnostic tests and Plots

Tables 2.4, 2.5 and 2.6 show the Heidelberg and Welch Diagnostic test results for 50000, 100000 and 300000 iterations, respectively.

Table 2.4: Heidelberg and Welch Diagnostic test output for 50000 iteration

Variable	**Stationarity**	**p-value**
β_0	passed	0.789
β_1	passed	0.880
β_2	passed	0.559
β_3	passed	0.411
ϕ_x	passed	0.746
γ^2	passed	0.126
δ^2	passed	0.805

Table 2.5: Heidelberg and Welch Diagnostic test output for 100000 iteration

Variable	**Stationarity**	**p-value**
β_0	passed	0.586
β_1	passed	0.659
β_2	passed	0.252
β_3	passed	0.485
ϕ_x	passed	0.713
γ^2	passed	0.196
δ^2	passed	0.345

Table 2.6: Heidelberg and Welch Diagnostic test output for 300000 iteration

Variable	**Stationarity**	**p-value**
β_0	passed	0.453
β_1	passed	0.416
β_2	passed	0.568
β_3	passed	0.936
ϕ_x	passed	0.389
γ^2	passed	0.102
δ^2	passed	0.486

The results from the tables confirm that the Markov Chain converges to the stationary state for all the parameters, for all the iterations. The trace plots and autocorrelation plots for β_3 (coefficient of the interaction term) were produced to further assist the diagnosis about the convergence of the chains.

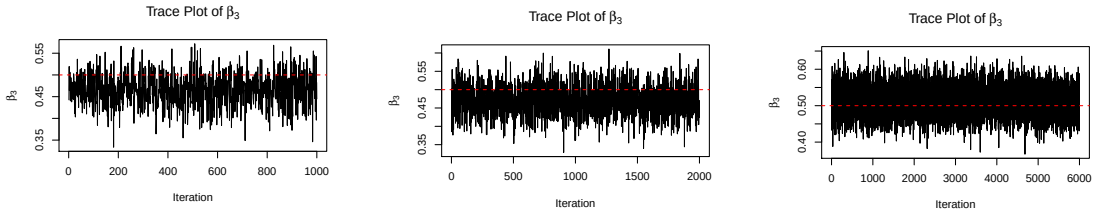

Figure 2.2: Trace plots of β_3 estimates with respect to different iteration numbers. From the left, the first, second and third trace plot has generated under 50000, 100000 and 300000 iterations respectively.

Since trace plots shows precisely where the chain has been exploring and roughly flat around, we can observe from Figure 2.2 that for 50000 iteration, the Markov Chain

is approximately flat around at 0.45 when the true value considered was 0.5. The chain became closer to 0.5 (approximately 0.47) for 100000 iterations and for 300000 iterations, the chain mixed better, become more fuzzy and flat around almost at 0.5.

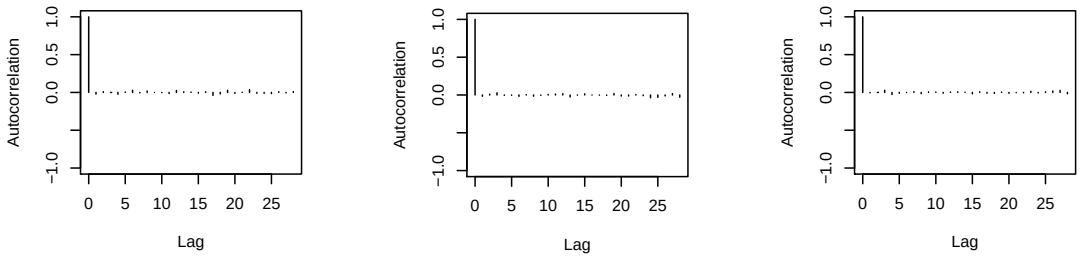

Figure 2.3: Autocorrelation plots of β_3 estimates with respect to different iteration numbers. From the left, the first, second and third trace plot has been generated for 50000, 100000 and 300000 iterations, respectively.

From the autocorrelation plot for β_3, in Figure 2.3, the Markov Chain shows slight correlations between the MCMC draws for 500000 and 100000 iterations, which is an indication of slow convergence. However, 300000 iterations eliminates the correlations and ensure swift convergence.

2.1.2 Sample size

In the sample sizes (n) scenario 100, 1000 and 10000 subjects were iterated 50000 times in order to monitor the behaviour of MCMC estimates. The true values of the parameters considered similar as before (section 2.1).

The following tables and graphs show the summary of the simulation results. The estimate is the sample mean of the distinct sample sizes.

Table 2.7: MCMC summary for 100 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.481	0.203	0.027	0.721
β_1	0.5	0.496	0.108	0.016	0.681
β_2	0.5	0.560	0.148	0.030	0.692
β_3	0.5	0.120	0.108	0.016	0.743
ϕ_x	0.0	0.056	0.129	0.023	0.510
γ^2	1.0	1.356	0.406	0.171	0.506
δ^2	1.0	0.846	0.199	0.038	0.658

Table 2.8: MCMC summary for 1000 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.512	0.107	0.005	0.743
β_1	0.5	0.601	0.108	0.007	0.741
β_2	0.5	0.468	0.050	0.003	0.721
β_3	0.5	0.512	0.041	0.002	0.787
ϕ_x	0.0	-0.013	0.035	0.001	0.621
γ^2	1.0	1.042	0.063	0.005	0.863
δ^2	1.0	0.965	0.055	0.003	0.844

Table 2.9: MCMC summary for 10000 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.520	0.016	0.000	0.826
β_1	0.5	0.500	0.012	0.000	0.865
β_2	0.5	0.491	0.014	0.000	0.941
β_3	0.5	0.502	0.012	0.000	0.953
ϕ_x	0.0	-0.005	0.011	0.000	0.628
γ^2	1.0	1.010	0.017	0.000	0.865
δ^2	1.0	1.022	0.027	0.000	0.782

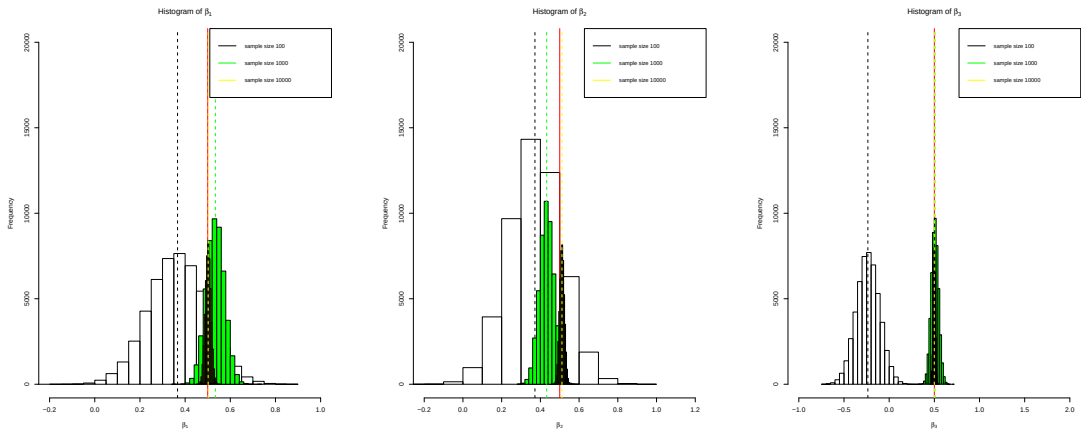

Figure 2.4: Histogram of the MCMC estimates of β_1, β_2 and β_3 with respect to different sample size. The vertical red solid line represents the true parameter value and the black, green and yellow line indicates the sample size 100, 1000 and 10000, respectively.

Tables 2.7, 2.8 and 2.9 and Figure 2.4 show the impact of sample size on the Bayesian

estimates. Sample size has been expected to provide a significant impact on the estimates, that, considering the table values, meet the expectations. With the increment of n, the variability of the estimates decreased remarkably; as well as the MSE. Besides, the increasing sample sizes improves 95% empirical coverage probabilities of the parameters. More interestingly, as the last graph shows, convergence of β_3 requires large sample size for better convergence.

2.1.2.1 Diagnostic test and Plots

Tables 2.10, 2.11 and 2.12 show Heidelberg and Welch Diagnostic test results for 100, 1000 and 10000 sample sizes, respectively.

Table 2.10: Heidelberg and Welch Diagnostic test output for 100 sample size

Variable	**Stationarity**	**p-value**
β_0	passed	0.364
β_1	passed	0.219
β_2	passed	0.735
β_3	passed	0.618
ϕ_x	passed	0.559
γ^2	passed	0.166
δ^2	passed	0.904

Table 2.11: Heidelberg and Welch Diagnostic test output for 1000 sample size

Variable	**Stationarity**	**p-value**
β_0	passed	0.783
β_1	passed	0.880
β_2	passed	0.559
β_3	passed	0.411
ϕ_x	passed	0.746
γ^2	passed	0.126
δ^2	passed	0.805

Table 2.12: Heidelberg and Welch Diagnostic test output for 10000 sample size

Variable	**Stationarity**	**p-value**
β_0	passed	0.635
β_1	passed	0.526
β_2	passed	0.510
β_3	passed	0.168
ϕ_x	passed	0.933
γ^2	passed	0.177
δ^2	passed	0.304

The results from the tables confirm that the Markov Chains converge to the stationary states for all the parameters, for all the sample sizes. The trace plot and autocorrelation plots has been generated for β_3 (coefficient of the interaction term) for further assessment of the diagnosis about the convergence of the chains.

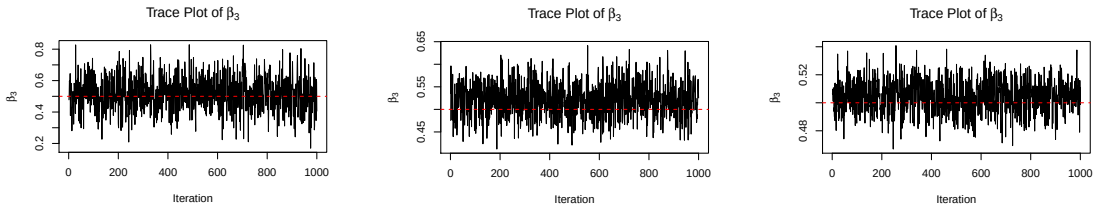

Figure 2.5: Trace plots of β_3 estimates with respect to different sample sizes. The first, second and third trace plots were generated under 100, 1000 and 10000 sample sizes.

From the trace plots of β_3 in Figure 2.5 we can observe that, with 100 subjects, Markov Chain explores values around 0.2 to 0.8. As the sample size increases to 10000, the range of exploring shrinks between 0.48 to 0.52 with less fluctuations.

