Formal and Informal Approaches to Adjusting for Exposure Mismeasurement

by

Tian Shen

B.Sc., The University of New Brunswick, 2006

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

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2009

Abstract

In many research areas, measurement error frequently occurs when investigators are trying to analyze the relationship between exposure variables and response variable in observational studies. Severe problems can be caused by the mismeasured exposure variables, such as loss of power, biased estimators, and misleading conclusions. As the idea of measurement error is adopted by more and more researchers, how to adjust for such error becomes an interesting point to study. Two big barriers in solving the problems are as follows. First, the mechanism of measurement error (the existence and magnitude of the error) is always unknown to researchers. Sometimes only a small piece of information is available from previous studies. Moreover, the situation can be worsen when the study conditions are changed in the present study, which makes previous information not applicable. Second, some researchers may still argue about the consequences of ignoring measurement error due to its uncertainness. Thus, the adjustment for the mismeasurement turn to be a difficult, or impossible task.

In this thesis, we are studying situations where the binary response variable is precisely measured, but with a misclassified binary exposure or a mismeasured continuous exposure. We propose formal approaches to estimate unknown parameters under the non-differential assumption in both exposure conditions. The uncertain variance of measurement error in the continuous exposure case, or the probabilities of misclassification in the binary exposure case, are incorporated by our approaches. Then the posterior models are estimated via simulations generated by the Gibbs sampler and the Metropolis - Hasting algorithm.

Meanwhile, we compare our formal approach with the informal or naive approach in both continuous and exposure cases based on simulated datasets. Odds ratios on log scales are used in comparisons of formal and informal approaches when the exposure variable is binary or continuous. General speaking, our formal approaches result in better point estimators and less variability in estimation. Moreover, the 95% credible, or confidence intervals are able to capture the true values more than 90% of the time.

At the very end, we apply our ideas on the QRS dataset to seek consistent conclusions draws from simulated datasets and real world datasets, and we are able to claim that overall our formal approaches do a better job regardless of the type of the exposure variable.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First of all, I would like to give special thanks to my supervisor, Professor Paul Gustafson, since without his thoughtful support and guidance, I would not able to complete this thesis. It was an honor to work with him and he is the best supervisor that one could ever ask for. I would also thank Professor Lang Wu, for agreeing to be my second reader and also for his excellent teaching throughout my study period.

I express sincere gratitude to Professors John Pekau, Ruben Zamar, Arnaud Doucet, Michael Schulzer, Jiahua Chen and Matias Salibián-Barrera for their constant support and outstanding teaching.

I am grateful to everyone, who makes the department a wonderful place to study. Special thanks to Mike Danilov, Ehsan Karim, Derrick Lee, Hernan Epstein, Xu Chen , Liu Zhong and Xu Wang for their help and friendship, and to Peggy Ng, Viena Tran, Elaine Salameh for their support and understanding of my work.

Tian Shen

Vancouver, BC, July 2009

To My lovely Parents

Chapter 1

Introduction

In many research areas, statistical methods are used to analyze the relationships between two or more variables. For example, in the epidemiology area, statistical models are used to understand or study the relationship between an outcome variable Y and an explanatory variable X. For instance Y can be presence or absence of heart disease, and X can be presence or absence of smoking, where Y and X are both binary variables, or X can be the blood pressure, which is a continuous variable. There are four types of designs that are most often applied, which are the cross-sectional study, cohort study, case-control study and randomized controlled clinical trial. In this thesis, we are focused on the case-control study, which is also referred to as a retrospective study. We randomly select subjects from "case" and "control" groups, then compare the health outcomes for the two groups based on selected subjects. The explanatory variable X is often measured by some instruments. When X is precisely measured, the instrument is called a gold standard. However, due to the high cost and lack of such precise instrument, X is often measured imprecisely. If X is a categorical variable, imprecise measurements are called misclassifications, while if X is a continuous variable, they are called measurement errors. In this thesis, we are working on both discrete and continuous cases of X.

1.1 Misclassification and Measurement Error

Generally speaking, misclassification means grouping a subject into a wrong category . For example, a person who smoked one pack of cigarette in the day of the experiment might be accidently grouped as a heavy smoker while this person might barely smoke in other days. Thus, rather than recording X itself, a surrogate variable X^* is often

recorded instead. The misclassification probability (say p_{ij}) defines the probability of classifying a subject into group i while its true status is in group j.

When X is a continuous variable, by the definition of the classical error model in the measurement error literature, a surrogate variable X^* is the linear term of X plus an error term. For example, a recorded patient's blood pressure might be higher than his/her true values due the equipment error. As defined in Chapter 3 in this thesis, $X^* = X + Z$, where Z is the error term and E($Z|X$)=0.

Measurement error and misclassification can be categorized into differential and non-differential cases. If the distribution of the surrogate variable X^* only depends on the true value X but not the health outcome Y, then the mismeasurement is classified as non-differential. Otherwise, the mismeasurement is categorized as differential.

Due to the misclassification and measurement error , usually we only have precisely measured health outcome Y , the surrogate variable X^* and some other precisely measured covariates U in the data. Since the goal of most studies is to understand the relationship between X and Y, conclusions obtained from X^* and Y instead could be very misleading. Thus, the study of measurement error and misclassification is significantly important.

1.2 Overview of Current Available Methods

In the literature, especially in the biomedical research, many methods were proposed to deal with misclassification and measurement error. Barron (1977) proposed the matrix method that estimate the expectation of cell counts by using their margins, and the odds ratio can be estimated later on based on the cell counts. By reparameterizing the misclassification, Marshall (1990) presented the inverse matrix method to retrieve the odds ratio. However, Lyles (2002) pointed out that the inverse matrix method is just a maximum likelihood estimate method that corrects the biased odds ratio due to misclassification. The efficiency of the matrix method and inverse matrix method were compared under the assumption of differential misclassification (the degree of measurement error is different across different groups) by Morrissey and Spiegelman (1999). They concluded

that the inverse matrix is more efficient than the matrix method; nevertheless, the sensitivity, specificity, and probability of exposure are some key determinants of the efficiency. Later, other methods like simulation extrapolation (SIMEX) and logistic regression model are also developed to approach the misclassification problem (Kuchenhoff, Mwalili, and Lesaffre, 2006; Skrondal and Rabe-Hesketh, 2004). With the improvement of computational capability of computers and enhanced simulation techniques, the Bayesian analysis becomes another prospective method to study the misclassification problem.

Carroll, Ruppert, Stefanski, and Cainiceanu (2006) grouped methods that deal with measurement error into functional and structural models, where X is considered as fixed or random with minimum assumptions of distributions in the functional models while X is considered as random variables in the structural models. Two general methods used in the functional model category are the regression calibration (Pierce and Kellerer, 2004) and SIMEX (Cook and Stefanski, 1994) methods. Carroll, Ruppert, Stefanski, and Cainiceanu (2006) stated that even though these two methods are very simple, they are only consistent in some special cases such as linear regression and they have limited contributions in reducing the bias caused by the measurement error. Disregarding some limitations of the regression calibration and SIMEX methods, they are still widely used since both of them are very easy to implement by using the existing software packages, which reduces the potential difficulties in the analysis set-up part. In the structural method category, the Expectation-Maximization algorithm in the frequentist perspective and Markov chain Monte Carlo algorithm in the Bayesian perspective are two useful algorithms to solve the measurement error problems.

Bayesian methods are capable of dealing with biases induced by both misclassification and measurement error. One great advantage of Bayesian methods is that they can fully incorporate the uncertainties of parameters. Though fully specified models are often required, this kind of information plus some knowledge of mismeasurement are often

available to medical researchers before they conduct studies. When an observed dataset is available, the natural existence of prior information make Bayesian analysis an appealing method, since inference now can be made through the prior and present information. In this thesis, we are focus on using the Gibbs sampler and Metropolis-Hastings algorithm in Bayesian methods to solve the misclassification and measurement error problems.

1.3 Bayesian Analysis

Bayesian inference is statistical inference in which data are used to update or to newly infer quantities that are observed or wished to learn by using probability model. The "combination" of Markov Chain Monte Carlo (MCMC) algorithm and Bayesian inferences is often used in solving the mismeasurement related problems.

1.3.1 Bayes Rule

To understand the Bayesian analysis, it is essential to understand the fundamental principle of the analysis - Bayes rule. Assume there is a independently and identically distributed dataset $y = (y_1, y_2, ..., y_n)$ with unknown parameter θ and following distribution f_θ. Bayesian analysis is to conclude a parameter θ based on the observed data, so we denote the conditional distribution on the observed data as $f(\theta|y)$, which is called the posterior distribution. Meanwhile, the sampling distribution of the data set is:

$$f(y|\theta) = \prod_{i=1}^{n} f_\theta(y_i)$$

Then, a joint probability distribution of θ and y can be defined as:

$$f(\theta, y) = f(\theta) \times f(y|\theta)$$

where $f(\theta)$ refers the prior distribution of θ. Applying the basic property of conditional distribution (Bayes' rule), the posterior distribution turns to:

$$f(\theta|y) = \frac{f(\theta, y)}{f(y)} = \frac{f(\theta) \times f(y|\theta)}{f(y)} \tag{1.1}$$

Since the $f(y)$ is a constant term, independent of the parameter θ, we can express the unnormalized posterior density as $f(\theta|y) \propto f(\theta) \times f(y|\theta)$.

To estimate the parameter θ, we can calculate the expected value of θ under the posterior density (Gelman et al., 2004) by taking the integral:

$$\hat{\theta} = \int \theta f(\theta|y) d\theta = \frac{\int \theta f(\theta) f(y|\theta) d\theta}{\int f(\theta) f(y|\theta) d\theta}$$

If there is a closed form of the posterior distribution, the estimation of θ would be very easy to carry out . However, sometimes there is no closed form of the posterior density, we have to use the other numerical methods, such as Markov Chain Monte Carlo algorithms.

1.3.2 Prior Distribution

As an "absent" term in classical statistical analysis, the prior distribution plays a major role in the Bayesian analysis. There are mainly two types of prior distributions: informative priors and non-informative priors. A non-informative prior also can be called a "flat" prior that has a limited impact on the posterior distribution. Researchers use non-informative priors usually because they don't have very much knowledge about the parameter in the previous study or they don't want the inference to be greatly affected by some external sources. Moreover, the determination of "flat" prior is not trivial since everyone has a different definition of "flat".

The informative prior, information that gained from previous study , on the other hand, may provide a strong influence on the posterior distribution. However, the influence is somehow related to the sample size, number of MCMC iterations and the form of the prior. In most cases, strong priors are not necessary when the sample size is big.

Nevertheless, in some cases, the priors are needed to be strong regardless of the sample size.

1.3.3 Markov Chain Monte Carlo Algorithm

The computation of the posterior distribution is always a problem since the density of the posterior sometimes is very complicated or even high dimensional. Because of this problem, the usage of Bayesian analysis has been very limited. When the MCMC algorithm was first applied in 1990, the limitation of Bayesian analysis then vanished and Bayesian methods becomes increasingly popular.