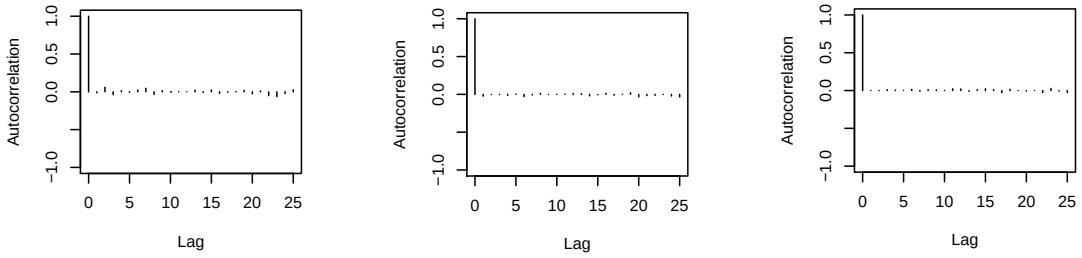

Figure 2.6: Autocorrelation plots of β_3 estimates with respect to different iteration numbers. From the left, the first, second and third trace plot has generated for 100, 1000 and 10000 sample sizes, respectively.

From the auto correlation plots of β_3, it is clear that the Markov Chain shows slight correlation among draws for a small sample size 100. As we increase the size to 10000, the correlation become invisible and secure quick convergence.

Chapter 3

Interaction model with continuous covariate with measurement error

The presence of measurement error in the model distorts the Bayesian estimates. These distortions are expected to be minimized through taking validation subsample and replication data. The primary inferential goal in this chapter is to diagnose the impact of measurement error on the performance of Bayesian parameters under different frameworks. Moreover, we investigate the performance of the estimates under validation subsamples and replicates.

Let Y be the response variable and X be the true but unobserved continuous explanatory variable, subject to the measurement error. Let W be the surrogate explanatory variable for X, and z is the other precisely measured continuous explanatory variable.

Then the desired regression model can be written as

$$Y|X \sim N(\beta_0 + \beta_1 X + \beta_2 z + \beta_3 X z, \delta^2),$$

where,

$$X \sim N(\phi_x, \gamma^2),$$
$$W|X \sim N(X, \omega^2).$$

When we consider W as the surrogate variable for X then based on the observed variable as Gustafson (2004) we can obtain the model and the solutions for the unknown parameters as

$$Y|W \sim N(\beta_0^* + \beta_1^* W + \beta_2^* z + \beta_3^* W z, \delta^{*2}),$$
$$W \sim N(\phi^*, \gamma^{*2}).$$

In here,

$$\phi^* = \phi_x,$$
$$\gamma^{*2} = \omega^2 + \gamma^2,$$
$$\delta^{*2} = \delta^2 + ((\beta_1 + \beta_3 z)^2 \omega^2 \gamma^2)/(\omega^2 + \gamma^2),$$
$$\beta_0^* = \frac{\beta_1 \phi_x}{(1 + \gamma^2/\omega^2)},$$
$$\beta_1^* = \frac{\beta_1}{(1 + \omega^2/\gamma^2)},$$
$$\beta_2^* = \beta_2 + \frac{\beta_2 \phi_x}{(1 + \gamma^2/\omega^2)},$$
$$\beta_3^* = \frac{\beta_3}{(1 + \omega^2/\gamma^2)}.$$

If we consider $\omega^2 = 0$, then $\delta^{*2} = \delta^2$, $\beta_0^* = 0$, $\beta_1^* = \beta_1, \beta_2^* = \beta_2$ and $\beta_3^* = \beta_3$. This confirms that in the absence of measurement error, the model is identified. However, in the presence of measurement error (ω^2), the above equations imply that there are more parameters than available data. In this case we can get extra data either from validation subsamples or replicates.

The joint density of unobserved quantities given the observed ones, according to Bayesian method, can be written as

$$f(\eta|y,w) \propto \prod_{i=1}^{n}\{f(y_i|x_i,\tilde{\beta},\delta^2)f(w_i|x_i,\omega^2)f(x_i|\phi_x,\gamma)\}f(\tilde{\beta})f(\phi_x)f(\gamma^2)f(\delta^2)f(\omega^2),$$

where, $\eta = (\beta_0, \beta_1, \beta_2, \beta_3, \phi_x, \delta^2, \gamma^2, \omega^2)$ is the vector of unknown parameters. Consider applying improper priors for the regression coefficients β's and ϕ_x and proper priors of Inverse Gamma distribution for the variance components δ^2, γ^2 and ω^2.

$$f(\beta) \propto 1, \quad f(\phi_x) \propto 1, \quad \gamma^2 \sim IG(0.5, 0.5), \quad \delta^2 \sim IG(0.5, 0.5), \quad \omega^2 \sim IG(0.5, 0.5).$$

Therefore the posterior becomes

$$\begin{aligned} f(\eta|y,w) \propto & (\frac{1}{\delta^2})^{n/2} e^{-\Sigma_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i-\beta_2 z_i-\beta_3 x_i z_i)^2/2\delta^2} \\ & \times (\frac{1}{\omega^2})^{n/2} e^{-\Sigma_{i=1}^{n}(w_i-x_i)^2/2\omega^2} \\ & \times (\frac{1}{\omega^2})^{n/2} e^{-\Sigma_{i=1}^{n}(x_i-\phi_x)^2/2\gamma^2} \\ & \times (\frac{1}{\gamma^2})^{0.5+1} e^{-(0.5)/2\gamma^2} \\ & \times (\frac{1}{\delta^2})^{0.5+1} e^{-(0.5)/2\delta^2} \\ & \times (\frac{1}{\omega^2})^{0.5+1} e^{-(0.5)/2\omega^2}. \end{aligned}$$

For validation subsamples, the measurements of X are known for some of the subjects. Therefore, X can be partitioned into x_c and x_r, where, subscripts c and r denote complete and reduced cases, respectively. That is, we observe only (W, Y, z) for reduced case and observe (W, Y, z, X) for the complete cases. let $\eta_M = (\beta_0, \beta_1, \beta_2, \beta_3, \delta^2)$ is the vector of parameters in the error model, $\eta_R = (\phi_x, \gamma^2)$ is the vector of parameters in the response model and $\eta_E = \omega^2$ is the parameter in the exposure models. The posterior distribution, for validation subsamples, takes the form

$$\begin{aligned} f(x_r, \eta|y, w, x_c, z) \propto & \prod_{i=1}^{n} f(w_i|x_i, \eta_M) \\ & \times \prod_{i=1}^{n} f(y_i|x_i, z_i, \eta_R) \\ & \times \prod_{i=1}^{n} f(x_i, \eta_E) \\ & \times f(\eta_M, \eta_R, \eta_E). \end{aligned}$$

For replication study design, repeated measurements of W, say, m_i replicated measurements are made for the i^{th} subject. Then posterior distribution for the replication

design takes the form

$$\begin{aligned} f(x, \eta | y, w, z) \propto & \prod_{i=1}^{n} \prod_{j=1}^{m_i} f(w_{ij} | x_i, \eta_M) \\ & \times \prod_{i=1}^{n} f(y_i | x_i, z_i, \eta_R) \\ & \times \prod_{i=1}^{n} f(x_i, \eta_E) \\ & \times f(\eta_M, \eta_R, \eta_E). \end{aligned}$$

For the simulation purpose, this section turns the attention to analyze and describe the behaviour of the Bayesian naive estimates in presence of measurement error, while observing the impact of (a) MC iteration number and (b) the magnitude of measurement error. Moreover, we investigated the performance of validation subsamples (two scenarios) and replicates as the remedies of measurement error issues.

3.1 Simulation studies

3.1.1 Monte Carlo iteration number

We assumed a study design with 1000 subjects where MCMC algorithm has been implemented and iterated different times to ensure the convergence of the estimated parameters towards the true values. Under each iteration scenario, the true values considered $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0.5$, $\gamma^2 = 1$, $\delta^2 = 1$, $\phi_x = 0$, ω^2 (measurement error) $= 0.5$ and 1000 burn-ins iteration values. Moreover, for the validation design, we generated 40 (4% for the sample size) accurately measured X. For the replication design, two sets of independent Ws were generated from the standard normal distribution. The associated tables and graphs were produced accordingly as follows.

Table 3.1: Estimates of β_1 for different number of iterations, for validation subsamples, replicates as well as naive

Iteration	**Validation**	**Replication**	**Naive**
1000	0.552	0.498	0.342
5000	0.496	0.453	0.289
10000	0.476	0.502	0.322
20000	0.546	0.531	0.364
30000	0.522	0.525	0.349
40000	0.514	0.489	0.350
50000	0.477	0.501	0.321

Table 3.2: Estimates of β_2 for different number of iterations, for validation subsamples, replicates as well as naive

Iteration	**Validation**	**Replication**	**Naive**
1000	0.491	0.484	0.521
5000	0.491	0.501	0.477
10000	0.467	0.469	0.471
20000	0.498	0.494	0.502
30000	0.509	0.503	0.500
40000	0.536	0.530	0.549
50000	0.468	0.469	0.471

Table 3.3: Estimates of β_3 for different number of iterations, for validation subsamples, replicates as well as naive

Iteration	**Validation**	**Replication**	**Naive**
1000	0.450	0.455	0.310
5000	0.470	0.422	0.270
10000	0.499	0.476	0.337
20000	0.472	0.494	0.320
30000	0.492	0.510	0.339
40000	0.475	0.499	0.330
50000	0.500	0.475	0.338

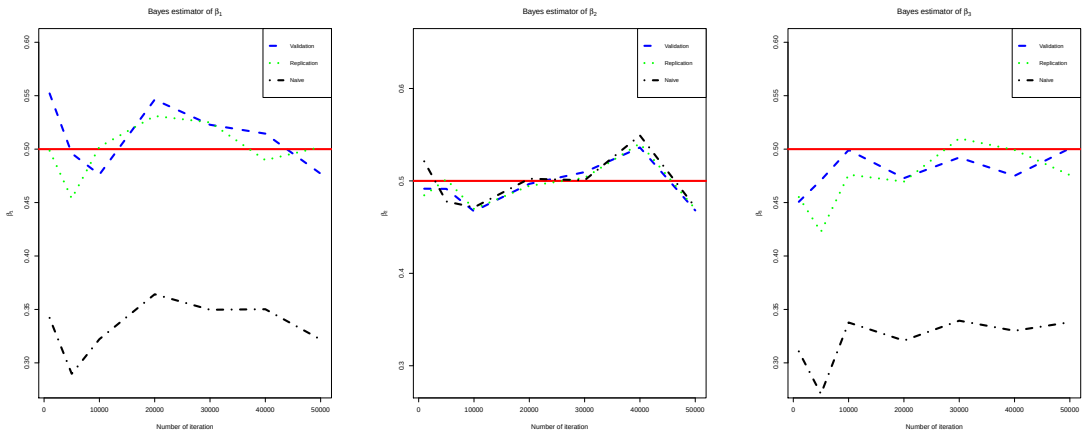

Figure 3.1: Line graph of the MCMC estimates of β_1, β_2 and β_3 with respect to different iteration numbers. The horizontal red solid line represents the true parameter value and the dashed blue, dotted green and black dotted dashed line indicates the validated, replicated and naive estimates, respectively.

Tables 5.7, 5.8, 5.9 and Figure 3.1 represent the mean MCMC estimates and the associated line graphs regarding to validation, replication and naive samples, respectively. For β_2, the coefficient of accurately measured variable, all methods of estimate perform similarly. However, in presence of error in X and therefore in Xz as well, the naive estimates perform poorly. On the contrary, both validation subsampling and replication mechanism, the Markov Chain nearly converged to the true value, as the iteration numbers increased. Interestingly, replication design performs better than the validation design for both erroneous variable X and interaction term, Xz. A possible reason for this may be, 4% validation subsample contains less information about the parameter than the replication does.

3.1.2 Amount of measurement error

It is expected that substantial measurement error will have a considerable impact on the estimation of parameters and as a result can make a misleading inference. By introducing different amounts of mismeasurement phenomenon in the model, one can have a sense of how measurement error can affect the associated regression coefficients. The density plots and estimated values from simulation of the parameter of interest can make this belief more visible. MCMC method has been carried out to simulate two data sets from the similar setting as iteration case while considering measurement error (ω^2) as 0.5 and 1 only.

Table 3.4: Effect of measurement error, $\omega^2 = 0.5$, on MCMC estimates from validation subsample

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.452	0.068	0.002	0.515
β_1	0.5	0.551	0.072	0.006	0.673
β_2	0.5	0.492	0.036	0.001	0.688
β_3	0.5	0.430	0.066	0.005	0.613
ϕ_x	0.0	0.013	0.040	0.002	0.783
γ^2	1.0	0.975	0.051	0.003	0.562
δ^2	1.0	0.955	0.051	0.002	0.683
ω^2	0.5	0.764	0.268	0.0248	0.695

Table 3.5: Effect of measurement error, $\omega^2 = 0.5$, on MCMC estimates using replication design

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.525	0.032	0.004	0.746
β_1	0.5	0.495	0.037	0.008	0.799
β_2	0.5	0.486	0.037	0.001	0.804
β_3	0.5	0.455	0.059	0.004	0.824
ϕ_x	0.0	0.018	0.039	0.002	0.862
γ^2	1.0	0.995	0.030	0.005	0.743
δ^2	1.0	0.981	0.030	0.009	0.723
ω^2	0.5	0.707	0.208	0.005	0.784

Table 3.6: Effect of measurement error $\omega^2 = 0.5$, on MCMC estimates using naive

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.425	0.012	0.002	0.697
β_1	0.5	0.342	0.019	0.001	0.577
β_2	0.5	0.521	0.010	0.002	0.752
β_3	0.5	0.310	0.051	0.010	0.517
δ^2	1.0	1.033	0.040	0.001	0.562

Table 3.7: Effect of measurement error, $\omega^2 = 1.0$, on MCMC estimates using validation subsample

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.525	0.044	0.002	0.715
β_1	0.5	0.561	0.087	0.010	0.624
β_2	0.5	0.491	0.037	0.002	0.645
β_3	0.5	0.424	0.090	0.007	0.567
ϕ_x	0.0	0.029	0.051	0.003	0.512
γ^2	1.0	0.973	0.063	0.005	0.752
δ^2	1.0	0.957	0.05	0.002	0.531
ω^2	1.0	1.056	0.075	0.006	0.516

Table 3.8: Effect of measurement error, $\omega^2 = 1.0$, on MCMC estimates using replication design

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.471	0.047	0.001	0.761
β_1	0.5	0.484	0.043	0.002	0.757
β_2	0.5	0.487	0.038	0.002	0.742
β_3	0.5	0.437	0.075	0.005	0.737
ϕ_x	0.0	0.032	0.050	0.003	0.586
γ^2	1.0	0.998	0.033	0.001	0.656
δ^2	1.0	0.988	0.027	0.001	0.742
ω^2	1.0	1.00	0.019	0.000	0.572

Table 3.9: Effect of measurement error, $\omega^2 = 1.0$, on MCMC estimates using navie

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.525	0.013	0.002	0.515
β_1	0.5	0.258	0.012	0.001	0.484
β_2	0.5	0.532	0.027	0.002	0.637
β_3	0.5	0.232	0.018	0.002	0.526
δ^2	1.0	1.066	0.070	0.003	0.621

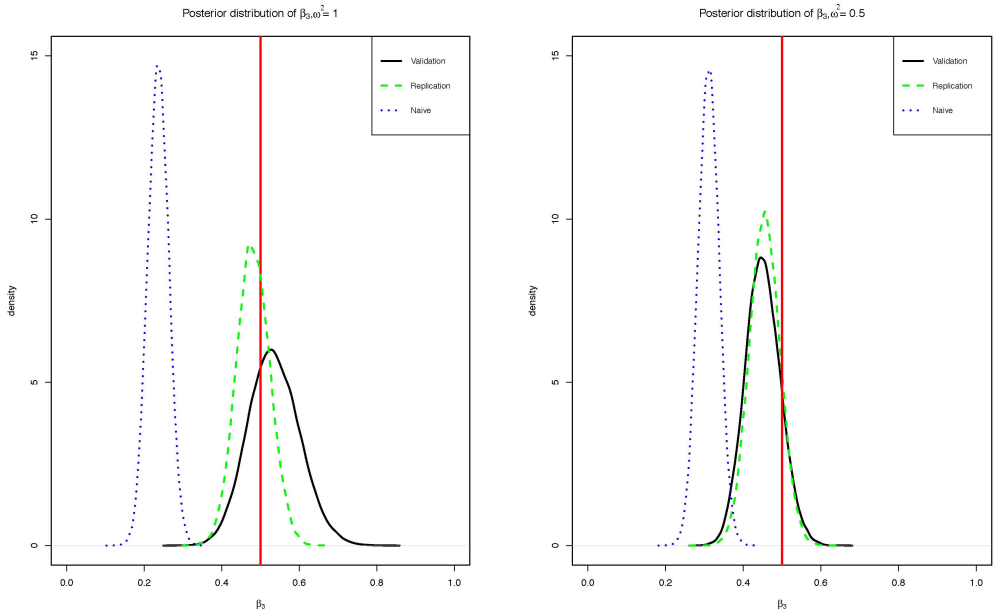

Figure 3.2: Posterior distribution of β_3 under naive, validation subsample and replication design for measurement error 0.5 and 1. The solid vertical red line indicates the true value of the parameters. The solid black, dashed green and dotted blue curves identifies the posterior distribution of validation subsample, replication and naive design.

Focusing on a higher measurement error scenario ($\omega^2 = 1$), naive provided a poor estimation of the parameter (with less variability), when replication and validation subsample provided better estimates. Decreasing the error from 1 to 0.5 affects both the location and the width of the posterior densities of β_3, and this adjustment moves the posterior distributions closer to the true value with less variability.

In overall, for both ω^2 cases, replication design provides a better estimates and convergence to the true value of β_3. More probable reason for this context is, we considered only 4% of the validation subsample while repeated samples has been taken for replication data to adjust measurement error.

3.1.3 Validation subsample

Since the continuous explanatory variable X is not precisely measured and replaced by a noisy surrogate W, therefore one possible adjustment for the bias caused by mismeasured variable would be considering validation subsample. We simulated data sets for two validation cases (4% and 20% of the sample size) considering 1000 Monte Carlo subjects that iterated 50000 times.

Starting the simulation framework with 4% validation subsample (this refers only 4% of the subjects consist of the accurate information of the X variable) and eventually rising the sample information to 20%, we investigated the behaviour of Bayesian estimates.

Table 3.10: Bayesian estimates for validation subsamples (4% and 20%), replicates as well as naive

Parameter	**True value**	**Validation 4%**	**Validation 20%**	**Replication**	**Naive**
β_0	0.5	0.525	0.505	0.513	0.516
β_1	0.5	0.587	0.508	0.486	0.252
β_2	0.5	0.531	0.525	0.518	0.553
β_3	0.5	0.415	0.482	0.458	0.261
ϕ_x	0.0	-0.006	0.057	0.017	-
γ^2	1.0	0.994	1.038	0.989	-
δ^2	1.0	0.982	0.994	0.968	1.114
ω^2	1.0	1.046	0.995	1.005	-

Table 3.11: MCMC estimates under 4% validation subsample

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.525	0.044	0.002	0.715
β_1	0.5	0.587	0.058	0.005	0.671
β_2	0.5	0.531	0.047	0.002	0.655
β_3	0.5	0.415	0.050	0.003	0.643
ϕ_x	0.0	-0.006	0.042	0.002	0.729
γ^2	1.0	0.994	0.059	0.004	0.582
δ^2	1.0	0.982	0.034	0.001	0.811
ω^2	1.0	1.046	0.071	0.006	0.532

Table 3.12: MCMC estimates under 20% validation subsample

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.505	0.074	0.002	0.685
β_1	0.5	0.458	0.057	0.004	0.753
β_2	0.5	0.525	0.044	0.002	0.715
β_3	0.5	0.520	0.044	0.002	0.782
ϕ_x	0.0	0.057	0.070	0.005	0.756
γ^2	1.0	1.038	0.052	0.003	0.766
δ^2	1.0	0.994	0.027	0.001	0.624
ω^2	1.0	0.995	0.035	0.001	0.657

Table 3.13: MCMC estimates under replication design

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.513	0.044	0.002	0.764
β_1	0.5	0.486	0.054	0.003	0.677
β_2	0.5	0.518	0.039	0.002	0.698
β_3	0.5	0.458	0.049	0.003	0.786
ϕ_x	0.0	0.017	0.041	0.002	0.782
γ^2	1.0	0.989	0.035	0.001	0.754
δ^2	1.0	0.968	0.040	0.001	0.823
ω^2	1.0	1.005	0.020	0.000	0.842

Table 3.14: MCMC naive estimates

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.516	0.238	0.003	0.615
β_1	0.5	0.252	0.248	0.012	0.581
β_2	0.5	0.553	0.063	0.003	0.789
β_3	0.5	0.261	0.239	0.011	0.592
δ^2	1.0	1.114	0.117	0.005	0.548

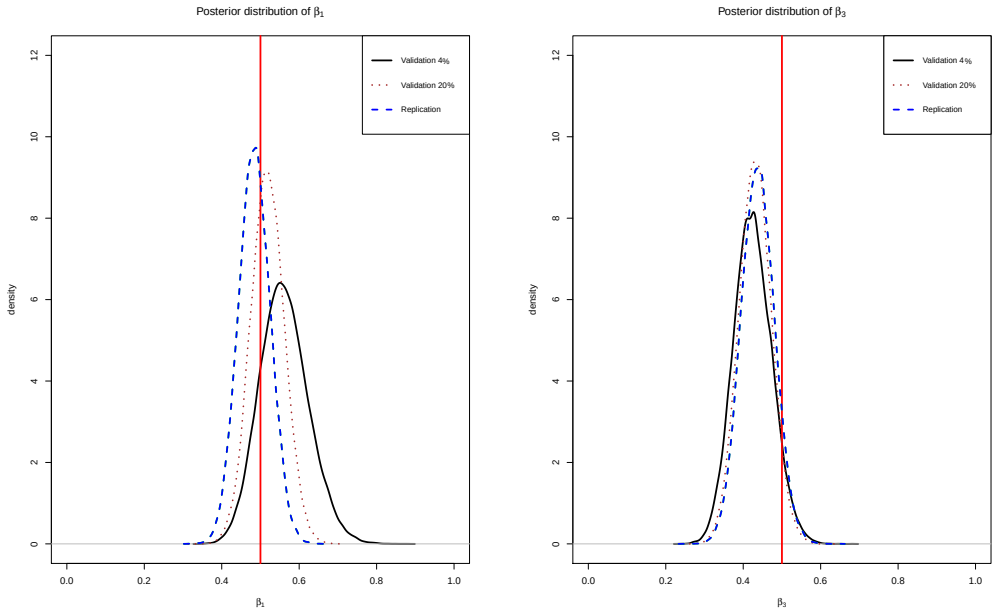

Figure 3.3: Posterior distributions of β_1 and β_3 for 4% and 20% validation subsample and replication design. The vertical red solid line is the true mean and the solid black and dotted red curves represents the posterior densities resulting from the 4% and 20% of validation subsample. And the blue dashed curve indicates the posterior distribution of replication design.