In the following two sections, two particular MCMC algorithms are introduced: Metropolis-Hastings algorithm (Greenberg and Chib, 1995) and Gibbs sampler (Casella and George, 1992).

Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is an iteration method for generating a sequence of samples from a probability distribution when directly sampling is difficult or impossible. It uses an acceptance/rejection rule to cover the target distribution. The basic algorithm follows:

1. Randomly choose a starting point θ^0
2. For the i^{th}iteration from i=1,2,...,
   (a) Sample a proposal θ^* from a jumping distribution at iteration i, $p(\theta^*|\theta^{i-1})$
   (b) Calculate the ratio of densities,

$$r = \frac{f(\theta^*)p(\theta^{i-1}|\theta^*)}{f(\theta^{i-1})p(\theta^*|\theta^{i-1})}$$

   where $f(\theta)$ is the target distribution.

(c) Set

$$\theta^i = \begin{cases} \theta^* & \text{with the probability } \min(r, 1) \\ \theta^{i-1} & \text{otherwise} \end{cases}$$

Note that the acceptance rate shouldn't be close to 0% or 100%, otherwise, the random walk would either move too slowly (cover the target population too slowly) or stand still for too long (likely to stay in certain regions). Thus, proper adjustments of the jumping distribution sometimes are necessary.

Gibbs sampler

The Gibbs sampler is a special case of MH algorithm. It is applied when the explicit form of the joint distribution is unknown but the conditional distribution of each variable is known. Assume we divide the parameter vector θ into p components, say $\theta = (\theta_1, \theta_2, ..., \theta_p)$. Each iteration of Gibbs sampler cycles thought all the components of θ. Each component is drawn based on the values of all others. Thus, the algorithm that generates a Markov chain at iteration i is:

1. Randomly set the initial values for all the components $\theta = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$.

2. For the i^{th}iteration from i=1,2,...,
   for j=1...p, sample $\theta_j^{(i)}$ from the distribution of $f(\theta_j|\theta_{-j}^{(i)})$,
   where $\theta_{-j}^{(i)}$ represents all the components of θ, except for θ_j, at their current values:
   $\theta_{-j}^{(i)} = (\theta_1^{(i)}, \theta_1^{(i)}, ..., \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, ..., \theta_p^{(i-1)})$.

Though the above two MCMC algorithms looks very simple, sometime the technical details can be very challenging. In this thesis, we will apply Gibbs sampler algorithm in Chapter 2 and Metropolis-Hastings algorithm in Chapter 3.

Chapter 2

Simulation Study for Categorical Exposure Variable

2.1 Introduction

Assume a researcher is interesting in study the relationship between a binary exposure variable X and a continuous response variable Y. The exposure variable X is often coded as 0 or 1, where $X = 0$ refers to "unexposed" and $X = 1$ refers to "exposed" in most epidemiological situation. Instead of observing X, a surrogate variable X^* is measured. Under the non-differential misclassification assumption, X and X^* are conditionally independent of Y, and the specificity and sensitivity are used to measure the magnitude of the misclassification (Gustafson, 2004). Then, the sensitivity $SN = Pr(X^* = 1|X = 1)$ is the probability of correctly classifying a truly "exposed" subject, and the specificity $SP = Pr(X^* = 0|X = 0)$ is the probability of correctly classifying a truly "unexposed" subject. In the following subsections, we will introduce two approaches to analyze the relationship between discrete exposure variables and health outcome Y.

2.2 2× 3 Table–Formal Analysis

Though there are only two conditions of the true exposure X, Yes or No, sometimes the surrogate variable X^* has three conditions instead as: Unlikely, Maybe and Likely, due to possible instrument error (as displayed in Table 2.1).

		Assessed Exposure		
		Unlikely	Maybe	Likely
True Exposure	No	p_{00}	p_{01}	p_{02}
	Yes	p_{10}	p_{11}	p_{12}

Table 2.1: *2×3 table due to the misclassification and measurement error*

Note that the p_{ij} in the table defines the probability of classifying a subject into group j while its true status is in group i.

If the truly exposed condition, X, is known, then the exact relationship (referred as a "true" result) of X and Y is able to be analyzed. However, in reality, X is often inaccessible, and researchers only know its surrogate, X^*. Even though, researchers are analyzing the relationship between X^* and Y, they also tend to conclude it as the relationship between X and Y. It is very dangerous to make such conclusion since it can be very biased. The first analysis method is termed as a formal analysis: analysis that acknowledge the existences of 2×3 table (Table 2.1) structure in the exposure condition. Specifically, the analysis is carried out based on Table 2.2.

		Exposure		
		Unlikely	Maybe	Likely
Health Outcome	controls	n_{00}	n_{01}	n_{02}
	cases	n_{10}	n_{11}	n_{12}

Table 2.2: *A 2×3 table for formal analysis*

where n_{ij} in the table is the number of subjects that fall in the condition (e.g. n_{00} is the number of subjects that "Unlikely" have the exposure in controls). In the literature, there are fewer studies that involves the formal analysis, which makes it an interesting point to study. The second approach is termed as an informal analysis: analysis that tend to ignore the 2×3 table structure in the exposure condition. More details of informal analysis are going to be talked about in the next section.

2.3 Informal Analysis

When we assume the mismeasurement is non-differential, sensitivity and specificity of X^* for X can be used to described the magnitude of the misclassification. The closer the values of SN and SP are to one, the less severe the misclassification is.