The significant impact of both validation subsamples (4% and 20%) on the behaviour of posterior densities of β_1 and β_3 can be seen from Figure 3.3 and the associated tabulated values. 4% validation subsample displays wider posterior densities, implies larger variance for β_1. Indeed, increasing the validation subsample from 4% to 20% improves the estimates and therefore assisting MCMC method for convergence towards the true value with less variability. A reasonable explanation can be because, 20% validation subsample contains more information for variable X than 4%. Besides, from the tables we can acquire that naive estimate provided poor inference when, replication data yield better performance.

Interestingly, the curve for β_3 has nearly similar variability and location under both validation subsampling (4% and 20%) and replication cases, illustrates the challenging nature of capturing information from the interaction term.

In overall scheme, we can conclude that, increasing the validation subsample highly secure the estimates to converge to the true value compared to the replication design. However, in application, exactly measuring X for 20% of the sample may be very expensive. That is why it is sometimes better to have more inaccurate replicates of X than large number of accurate ones.

Following the posterior densities of all parameters are presented for a 20% validation subsample.

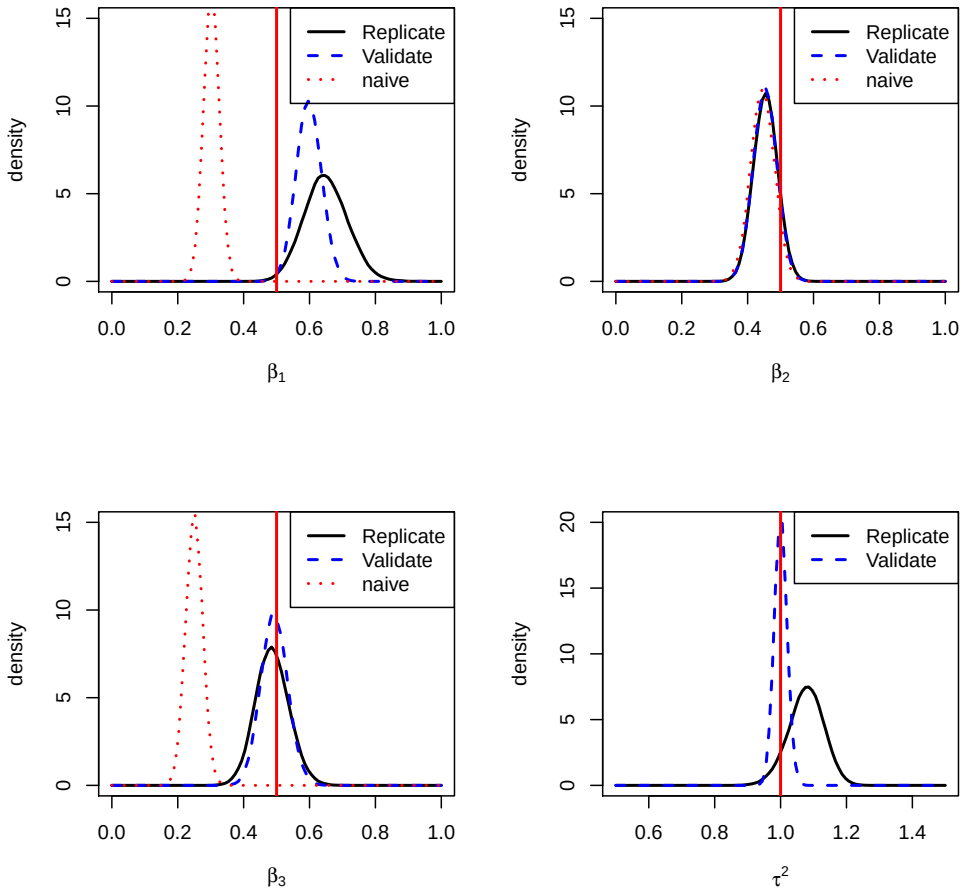

Figure 3.4: Posterior distributions of β_1, β_2, β_3 and ω^2. Here the solid black, dashed blue and dotted red curve gives the posterior density from the replication, validation subsampling and naive, respectively.

Since X is the mismeasured continuous variable, z is precisely measured continuous

variable, therefore the coefficient of z (β_2) is less likely to be affected by the measurement error presented in the model. The top right panel of β_2 in Figure 3.4 as it was expected, naive estimator performs similar to other two designs. However, measurement error has significant influence on the coefficient of X (β_1) and Xz (β_3) which is transparent from the graph as well. Naive estimates from the analysis for both β_1 and β_3 provides a density with less variability, nevertheless, this is unsuccessful for assuring almost sure convergence of the parameters. It is also noted that validated design provides slightly less variable density than the replicated ones.

The diagnostic plots and tests (not shown) satisfies the inference made from all frameworks in this chapter.

Chapter 4

Interaction model with discrete variable without misclassification

Let us consider a response model where one of the covariates is discrete

$$Y|X \sim (\beta_0 + \beta_1 X + \beta_2 z + \beta_3 X z, \delta^2)$$

Here, X is the binary variable with probability of success

$$p(X = 1) = r.$$

When X and z are both observed covariates, all the model parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \delta^2, r)$ can be uniquely estimated. Then the joint density of unobserved quantities given the observed ones is obtained as

$$f(\eta|y, x) \propto \prod_{i=1}^{n} \{f(y_i|x_i, \tilde{\beta}, \delta^2) p(x_i, r) f(\tilde{\beta})\} f(r) f(\delta^2),$$

where $\eta = (\beta_0, \beta_1, \beta_2, \beta_3, \delta^2, r)$ is the vector of unknown parameters of interest. Consider applying improper priors for the regression coefficients (β values) and proper priors of Inverse Gamma distribution for the variance component δ^2 and a non informative Beta distribution with known parameters for r. That is

$$f(\beta) \propto 1, \qquad \delta^2 \sim IG(0.5, 0.5), \qquad r \sim Beta(a, b).$$

Therefore the posterior becomes

$$\begin{aligned} f(\eta|y, x) \propto\ & r^{\Sigma_{i=1}^n x_i + na - n}(1-r)^{nb - \Sigma_{i=1}^n x_i} \\ & \times (\frac{1}{\delta^2})^{n/2} e^{-\Sigma_{i=1}^n (y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i - \beta_3 x_i z_i)^2 / 2\delta^2} \\ & \times (\frac{1}{\delta^2})^{0.5+1} e^{-(0.5)/2\delta^2} \end{aligned}$$

To investigate the performance of the posterior distributions under various scenarios, simulation studies have been conducted, where no misclassification has been considered in the desired model. The scenarios considered for (a) MC iteration number, (b) sample size (n) and (c) prior selection for r.

4.1 Simulation studies

4.1.1 Monte Carlo iteration number

The purpose of this section is to observe the converging trend of the MCMC estimates to the true value while varying the iteration numbers. A sample of size 1000 has been iterated 50000, 100000 and 300000 times for our study purpose. The discrete covariate X has been generated for two probability cases - rare case (probability $r = 0.05$) and

common case (probability $r = 0.5$) to observe the impact of r on the estimates. For each iteration and probability scenario, other model parameters were assigned as $\beta = 0.5$ with $\delta^2 = 1$, and 1000 burn-ins.

4.1.1.1 Rare case

The MCMC estimated values of the unknown parameters and the associated graphs has been produced to understand impact of iteration numbers as well as the magnitude of probability r.

Table 4.1: Summary of MCMC estimates for 50000 iteration

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.048	0.006	0.000	0.515
β_0	0.50	0.471	0.042	0.002	0.673
β_1	0.50	0.125	0.261	0.071	0.546
β_2	0.50	0.464	0.048	0.002	0.626
β_3	0.50	0.240	0.264	0.077	0.756
δ^2	1.00	1.002	0.045	0.002	0.747

Table 4.2: Summary of MCMC estimates for 100000 iteration

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.056	0.010	0.000	0.741
β_0	0.50	0.501	0.030	0.001	0.634
β_1	0.50	0.483	0.142	0.028	0.781
β_2	0.50	0.507	0.033	0.001	0.792
β_3	0.50	0.343	0.212	0.053	0.794
δ^2	1.00	0.873	0.131	0.009	0.737

Table 4.3: Summary of MCMC estimates for 300000 iteration

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.061	0.014	0.000	0.654
β_0	0.50	0.484	0.036	0.001	0.836
β_1	0.50	0.458	0.221	0.053	0.896
β_2	0.50	0.513	0.035	0.001	0.723
β_3	0.50	0.546	0.049	0.031	0.781
δ^2	1.00	1.011	0.016	0.003	0.941

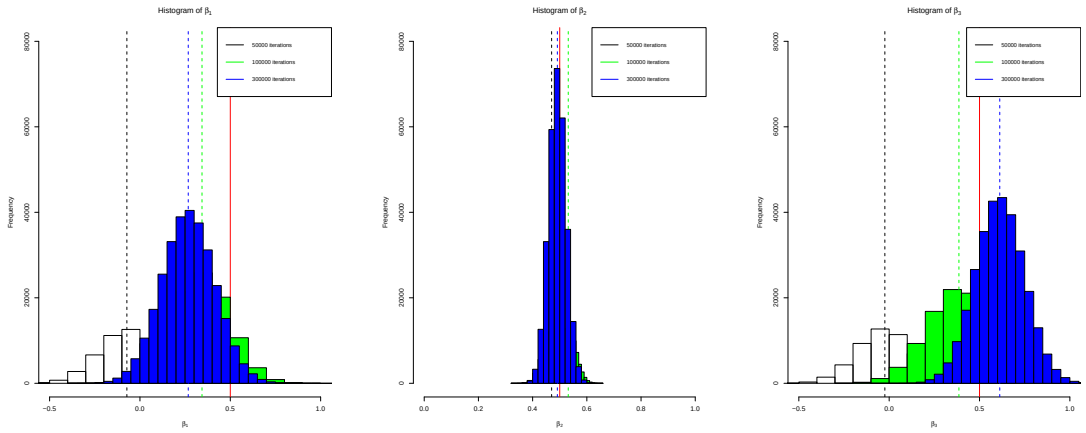

Figure 4.1: Histogram of the MCMC estimates of β_1, β_2 and β_3 with respect to different iteration numbers under rare probability. The vertical red solid line represents the true parameter value and the dashed black, green and blue line indicates the 50000, 100000 and 300000 iteration estimates of β values, respectively.

Table 4.1, 4.2 and 4.3 display the estimated MCMC parameters with S.D., MSE as well as the empirical 95% coverage probability values from the simulation studies. According to the tabulated values, parameters began to approach to the true value with the increment of iterations number. Besides, the MSE and S.D. values began to decrease. However, β_1, the coefficient of X required more iterations to converge towards the true value. The histograms help to visualize the converging behaviour of MCMC estimates towards their true values, more clearly. The white, green and blue histograms represents 50000, 100000 and 300000 iterations, respectively. With the increment of iteration number, the MCMC means start converging to their true means. As predictor X has few numbers of observations, it costs more iteration for β_1 to reach near to the true value. On the contrary, the estimated MCMC β_2's were closer to the true mean. Moreover, for the interaction term, it took more iterations for β_3 to reach the true value. However, with the increment of iteration number, the variability starts to shrink in all

β cases.

4.1.1.2 Common probability

A similar set up has been considered for the common case of X variable. With a higher probability ($r = 0.5$), X has more successes to estimate the parameters. Therefore, the convergence for all parameters are faster than the rare case ($r = 0.05$).

Table 4.4: Summary of MCMC estimates for 50000 iteration

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.5	0.522	0.027	0.000	0.765
β_0	0.5	0.595	0.105	0.009	0.736
β_1	0.5	0.321	0.189	0.023	0.785
β_2	0.5	0.588	0.099	0.008	0.714
β_3	0.5	0.347	0.164	0.019	0.632
δ^2	1.0	1.033	0.057	0.004	0.942

Table 4.5: Summary of MCMC estimates for 100000 iteration

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.5	0.483	0.023	0.000	0.854
β_0	0.5	0.462	0.057	0.004	0.871
β_1	0.5	0.555	0.083	0.008	0.841
β_2	0.5	0.512	0.045	0.002	0.864
β_3	0.5	0.475	0.066	0.006	0899
δ^2	1.0	0.959	0.059	0.004	0.876

Table 4.6: Summary of MCMC estimates for 300000 iteration

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.5	0.520	0.026	0.000	0.948
β_0	0.5	0.491	0.046	0.003	0.913
β_1	0.5	0.501	0.063	0.005	0.979
β_2	0.5	0.542	0.063	0.005	0.894
β_3	0.5	0.472	0.071	0.007	0.965
δ^2	1.0	1.000	0.044	0.002	0.928

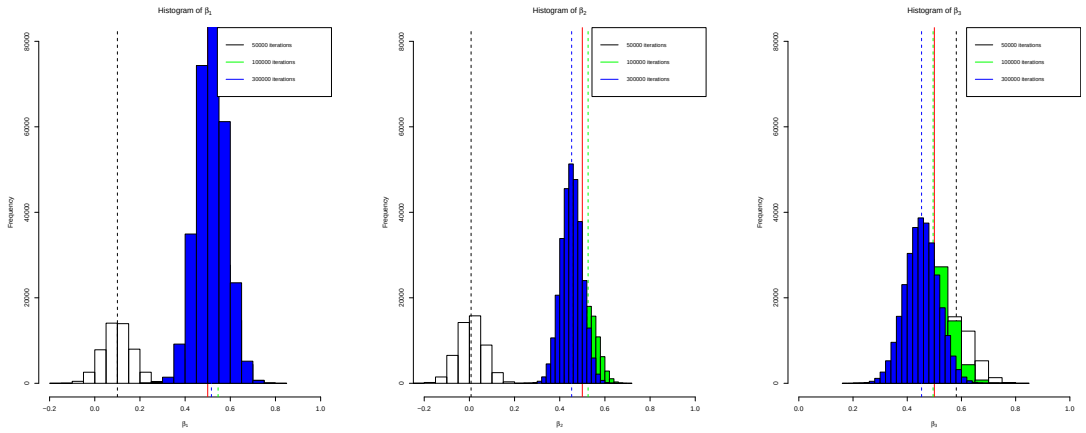

Figure 4.2: Histogram of the MCMC estimates for β_1, β_2 and β_3 with respect to different iteration numbers under common probability. The vertical red solid line represents the true parameter value and the dashed black, green and blue lines indicate the 50000, 100000 and 300000 iterated estimates of β values.

The histograms and the outputs in Table 4.4, Table 4.5 and Table 4.6 show that changing of the iteration number has a good impact on the MCMC estimates. Lower iteration generates poor estimates while higher iteration number minimize the distance between the true and estimated values of the parameters. In addition, β_2 achieves the lowest variability for 300000 MCMC iteration. Furthermore, the associated MSE and S.D. values began to decrease as the chain numbers increase.

4.1.2 Sample size

It is expected that the sample size has a significant effect on the posterior estimates. Therefore, the behaviour of the estimates were studied for the sample size (n) 100, 1000, and 10000. For these scenarios, we considered r to be 0.05 and iteration numbers to be 50000. The other parameters were as before. Histograms of estimated values for

all the coefficients (β values) were made to visualize the behaviour of the estimated parameters.

Table 4.7: Summary of MCMC estimates for 100 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.078	0.038	0.002	0.571
β_0	0.50	0.545	0.128	0.023	0.547
β_1	0.50	0.147	0.620	0.522	0.531
β_2	0.50	0.501	0.101	0.014	0.573
β_3	0.50	0.402	0.374	0.201	0.624
δ^2	1.00	1.344	0.396	0.168	0.574

Table 4.8: Summary of MCMC estimates for 1000 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.046	0.007	0.000	0.734
β_0	0.50	0.479	0.038	0.000	0.648
β_1	0.50	0.840	0.373	0.001	0.581
β_2	0.50	0.516	0.037	0.000	0.745
β_3	0.50	0.623	0.202	0.000	0.743
δ^2	1.00	1.046	0.066	0.000	0.862

Table 4.9: Summary of MCMC estimates for 10000 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.051	0.002	0.000	0.845
β_0	0.50	0.506	0.012	0.000	0.782
β_1	0.50	0.497	0.044	0.000	0.851
β_2	0.50	0.490	0.013	0.000	0.774
β_3	0.50	0.532	0.055	0.000	0.815
δ^2	1.00	0.980	0.023	0.000	0.951

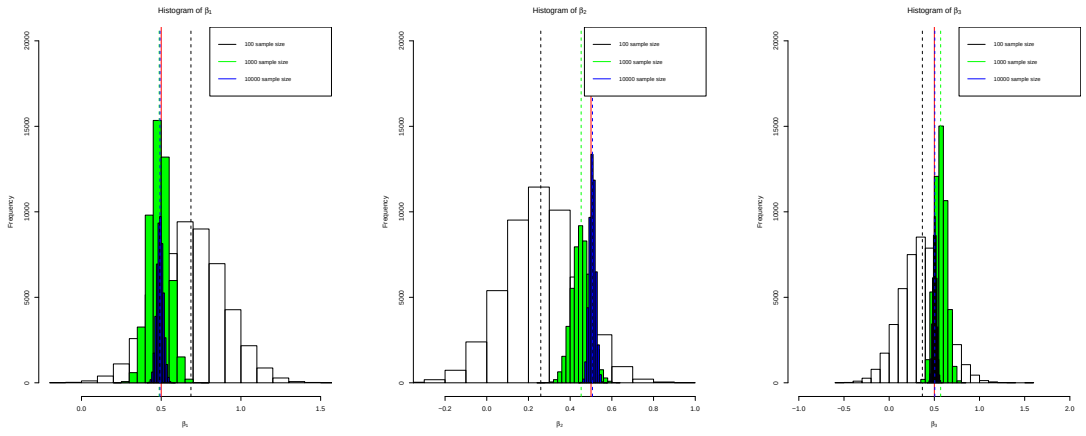

Figure 4.3: Histograms of the MCMC estimates of β_1, β_2 and β_3 with respect to different sample sizes for the rare case ($r = 0.05$). The vertical red solid line represents the true parameter value and the dashed black, green and blue lines indicate 100, 1000 and 10000 samples, respectively.

The estimated values of the unknown parameters are presented in Tables 4.7, 4.8 and 4.9. The dramatic change of the parameters and their variances is clear from the

histogram of β values in Figure 4.3,

All the β estimates with small sample size appeared to have large variances. Moreover, for small sample size the MCMC estimates were far from the true values. For larger sample sizes, the variances became small. Besides, the estimates converge to the true values with the increase in n. Therefore, larger sample size helped the MCMC estimates to converge faster to their true values with less variability.

4.1.3 Beta Prior

In this section the impact of prior selection for r on the estimation process was explored. For this purpose, Beta distributions with parameters 5 and 1 as the least informative, 1 and 1 (uniform between zero and one) as non informative and 2 and 5 as most informative, were selected. A sample of 1000 observations were iterated 50000 times to create the following tables and histograms. Other parameters were set the same as the last section.

Table 4.10: Summary of MCMC estimates for the least informative prior (Beta(5, 1))

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.050	0.015	0.000	0.848
β_0	0.05	0.540	0.059	0.004	0.872
β_1	0.5	0.036	0.072	0.007	0.841
β_2	0.5	0.575	0.058	0.004	0.871
β_3	0.5	0.412	0.071	0.007	0.913
δ^2	1.0	0.969	0.053	0.003	0.741

Table 4.11: Summary of MCMC estimates for the non informative prior (Beta(1, 1))

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.049	0.018	0.000	0.871
β_0	0.5	0.512	0.046	0.003	0.536
β_1	0.5	0.488	0.064	0.005	0.885
β_2	0.5	0.487	0.046	0.003	0.926
β_3	0.5	0.459	0.088	0.009	0.877
δ^2	1.0	1.109	0.046	0.003	0.984

Table 4.12: Summary of MCMC estimates for the most informative prior (Beta(2, 5))

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
r	0.05	0.049	0.018	0.000	0.983
β_0	0.5	0.530	0.053	0.003	0.944
β_1	0.5	0.454	0.077	0.007	0.917
β_2	0.5	0.451	0.046	0.003	0.834
β_3	0.5	0.482	0.066	0.006	0.921
δ^2	1.0	0.976	0.049	0.003	0.887

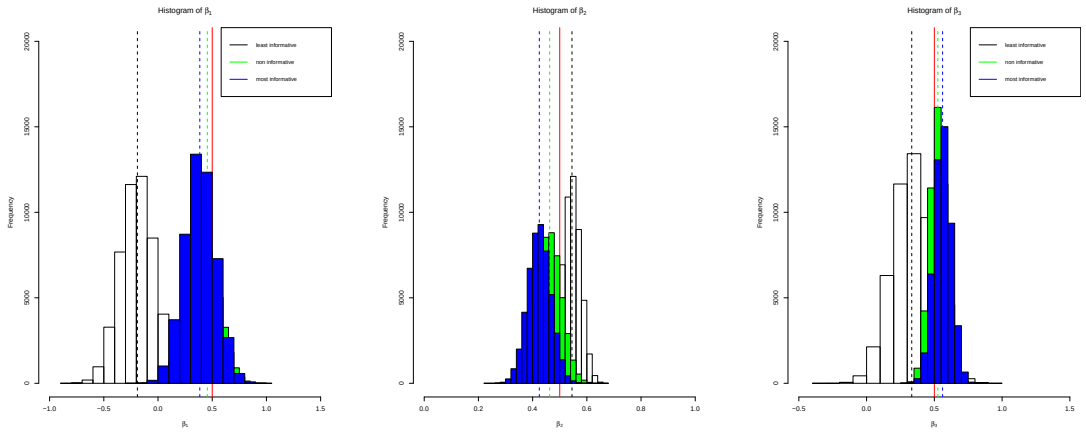

Figure 4.4: Histograms of the MCMC estimates for β_1, β_2 and β_3 under common probability. The vertical red solid line represents the true parameter value, and the dashed black, green and blue lines represent the least, non and the most informative priors, respectively.