As stated in Gustafson (2004), when the probability of the exposure is very rare, the effects of misclassification worsens much more with the decrease of specificity than the decreases of sensitivity, which means the impact of low specificity will be bigger than the impact of low sensitivity in a further analysis. This attracts particular attention in the epidemiological area for the study of rare disease since it implies that when the exposure is very rare, the analysis can be fully misleading even with a very small effects of misclassification. Thus, when some epidemiologists realize that they have the 2×3 table structure (as Table 2.1) in hand, they tend to group the "Maybe" and "Unlikely" groups together to form a 2×2 table as in Table 2.3: where $q_{00} = p_{00} + p_{01}$ and $q_{10} = p_{10} + p_{11}$.

	Assessed Exposure		
True Exposure		No	Yes
	No	q_{00}	q_{01}
	Yes	q_{10}	q_{11}

Table 2.3: *2×2 table by epidemiologists' rule*

They prefer such grouping so that more subjects are classified as unexposed, and this leads an increase of probability that a true negative is correctly classified, which means a large SP value. In such way, a low specificity could be avoided so that a small effect of misclassification won't lead a huge impact in the further analysis.

In the literature, most analyses are constructed based on the ignorance of the structure of Table 2.1. Thus, they are based on Table 2.4, and we would like to refer such analysis as informal analysis in the rest of the Chapter.

	Exposure		
		No	Yes
Health outcome	controls	$n_{00}+n_{01}$	n_{02}
	cases	$n_{10}+n_{11}$	n_{12}

Table 2.4: *A 2×2 table for informal analysis*

2.4 Odds Ratio

Suppose we have two groups of subjects, controls and cases, and denoted the number of subjects in each group as n_0 and n_1. Let r_0 denote the prevalence of exposure in the control group and r_1 denote the prevalence of exposure in the case group, i.e. $r_0 = Pr(X = 1|Y = 0)$ and $r_1 = Pr(X = 1|Y = 1)$. Thus, the odds ratio, which defines the relationship between X and Y is formed as:

$$\Psi = \frac{\frac{r_1}{1-r_1}}{\frac{r_0}{1-r_0}}. \tag{2.1}$$

In the informal analysis, the odds ratio is usually calculated as:

$$\widehat{OR}_{informal} = \frac{n_{12} \times (n_{00}+n_{01})}{n_{02} \times (n_{10}+n_{11})},$$

where the standard error of log odds ratio is formulated as:

$$se = \left(\frac{1}{n_{00}+n_{01}} + \frac{1}{n_{02}} + \frac{1}{n_{10}+n_{11}} + \frac{1}{n_{12}}\right)^{1/2}.$$

In the formal analysis, there are three different ways to calculate the odds ratio. Since the MCMC algorithm estimates the prevalence in both case and control groups (r_0 and r_1) at each iteration, they are going to be updated every time. Thus, the first way is to find the posterior mean of odds ratios, which is to calculate the odds ratio at each MCMC iteration, and then take the average of them. The second way is to find the posterior average of r_0, r_1 of each iteration, then use formula (2.1) to find the odds ratio. The

third way is to find the posterior mean of log odds ratio, which is to calculate average of the log of odds ratio at each iteration first and then transfer it by using exponential function. Mathematically, those three odds ratio can be written as:

$$\widehat{OR}^i = \frac{\frac{r_1^i}{1-r_1^i}}{\frac{r_0^i}{1-r_0^i}},$$

$$\widehat{OR}_1 = \frac{1}{m}\sum_i^m \widehat{OR}^i,$$

where i refers to the i^{th} iteration, and m is the total number of iterations.

$$\hat{r_0} = \frac{1}{m}\sum_i^m \hat{r_0}^i \ , \hat{r_1} = \frac{1}{m}\sum_i^m \hat{r_1}^i \tag{2.2}$$

$$\widehat{OR}_2 = \frac{\frac{\hat{r_1}}{1-\hat{r_1}}}{\frac{\hat{r_0}}{1-\hat{r_0}}}$$

where r_0, r_1 are calculated from equation (2.2). And also,

$$\widehat{OR}_3 = exp\left(\frac{1}{m}\sum_i^m log(\widehat{OR}^i)\right)$$

We are interested in comparing the odds ratios in both analysis. Since the odds ratio always falls into a very skewed distribution, we tend to compare the informal odds ratio and the formal odds ratio in the log scale, such that the distribution will be more symmetric. We are going to focus on comparing the $\widehat{OR}_{informal}$ with $\widehat{OR}_3$ in the log scale with respect to their point estimator and confidence/credible intervals as :

$$\log(\widehat{OR}_{informal}) \ = \ \log\left(\frac{n_{12} \times (n_{00} + n_{01})}{n_{02} \times (n_{10} + n_{11})}\right),$$

with its 95% confidence interval as:

$$CI_{informal} : \left(\log(\widehat{OR}_{informal}) - 1.96 * se, \log(\widehat{OR}_{informal}) + 1.96 * se\right).$$

The log formal odds ratio is as:

$$\log \widehat{OR}_3 = \frac{1}{m}\sum_{i}^{m} log(\widehat{OR}^i).$$

Note that the 95% credible interval for $\log \widehat{OR}_3$ can be simply obtained by finding the 2.5% and 97.5% quantiles of all $\log \widehat{OR}_3$ that obtained from each iteration.

In order to evaluate the performance of our proposed formal approach, we conduct simulation studies under three cases as: when the probability of classify, p_{ij} are known; when we only have prior information on p_{ij}; when we have some validation data, i.e. we have some subjects have both X and X^* measured.