Figure 4.4 shows that for β_1 and β_3, non and most informative beta priors provided similar histograms, where the least informative prior provided the worst results. For β_2, all priors perform quite similar. This implies that for the coefficient of z, the choice of prior for r is not as important as for the coefficients of X and Xz.

Chapter 5

Interaction model with discrete variable with misclassification

Let X be a binary exposure variable in the regression model that we discussed in Chapter 4. In here, X is unobservable and instead, a surrogate binary covariate W is measured that incorporates the error and leads to the term misclassification. The magnitude of misclassification is characterized in terms sensitivity (probability of correct classifying success) and specificity (correctly classifying failure). Therefore, the response model is as follows

$$Y|X \sim (\beta_0 + \beta_1 X + \beta_2 z + \beta_3 X z, \delta^2),$$

where,

$$P(X = 1) = r.$$

Moreover,

$$P(W = 1|X = 1) = u$$

$$P(W = 0|X = 0) = v,$$

where u is the sensitivity and v is the specificity.

Based on the misclassified W we have

$$Y|W \sim N(\beta_0^* + \beta_1^* W + \beta_2^* z + \beta_3^* W z, \delta^{2*}), \tag{5.1}$$

where

$$\delta^{2*} = \delta^2 + (\beta_1 + \beta_3 z)^2 Var(X|W),$$

and

$$\begin{aligned} Var(X|W) =& \frac{(1-u)r}{1-r_w}(1 - \frac{(1-u)r}{1-r_w}) - [\frac{ur}{r_w} + (1 - \frac{(1-u)r}{1-r_w})^2]W^2 \\ &+ [(\frac{ur}{r_w} - \frac{(1-u)r}{1-r_w})(1 - 2\frac{(1-u)r}{1-r_w})]W. \end{aligned}$$

The coefficients in model (5.1) are as follows

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_1(\frac{(1-u)r}{1-r_w}), \\ \beta_1^* &= \beta_1(\frac{ur}{r_w} + \frac{(1-u)r}{1-r_w}), \\ \beta_2^* &= \beta_2 + \beta_3\frac{(1-u)r}{1-r_w}, \\ \beta_3^* &= \beta_3(\frac{ur}{r_w} + \frac{(1-u)r}{1-r_w}). \end{aligned}$$

In here, r_w is the probability of success for W that is defined as

$$r_w = P(W = 1) = ru + (1 - u)(1 - r).$$

Since r_w varies between zero and one, with the following bounds are needed

$$min(u, 1 - v) \leq r_w \leq max(u, 1 - v).$$

The joint density of unobserved quantities given the observed ones is obtained as

$$f(\eta, x|y, w) \propto \prod_{i=1}^{n} \{f(y_i|x_i, , \tilde{\beta}, \delta^2)p(w_i/x_i, u, v)p(x_i, r)f(\tilde{\beta})\}f(r)f(\delta^2)f(u, v, r),$$

where $\eta = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3, r, \delta^2,\ v,\ u)$ is the vector of unknown parameters. Applying improper priors for the regression coefficients (β values) and proper priors of Inverse Gamma distribution for the variance component δ^2 and non informative uniform distributions for r, u and v, we have

$$\begin{aligned} f(\eta|y, w) \propto\ & r^{\Sigma_{i=1}^{n} x_i}(1 - r)^{n - \Sigma_{i=1}^{n} x_i} \\ & \times u^{\Sigma_{i=1}^{n} w_i}(1 - u)^{n - \Sigma_{i=1}^{n} w_i} \\ & \times v^{\Sigma_{i=1}^{n} x_i}(1 - v)^{n - \Sigma_{i=1}^{n} w_i} \\ & \times (\frac{1}{\delta^2})^{n/2} e^{-\Sigma_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i - \beta_3 x_i z_i)^2 / 2\delta^2} \\ & \times (\frac{1}{\delta^2})^{0.5+1} e^{-(0.5)/2\delta^2}. \end{aligned}$$

The behaviour of the posterior distributions under various scenarios has been observed through simulating suitable data set. The scenarios considered were (a) Monte Carlo

iteration number, (b) sample size (n), (c) prior selection and (d) sensitivity and specificity.

5.1 Simulation studies

5.1.1 Monte Carlo iteration number

In order to observe the impact of number of iterations on convergence, a sample of 1000 observations were replicated 50000, 100000 and 300000 times. The response model's parameters were kept the same as Chapter 4. Moreover, X is unobservable and W provides error-prone information for X in the model. Sensitivity and specificity were considered to be 0.9 and 0.3, respectively. The simulation studies were done for two separate scenarios of rare probability ($r = 0.05$) and common probability ($r = 0.5$).

5.1.1.1 Rare probability

It is expected that for small number of successes in the discrete variable X the convergence rate of Monte Carlo chains is slower (as comparing to larger number of successes). The associated tables and graphs are produced as follows.

Table 5.1: Summary of MCMC estimates for 50000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.50	0.498	0.039	0.005	0.671
β_1	0.50	0.627	0.180	0.003	0.767
β_2	0.50	0.468	0.051	0.002	0.538
β_3	0.50	0.498	0.191	0.004	0.543
r	0.05	0.051	0.007	0.002	0.874
δ^2	1.00	1.024	0.051	0.007	0.716
u	0.90	0.709	0.190	0.008	0.614
v	0.30	0.290	0.017	0.003	0.872

Table 5.2: Summary of MCMC estimates for 100000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.50	0.516	0.043	0.002	0.845
β_1	0.50	0.364	0.230	0.070	0.713
β_2	0.50	0.544	0.059	0.004	0.867
β_3	0.50	0.609	0.391	0.153	0.923
r	0.05	0.056	0.010	0.000	0.655
δ^2	1.00	0.980	0.048	0.003	0.961
u	0.90	0.710	0.190	0.005	0.873
v	0.30	0.289	0.017	0.000	0.832

Table 5.3: Summary of MCMC estimates for 300000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.50	0.472	0.048	0.00	0.567
β_1	0.50	0.389	0.251	0.00	0.856
β_2	0.50	0.519	0.044	0.00	0.777
β_3	0.50	0.619	0.222	0.00	0.945
r	0.05	0.046	0.007	0.00	0.993
δ^2	1.00	0.976	0.049	0.00	0.981
u	0.90	0.708	0.191	0.00	0.876
v	0.30	0.291	0.016	0.00	0.782

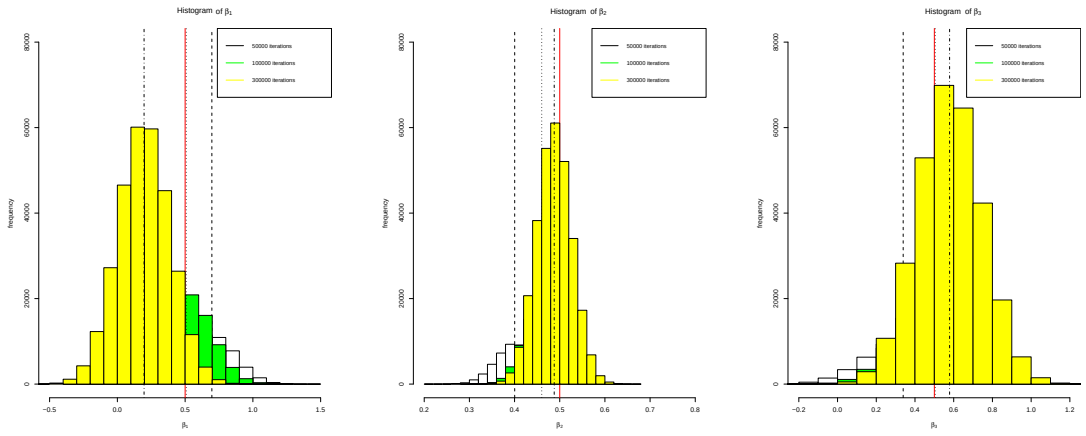

Figure 5.1: Histogram of the MCMC estimates of β_1, β_2 and β_3 with respect to different iteration numbers under rare probability in presence of misclassification. The vertical red solid line represents the true parameter value and the dashed, dotted and dotted dashed lines indicates the 50000, 100000 and 300000 replicated estimates of β values.

The tabulated values and the graphs for the regression coefficients have been presented in the Tables 5.1, 5.2 and 5.3. Figure 5.1 helps to visualize the convergence pattern at a glance. The red solid lines represent true value of the parameters, and white, green and yellow with dashed, dotted and dotted dashed horizontal lines, indicate the estimated values. From the histograms and tabulated values of the coefficient of X, (i.e β_1), we can observe that the increasing number of iterations did not help the MCMC estimates to converge the true parameter value. This is due to the fact that number of successes is very low (only 50). Moreover, for the rest of the observations, misclassification rate is very high ($1-0.3 = 0.7$). Without correcting for the bias caused by misclassification, the naive estimator does not perform well. However, the coefficient of accurately measured variable Z (i.e. β_2) converges faster its true value, with the increment of iterations. Finally, β_3, the coefficient of interaction slowly converges to the true value as iteration number rises. This is because both X (misclassified) and Z (accurately measured) contribute both negatively and positively to the convergence of β_3.

5.1.1.2 Common probability

The simulation setup for the common probability ($r = 0.5$) are similar to the rare probability case . The simulation outputs are provided in Tables 5.4 to 5.6. The histograms of the estimated values for the coefficients are presented in Figure 5.2.