2.5 Case 1: When We Know p_{ij}s

In some cases, researchers may know the probabilities of classifying the assessed exposure to true exposure from previous experiments, i.e. the p_{ij}s are known. Under this condition, the posterior density is expressed as:

$$\begin{aligned} & f(r_0, r_1, X_1 ... X_n, | Y_1 ... Y_j, X_1^* ... X_n^*) \\ & = \prod_j r_0^{X_j(1-Y_j)}(1-r_0)^{(1-X_j)(1-Y_j)} r_1^{X_j Y_j}(1-r_1)^{(1-X_j)Y_j} \\ & \times \prod_j p_{00}^{I(X_j^*=0)(1-X_j)} p_{01}^{I(X_j^*=1)(1-X_j)} p_{02}^{I(X_j^*=2)(1-X_j)} \\ & \times \prod_j p_{10}^{I(X_j^*=0)X_j} p_{11}^{I(X_j^*=1)X_j} p_{12}^{I(X_j^*=2)X_j}, \end{aligned}$$

where I is an indicator function. We assume the prior distribution of r_0 and r_1 are uniform (the same assumption carries in the following two cases), and then we use the Gibbs sampling method to update the unknowns, r_0 and r_1 since their posterior distributions have familiar distributions that we can recognize, namely Beta distributions. The posterior distribution of X_j is viewed as:

$$f(X_j|r_0, r_1, X_1, ...X_{j-1}, X_{j+1}, ...X_n, Y_1...Y_j, X_1^*...X_j^*) \propto \frac{b}{b+a}$$

where

$$\begin{cases} a = (1-r_0)^{1-Y_j}(1-r_1)^{Y_j} p_{00}^{I(X_j^*=0)} p_{01}^{I(X_j^*=1)} p_{02}^{I(X_j^*=2)} \\ b = r_0^{1-Y_j} r_1^{Y_j} p_{10}^{I(X_j^*=0)} p_{11}^{I(X_j^*=1)} p_{12}^{I(X_j^*=2)} \end{cases} ,$$

Thus, at each iteration of MCMC algorithm, the probability of getting $X_j = 1$ by given everything else "known" is $b/(a+b)$.

2.6 Case 2: When p_{ij}s are Unknown

Though, in the previous section, we studied the case that p_{ij}s are available from other experiments, in reality, those values are often unknown. The maximum information that we have on them are some knowledge of their priors. We choose a Dirichlet distribution as the prior distribution for p_{ij}s for two reasons. First, the summation of p_{0j} and the summation of p_{1j} are both equal to 1, i.e. $p_{00} + p_{01} + p_{02} = 1$ and $p_{10} + p_{11} + p_{12} = 1$. Second, the Dirichlet distribution has the property of being conjugate, which means the posterior distribution of p_{ij}s would also come from the Dirichlet distribution family. Thus, by choosing Dirichlet distribution as the prior, the MCMC algorithm is easy to compute the updated p_{ij}s at each iteration, and the results are also easy to be interpreted.

Hence, we have:

$$p_{00}, p_{01}, p_{02} \sim Dir(c_{00}, c_{01}, c_{02}),$$

$$p_{10}, p_{11}, p_{12} \sim Dir(c_{10}, c_{11}, c_{12}).$$

Then, the posterior distribution is changed to:

$$\begin{aligned}
&f(r_0, r_1, X_1, ...X_n, p_{00}, p_{01}, p_{02}, p_{10}, p_{11}, p_{12} | Y_1, ...Y_n, X_1^*, ...X_n^*) \\
&\quad \propto \prod_j r_0^{X_j(1-Y_j)}(1-r_0)^{(1-X_j)(1-Y_j)} r_1^{X_jY_j}(1-r_1)^{(1-X_j)Y_j} \\
&\quad \times \prod_j p_{00}^{I(X_j^*=0)(1-X_j)} p_{01}^{I(X_j^*=1)(1-X_j)} p_{02}^{I(X_j^*=2)(1-X_j)} \\
&\quad \times \prod_j p_{10}^{I(X_j^*=0)X_j} p_{11}^{I(X_j^*=1)X_j} p_{12}^{I(X_j^*=2)X_j} \\
&\quad \times p_{00}^{c_{00}-1} p_{01}^{c_{01}-1} p_{02}^{c_{02}-1} * p_{10}^{c_{10}-1} p_{11}^{c_{11}-1} p_{12}^{c_{12}-1} \\
&\quad \propto \prod_j r_0^{X_j(1-Y_j)}(1-r_0)^{(1-X_j)(1-Y_j)} r_1^{X_jY_j}(1-r_1)^{(1-X_j)Y_j} \\
&\quad \times \prod_j p_{00}^{(I(X_j^*=0)(1-X_j)+c_{00}-1)} p_{01}^{(I(X_j^*=1)(1-X_j)+c_{01}-1]} p_{02}^{(I(X_j^*=2)(1-X_j)+c_{02}-1)} \\
&\quad \times \prod_j p_{10}^{(I(X_j^*=0)X_j+c_{10}-1)} p_{11}^{(I(X_j^*=1)X_j+c_{11}-1)} p_{12}^{(I(X_j^*=2)X_j+c_{12}-1)} \\
&\quad \propto r_0^{\sum_j X_j(1-Y_j)}(1-r_0)^{\sum_j(1-X_j)(1-Y_j)} r_1^{\sum_j X_jY_j}(1-r_1)^{\sum_j(1-X_j)Y_j} \\
&\quad \times p_{00}^{\sum_j I(X_j^*=0)(1-X_j)+c_{00}-1} p_{01}^{\sum_j I(X_j^*=1)(1-X_j)+c_{01}-1} p_{02}^{\sum_j I(X_j^*=2)(1-X_j)+c_{02}-1)} \\
&\quad \times p_{10}^{\sum_j I(X_j^*=0)X_j+c_{10}-1} p_{11}^{\sum_j I(X_j^*=1)X_j+c_{11}-1} p_{12}^{\sum_j I(X_j^*=2)X_j+c_{12}-1}
\end{aligned}$$

Note that $c_{00}, c_{01}, c_{02}, c_{10}, c_{11}$ and c_{12} are called hyper-parameters, and the specific values assigned will be discussed in the results section.

2.7 Case 3: Validation Data

Sometimes, the true exposure variable X might be possible to capture, but it is too expensive to get for all the subjects. As a result, only a small proportion of the subjects have the complete information of X, X^* and Y, whereas the majority of the subjects don't have the precisely measured exposure status, X. Table 2.5 presents the structure of the validation sample and the incomplete main data. While all counts corresponding to

Validation Data					**Main Data**			
		X*				X*		
		Unlikely	Maybe	Likely		Unlikely	Maybe	Likely
Y=0	X=0				Y=0			
	X=1							
Y=1	X=0				Y=1			
	X=1							

Table 2.5: *Validation data and main data*

levels of $X, Y, X*$ are fully recorded in the validation data, only counts that correspond to Y, X^* are recorded in the main data. We want to use the information from the validation data to impute X for the main data and to make inference on X and Y.
Our new posterior density is like:

$$
\begin{aligned}
& f(X_1, ...X_m, p_{00}, p_{01}, p_{02}, p_{10}, p_{11}, p_{12}, r_0, r_1 | X_{m+1}, ...X_n, X_1^*, ...X_n^*, Y_1, ...Y_n) \\
&= \prod_{j=1}^{m} f(Y_j|X_j) \times \prod_{j} f(X_j^*|X_j) \times \prod_{j} f(X_j) \prod_{m+1}^{j=n} f(Y_j|X_j) \\
&\quad \times \prod_{j} f(X_j^*|X_j) \times \prod_{j} f(X_j) \times f(p_{00}, p_{01}, p_{02}, p_{10}, p_{11}, p_{12}) \times f(r_0) \times f(r_1),
\end{aligned}
$$

where $j = 1, ...m$ is the non-validation data part and $j = m + 1, ...n$ is the validation data part.

The simulation process for this case does not change a lot regarding the change of the posterior distribution, and the only difference is that "known" X values in the validation data do not need to be updated, where the "unknown" X values, r_0, r_1 and $p_{ij}s$ are

updated the same as in the previous case.

2.8 Results

In order to gain information about the performance of MCMC algorithms in all three cases, the MCMC trace-plots, posterior mean, 95% equal-tail credible interval, estimated bias, estimated mean square error of each unknown parameters are checked in the following subsections. Moreover, when the prevalence in control and cases are small, i.e. r_0 and r_1 are relatively small, the odds ratio of "formal" and "informal" are compared later on to assess their performance.

In all cases, two sets of the prevalence are used as (1): $r_0 = 0.2, r_1 = 0.25$ and (2): $r_0 = 0.7, r_1 = 0.4$, and each one is combined with the "true" probability of classifying values as $p_{00} = 0.5, p_{01} = 0.3, p_{02} = 0.2$ and $p_{10} = 0.1, p_{11} = 0.3, p_{12} = 0.6$. Any hyper-parameters used in the specific cases will be defined in the later subsections along with detailed information of each case scenarios. Moreover, two simulation studies are performed regarding each scenario in each case. The first one focuses on studying estimation from one sample. The second concentrates on studying the sampling distributions of each estimator across 100 simulated datasets.

2.8.1 Results for Case 1

Two scenarios we have in this case are:

- Scenario 1: $(r_0, r_1) = (0.2, 0.25)$, true odds ratio=1.33, $(p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$, $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$;
- Scenario 2: $(r_0, r_1) = (0.7, 0.4)$, true odds ratio=0.283, $(p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$, $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

Two simulation studies are carried for each scenario, where studies for the first scenario will be talked about in details as an example. Note that the same procedures are applied

to the second scenario as well.

In the first study, a dataset of size 4000 (2000 controls and 2000 cases) was generated based on the given "true" value, $(r_0, r_1) = (0.2, 0.25)$, $(p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$, and $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. Then, we use 21000 iterations, where the first 1000 is the burn-in period to update unknown parameters. The choice of burn-in period is make based on visual inspection. Figure 2.1 shows the traceplot of MCMC algorithm for r_0 and r_1 after the first 1000 burn-in period. It shows that the Markov chains moving smoothly within the target area.