Table 5.4: Summary of MCMC estimates for 50000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.502	0.054	0.004	0.876
β_1	0.5	0.507	0.077	0.008	0.885
β_2	0.5	0.554	0.075	0.006	0.893
β_3	0.5	0.457	0.087	0.010	0.924
r	0.5	0.488	0.019	0.000	0.853
δ^2	1.0	1.023	0.051	0.003	0.814
u	0.9	0.796	0.104	0.002	0.817
v	0.3	0.203	0.097	0.002	0.846

Table 5.5: Summary of MCMC estimates for 100000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.469	0.062	0.005	0.987
β_1	0.5	0.644	0.163	0.023	0.877
β_2	0.5	0.501	0.057	0.004	0.843
β_3	0.5	0.436	0.102	0.013	0.614
r	0.5	0.495	0.016	0.000	0.884
δ^2	1.0	0.920	0.089	0.006	0.871
u	0.9	0.799	0.101	0.002	0.766
v	0.3	0.200	0.100	0.002	0.878

Table 5.6: Summary of MCMC estimates for 300000 iterations

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.484	0.023	0.00	0.856
β_1	0.5	0.522	0.033	0.00	0.995
β_2	0.5	0.452	0.050	0.00	0.697
β_3	0.5	0.563	0.067	0.00	0.892
r	0.5	0.499	0.005	0.00	0.915
δ^2	1.0	0.994	0.015	0.00	0.893
u	0.9	0.799	0.100	0.00	0.798
v	0.3	0.200	0.099	0.00	0.673

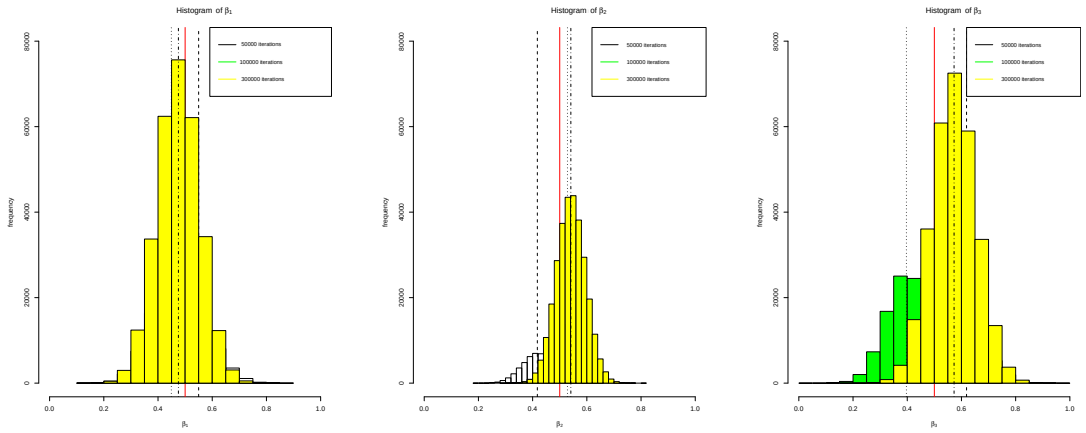

Figure 5.2: Histogram of the MCMC estimates of β_1, β_2 and β_3 with respect to different iteration numbers under common probability in presence of misclassification. The vertical red solid line represents the true parameter value and the black, green and yellow histograms and the corresponding dashed, dotted and dashed dotted lines indicate 50000, 100000 and 300000 iterations, respectively.

From the graph and table of the parameters the convergence trend is clearly visible. All the estimated MCMC parameters β_1, β_2 and β_3, start converging to the true parameter values with the increment of the replication number that meet the prediction as well. Besides, the low standard deviation value with higher iteration make the decision that high iteration can provide fast convergence.

5.1.2 Sample size

Sample size is considered to be one of the most important factor in any statistical analysis. It may have significant effect on the convergence of the Markov Chain and its convergence to the true value of the parameters. The simulation study was conducted for sample sizes 100, 1000 and 10000 with 50000 iterations. Other parameter values were 0.5 for all β values, $r = 0.05$, $u = 0.9$, $v = 0.3$ and $\delta^2 = 1$.

Table 5.7: Summary of MCMC estimates for 100 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.50	0.488	0.126	0.022	0.777
β_1	0.50	0.475	1.293	1.949	0.812
β_2	0.50	0.612	0.167	0.035	0.837
β_3	0.50	0.598	1.578	3.085	0.887
r	0.05	0.039	0.021	0.000	0.643
δ^2	1.00	0.929	0.157	0.031	0.853
u	0.90	0.706	0.199	0.017	0.854
v	0.30	0.293	0.045	0.002	0.771

Table 5.8: Summary of MCMC estimates for 1000 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.50	0.520	0.005	0.002	0.842
β_1	0.50	0.327	0.039	0.069	0.547
β_2	0.50	0.492	0.004	0.002	0.729
β_3	0.50	0.486	0.161	0.036	0.764
r	0.05	0.058	0.011	0.000	0.979
δ^2	1.00	0.946	0.068	0.004	0.871
u	0.90	0.710	0.090	0.005	0.881
v	0.30	0.289	0.017	0.000	0.615

Table 5.9: Summary of MCMC estimates for 10000 sample size

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.50	0.497	0.002	0.00	0.884
β_1	0.50	0.549	0.075	0.00	0.884
β_2	0.50	0.506	0.004	0.00	0.693
β_3	0.50	0.487	0.056	0.00	0.778
r	0.05	0.049	0.002	0.00	0.858
δ^2	1.00	1.014	0.020	0.00	0.854
u	0.90	0.709	0.090	0.00	0.533
v	0.30	0.290	0.010	0.00	0.734

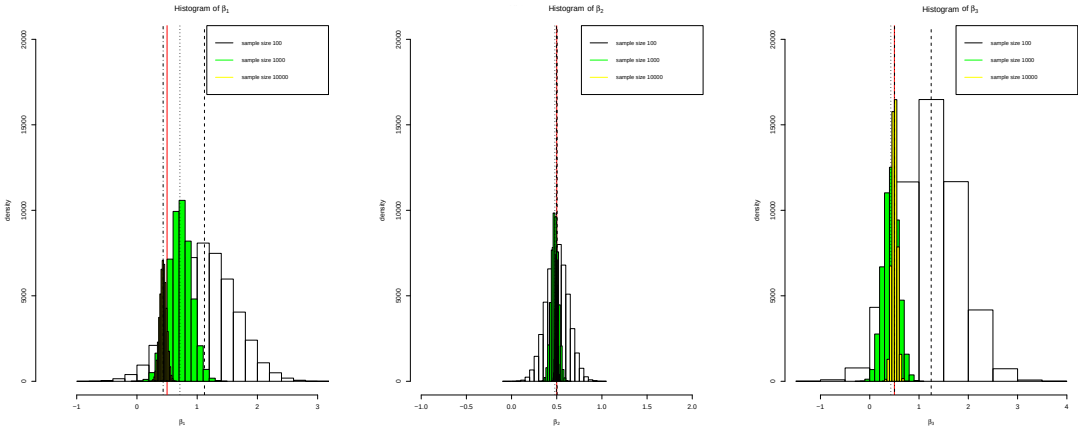

Figure 5.3: Histogram of the MCMC estimates of β_1, β_2 and β_3 for different sample sizes under rare probability. The vertical red solid line represents the true parameter value and the dashed black, green and yellow histograms indicate 100, 1000 and 10000 sample size estimates of β values.

Figure 5.3 represents histograms of β_1, β_2 and β_3, for the three sample sizes. A drastic change we can notice from the plots. Higher sample size confirms faster convergence for all the parameters. Most importantly, the variance for all β values of the histogram becomes almost zero. Interestingly, MCMC estimates for β_2 hits the true value for all sample setup compared to the other two coefficients β_1 and β_3. The incorporated error in the variable X as well in the interaction term may slower the convergence. This confirms the fact that generally, increasing sample size does not help correcting bias caused by measurement error.

5.1.3 Beta Prior

Another important simulation study done was varying beta prior for r from the least informative to the most informative, to explore the convergence of MCMC for the

response model coefficients. A sample of size 1000 was replicated 50000 times with β values set to be 0.5, and r was set to be 0.05. Sensitivity was set to be 0.9 and specificity 0.3 with 1000 burn-ins for generating the estimated MCMC values and related histograms of regression coefficients.

Table 5.10: Summary of MCMC estimates for the least informative prior (Beta (5, 1))

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.526	0.061	0.005	0.888
β_1	0.5	0.466	0.084	0.010	0.642
β_2	0.5	0.540	0.066	0.005	0.934
β_3	0.5	0.399	0.126	0.017	0.544
r	0.05	0.051	0.023	0.000	0.933
δ^2	1.0	0.975	0.050	0.003	0.912
u	0.9	0.803	0.097	0.002	0.896
v	0.3	0.196	0.104	0.002	0.634

Table 5.11: Summary of MCMC estimates for non informative prior (Beta(1, 1))

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.528	0.062	0.005	0.822
β_1	0.5	0.405	0.122	0.017	0.754
β_2	0.5	0.488	0.058	0.004	0.986
β_3	0.5	0.507	0.077	0.008	0.865
r	0.05	0.052	0.028	0.000	0.855
δ^2	1.0	1.061	0.078	0.007	0.772
u	0.9	0.804	0.096	0.002	0.814
v	0.3	0.195	0.105	0.002	0.463

Table 5.12: Summary of MCMC estimates for the most informative prior (Beta(2, 5))

Parameter	**True value**	**Estimate**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.570	0.089	0.008	0.883
β_1	0.5	0.411	0.117	0.016	0.756
β_2	0.5	0.452	0.074	0.007	0.871
β_3	0.5	0.507	0.079	0.009	0.835
r	0.05	0.050	0.015	0.000	0.843
δ^2	1.0	0.930	0.081	0.006	0.677
u	0.9	0.800	0.100	0.002	0.455
v	0.3	0.199	0.101	0.002	0.562

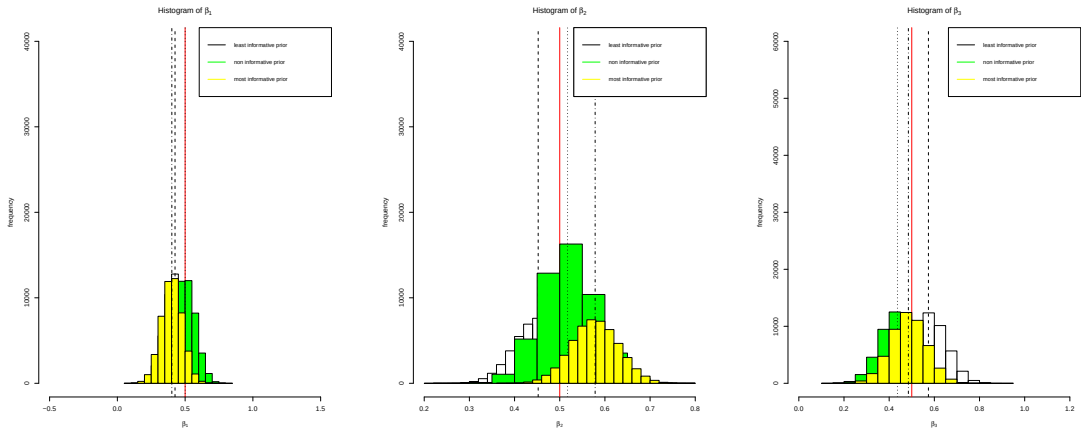

Figure 5.4: Histogram of the MCMC estimates for β_1, β_2 and β_3 with respect to different beta priors under rare probability. The vertical red solid line represents the true parameter value and the black, green and yellow histograms indicate estimates of β values for the least, non and the most informative priors, respectively.

Tables 5.10, 5.11 and 5.12 and Figure 5.4 represent the MCMC estimates and associated graphs. Interestingly, prior information for β_1 (the coefficient of erroneous variable), seem to have no improvement on the convergence on MCMC estimates. Convergence of estimated values for β_2 were not affected by the choice of prior for r, either. However, for β_2, the convergence process is satisfactory, implying that inference about the coefficient of the accurately measured variable is not affected by the error in another variable. In addition, convergence in the MCMC estimates for β_3 seem to require the most information from prior that is reflected in the histogram of β_3.

5.1.4 Sensitivity and Specificity

For analyzing the effect of sensitivity and specificity, we looked at two scenarios of (a) high sensitivity, low specificity ($u = 0.9$, $v = 0.3$) and (b) both low ($u = 0.3$, $v = 0.3$).