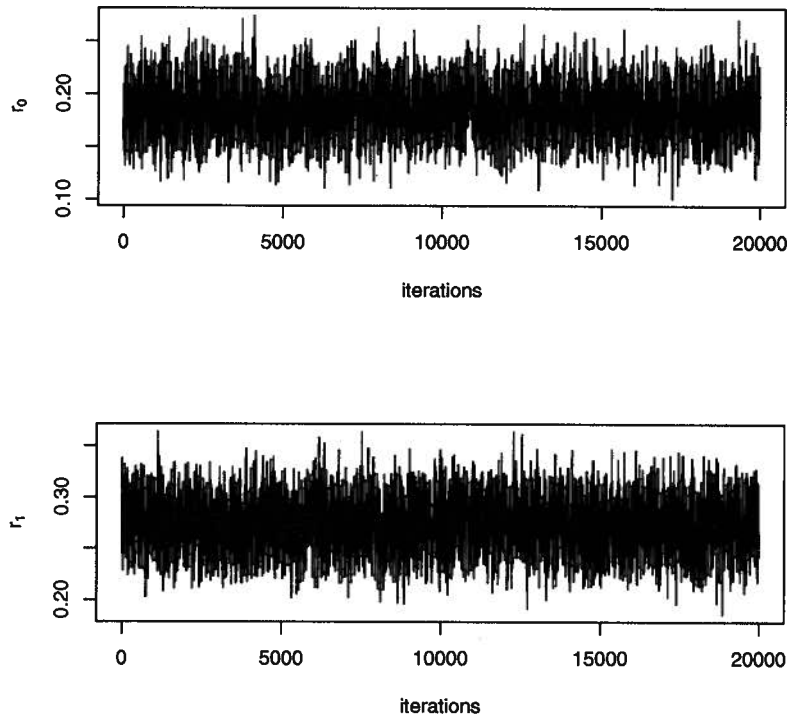

Figure 2.1: *Traceplots of r_0 and r_1 from MCMC algorithm in scenario 1 case 1. The traceplots show the 20000 iterations after 1000 burn-in period. The true values of r_0 and r_1 are 0.2 and 0.25 respectively.*

Table 2.6 shows the true values, posterior means and the 95% credible intervals for r_0 and r_1. Both 95% credible interval of r_0 and r_1 covers the "given" values, which indicates

that for this particular generated dataset, our approach works well.

	true value	posterior mean	95% CI
r_0	0.2	0.18	(0.15, 0.23)
r_1	0.25	0.27	(0.23, 0.31)

Table 2.6: *True values, posterior means, 95% credible intervals of* r_0 *and* r_1. *These are results from the first simulation study (one dataset simulation) for scenario 1 in case 1.*

The second study will repeats the first study 100 times, which enable us to investigate the sampling distributions of r_0 and r_1. Figure 2.2 is the histogram of 100 posterior means of r_0 and r_1. It demonstrates that the sampling distribution of $\hat{r_0}$ and $\hat{r_1}$ are approximately normally distributed and centered around the "true" values, majority values are around the "true" values.

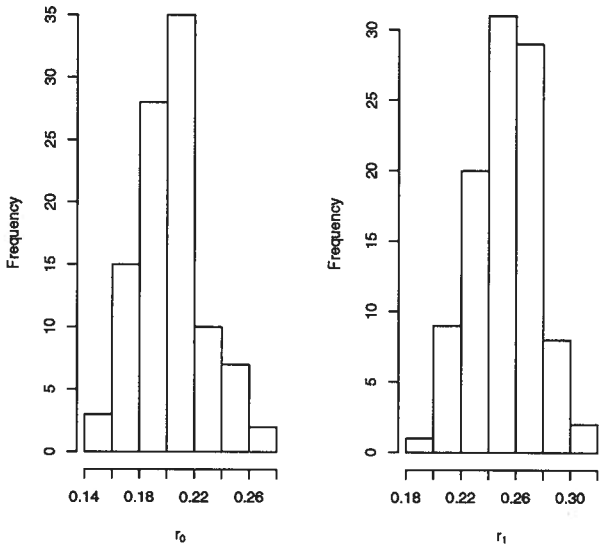

Figure 2.2: *Histogram of 100 posterior means of r_0 and r_1 in the second simulation study for scenario 1 case 1. The "true" values of r_0 and r_1 are 0.2 and 0.25 respectively.*

Table 2.7 confirms our observation by showing the 95% credible interval coverage rates are very high and the average lengths of the credible intervals are very small.

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.0024	0.0025	91	0.09
r_1	0.0028	0.0023	95	0.091

Table 2.7: *Bias, mean square error (MSE), coverage of 95 % CI and the average width of r_0 and r_1 for scenario 1 case 1. All results are based on 100 datasets, and their true values are 0.2 and 0.25 respectively.*

Since desirable results are obtained from the statistical procedures and stabilized iter-

ations are observed from the traceplot, it is reasonable to conclude that our approach works well when the prevalence for both control and cases are relatively small.

Table 2.8 shows the first simulation study result for scenario 2, where $r_0 = 0.7$ and $r_1 = 0.4$ and p_{ij} values remain the same. Again, the result shows that for the particular

	true value	posterior mean	95% CI
r_0	0.7	0.69	(0.63,0.72)
r_1	0.4	0.41	(0.36, 0.45)

Table 2.8: *True values, posterior means, 95% credible intervals of r_0 and r_1. These are results from the first simulation study (one dataset simulation) for scenario 2 in case 1.*

generated dataset, we are able to get reasonable estimators. However, in order to know how the model works in general, we need to review the results from the second simulation study.

Figure 2.3 and Table 2.9 are histogram and statistical results from the second simulation study for scenario 2 in case 1. All results from two studies for scenario 2 suggest that our proposed approach works equally well when the prevalences are considerably larger.