Again for this section, we considered r to be 0.05 with a sample of size 1000. All other parameters were set as the previous section. The following tables and graphs present the results.

Table 5.13: Summary of MCMC estimates for $u = 0.9$, $v = 0.3$

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.489	0.056	0.004	0.781
β_1	0.5	0.474	0.146	0.020	0.763
β_2	0.5	0.478	0.060	0.005	0.814
β_3	0.5	0.498	0.078	0.008	0.792
r	0.05	0.050	0.016	0.000	0.807
δ^2	1.0	0.934	0.037	0.003	0.825
u	0.9	0.799	0.101	0.002	0.681
v	0.3	0.200	0.100	0.002	0.727

Table 5.14: Summary of MCMC estimates for $u = 0.3$, $v = 0.3$

Parameter	**True value**	**Estimator**	**MSE**	**S.D.**	**Emperical 95% coverage probability**
β_0	0.5	0.467	0.062	0.005	0.672
β_1	0.5	0.371	0.112	0.025	0.685
β_2	0.5	0.547	0.071	0.006	0.793
β_3	0.5	0.726	0.110	0.014	0.548
r	0.05	0.046	0.033	0.001	0.583
δ^2	1.0	1.015	0.048	0.003	0.608
u	0.3	0.511	0.212	0.006	0.542
v	0.3	0.487	0.188	0.005	0.620

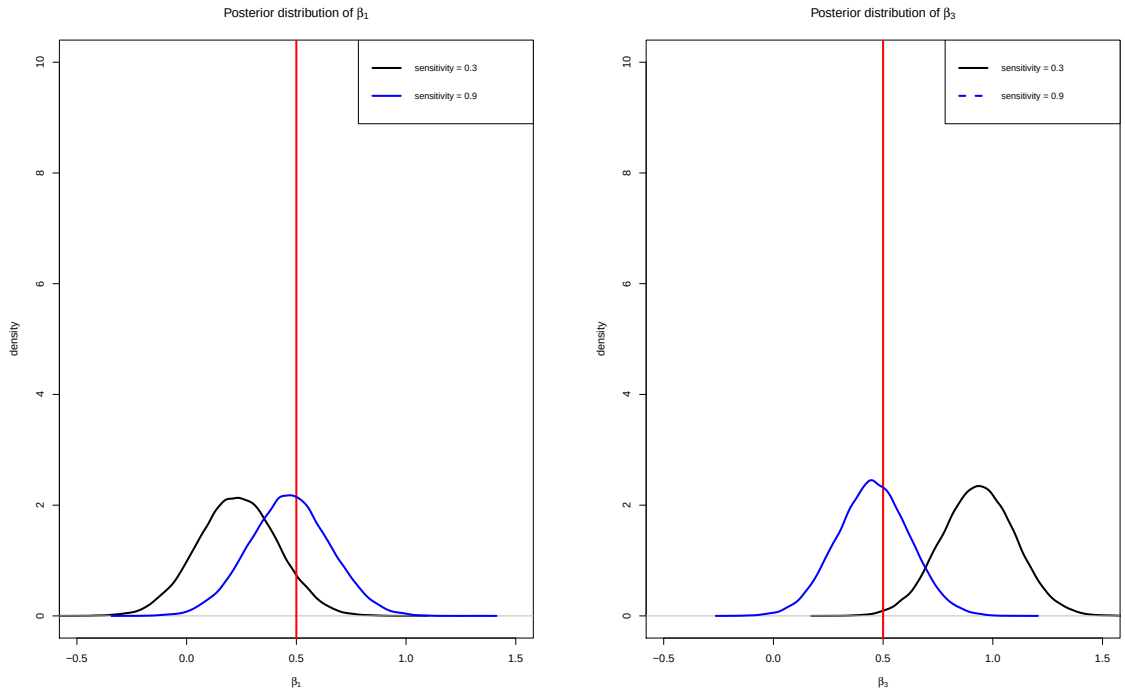

Figure 5.5: Posterior distribution of β_1 and β_3 with respect to different sensitivity under rare probability. The vertical red solid line represents the true parameter value and the black and blue curves indicates the distribution for sensitivity 0.3 and 0.9, respectively.

Keeping the specificity at 0.3, we changed the sensitivity from 0.3 to 0.9 and summarized the Bayesian estimates in tables 5.13 and 5.14 and in Figure 5.5. We observe that when we have lower number of success in the variable with lower rate of specificity and higher rate of sensitivity, the MCMC estimates perform well with lower MSE and S.D. value. However, considering both the sensitivity and specificity at a lower rate, we observe that the coefficient of the erroneous covariate and interaction term deviated from the true value. Moreover, the MSE and S.D. values were higher for this scenario.

Chapter 6

Discussion and Conclusion

In statistical models, presence of measurement errors in variables are common problems in practice. When we observe a data that is not measured correctly and the measurement errors are not taken into account, statistical inference is incorrect and provides a misleading conclusion. Therefore, it is important for the researcher to address measurement errors in order to obtain valid statistical inference. Moreover, presence of interaction terms in the model are common in many research areas. Erroneous variable incorporated with interaction, makes the analysis more complicated to deal with. One of the recent techniques is Bayesian methods that incorporates the prior knowledge about parameters. Only a few studies have implemented Bayesian techniques into interaction models with error in covariates.

Our primary goal was to monitor the behaviour of the Bayesian estimations of the model parameters in presence of measurement error for both discrete and continuous covariates. More specifically, we paid more attention to the behaviour of estimators for the coefficient of the interaction term in the models.

We started with the continuous case where no measurement error was considered. More specifically, in Chapter 2, we applied Bayesian techniques to the linear regression model

with a continuous random covariate without measurement error interacting with a nonrandom accurately measured covariate. We studied the behaviours of estimates under different scenarios when (a) number of Monte Carlo iterations and (b) sample size changed. Moreover, we analyzed the convergence of Markov Chains using graphical and testing hypothesis methods. We observed that while increasing both iteration numbers and the sample size, improved the convergence of the MC estimates, the coefficient of the interaction term required more iterations to perform as well as the other coefficients.

Moving to the mismeasured continuous covariate case, in Chapter 3, we studied the behaviour of the Bayesian estimators for the naive estimator that ignores the error in the error-prone covariate. We also considered validation subsample and replication as an adjustment for the bias caused by measurement error, where those estimates were also compared with the naive. Moreover, we studied the behaviours of estimates under different scenarios when (a) number of Monte Carlo iterations, (b) percentage of validation subsample and (c) magnitude of measurement error. We observed that generally, lowering measurement error as well as increasing number of iterations improved the convergence of the estimates, for the coefficient of the interaction term affected both the location and the width of the posterior densities, and therefore, the adjustment moved the posterior distributions closer to the true value with less variability. Moreover, 4% of the sample as the validated data was not as good as two replicates of error-prone covariate. However, increasing the validation subsample size to 20% made a significant improvement in the performance of the estimators. In application, however, there is a trade-off between cost and effect of larger sample.

For the discrete case, in Chapter 4, we first considered a linear regression model with an accurate discrete random covariate interacting with a nonrandom accurately measured covariate. More specifically, we studied the behaviours of Bayesian estimates under different scenarios when (a) number of Monte Carlo iterations for two cases of rare and

common probability, (b) sample size and (c) prior for the probability of success for the discrete random variable, changed. In here, we interestingly, observed that the non informative and most informative priors improved the performance of the estimator of the interaction term. However, for the coefficient of the nonrandom variable, the choice of prior did not make any significant difference.

Moving to misclassified discrete covariate, in Chapter 5, we studied the performance of the estimators for the naive one that ignores the misclassification in the error-prone covariate. Moreover, we studied the behaviours of estimates under different scenarios when (a) number of Monte Carlo iterations or two cases of rare and common probability, (b) sample size, (c) prior for the probability of success for the discrete random variable and (d) specificity and sensitivity, changed. The most interesting result of this chapter was to observe that for the rare case with lower rate of specificity and higher rate of sensitivity, the MCMC estimates perform well with lower MSE. However, considering both the sensitivity and specificity at a lower rate, we observed that the coefficient of the erroneous covariate and interaction term deviated from the true value.

Although our study included many interesting scenarios, there are still gaps that can be filled. Further investigation is required to evaluate, for example, the impact of choices of prior for the model coefficients, choices of distributions for the error-prone covariate on the estimation process.

Bibliography

[1] Berkson (1950) Are there two regressions?. *Journal of the american statistical association*; vol. 45

[2] Stephen P. Brooks (1998) Markov chain Monte Carlo method and its application. *The Statistician*; vol. 47

[3] Buzas J.S., Stefanski L.A., Tosteson T.D. (2014) Measurement Error. In: Ahrens W., Pigeot I. (eds) Handbook of Epidemiology.

[4] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, Ciprian M. Crainiceanu (2006) Measurement Error in Nonlinear Models: A Modern Perspective.

[5] Carroll, Raymond J and Li (1992) Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association*; vol. 87

[6] J. R. Cook and L. A. Stefanski (1994) Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*; vol. 89

[7] Gabriela Espino-Hernandez and Paul Gustafson (2011) Bayesian adjustment for measurement error in continuous exposures in an individually matched case-control study. *BMC Medical Research Methodology*; vol. 12

[8] D. Fouskakis, I. Ntzoufras, and D. Draper (2009) Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *The Annals of Applied Statistics*; vol. 3

[18] Hamra, Ghassan and MacLehose, Richard and Richardson, David (2013) Markov chain Monte Carlo: an introduction for epidemiologists. *International journal of epidemiology*; vol. 42

[10] Paul Gustafson (2003) Measurement error and misclassification in statistics and epidemiology.

[11] Gustafson (2004) Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments

[18] Gustafson (2005) On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*; vol. 20

[13] Gustafson, P. (2012) On the behaviour of Bayesian credible intervals in partially identified models. *Electronic Journal of Statistics*; vol. 6

[14] Gustafson, P., McCandless, L. (2014) Commentary: priors, parameters, and probability: a Bayesian perspective on sensitivity analysis. *Epidemiology*; vol. 25

[15] Holliday, Katelyn M and Avery, Christy L and Poole, Charles and McGraw, Kathleen and Williams, Ronald and Liao, Duanping and Smith, Richard L and Whitsel, Eric A (2014) Estimating personal exposures from ambient air-pollution measures: Using meta-analysis to assess measurement error. *Epidemiology (Cambridge, Mass.)*; vol. 25

[16] Lindley (1983) Theory and Practice of Bayesian Statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*; vol. 32

[17] Muff, Stefanie and Riebler, Andrea and Held, Leonhard and Rue (2015) Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*; vol. 64

[18] Pearson, K. (1902) On the mathematical theory of errors of judgment. *Phil. Trans. Roy. Soc. Lond. A*; vol. 198

[19] Richardson, S., Leblond, L., Jaussent, I. and Green, P. J. (2002) Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society*; vol. 165

[20] Wald (1940) The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, vol. 11

[21] Walters, Carl and Ludwig, Donald (1994) Calculation of Bayes posterior probability distributions for key population parameters. *Canadian Journal of Fisheries and Aquatic Sciences*; vol. 51

[22] Wayne A. Fuller (2008) Measurement Error Models.