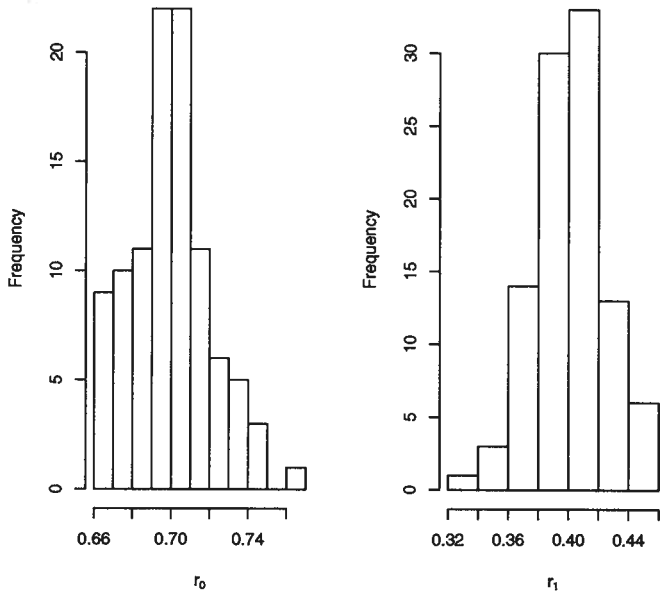

Figure 2.3: *Histogram of 100 posterior means of r_0 and r_1 in the second simulation study for scenario 2 case 1. The true values of r_0 and r_1 are 0.7 and 0.4 respectively.*

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.000068	0.0021	98	0.085
r_1	0.00074	0.0023	94	0.092

Table 2.9: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of r_0 and r_1 for scenario 2 case 1. All results are based on 100 datasets, and their true values are 0.7 and 0.4 respectively.*

2.8.2 Results for Case 2

In this case, there are also two scenarios as:

- Scenario 1: $(r_0, r_1) = (0.2, 0.25)$, true odds ratio=1.33, $(p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$, $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$;
- Scenario 2 : $(r_0, r_1) = (0.7.0.4)$, true odds ratio=1.33, $(p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$, $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

Remember that in this case, since p_{ij}s are unknown, we need to specify prior distributions for them. The hyper-parameters c_{ij} are chosen particularly as $(c_{00}, c_{01}, c_{02}) = (55, 30, 15)$, $(c_{10}, c_{11}, c_{12}) = (10, 25, 65)$ for both scenarios. Those values are chosen so that the prior distribution are centrated nearly around the "true" values of p_{ij}. Since any single component of a Dirchlet vector has a Beta distribution, we are able to see how concentrated these priors are by using the Beta distribution with certain hyper-parameters. For example, previous information (prior) states that p_{00} is from a Beta distribution with shape $\alpha = 55, \beta = 45$. Figure 2.4 displays the the true p_{ij} values with its corresponding Beta density function. From the figure, we can see that most "true" p_{ij} values are close to the centre of the the density function (especially with p_{01} and p_{10}) and the range of the x-axis are pretty narrow, which suggests that the priors that we use in this case are concentrated ones. We use the concentrated priors here is because by simulation studies, we realized that the no matter the sample size is large or not, a strong prior is crucial in this case (as mentioned in Section 1.3.2).

Same as in case 1, two simulation studies are carried for each scenario, where studies for the first scenario will be talked about in detail as an example.

In the first study for scenario 1, a dataset of size 4000 (2000 controls and 2000 cases) was generated based on the given "true" value. Then, we use 51000 iterations (the first 1000 is the burn-in period) to update unknown parameters.

Figure 2.5 show the traceplot of MCMC algorithm for r_0, r_1 and p_{ij}s after the first 1000

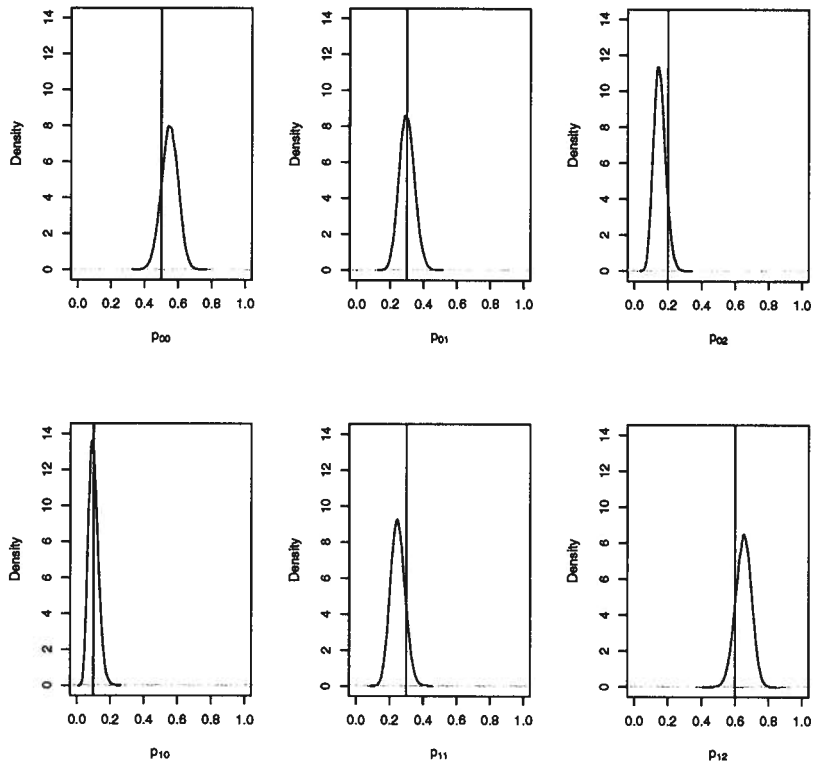

Figure 2.4: *Density plots of "true" p_{ij} values with its corresponding Beta density function. The vertical lines in the graph indicate the "true" values. The Beta density functions (from left to right) are: Beta (55,45), Beta(30,70), Beta(15, 85); Beta(10, 80), Beta(25, 80) and Beta(65,35).*

burn-in period. Again, the Markov chains move smoothly within target range and no chain is fixed at a particular value. We also observe that the generated sample Markov Chain is somehow more stabilized in some unknown parameters (e.g. p_{01}, p_{11}) than others (e.g. r_0 and r_1) after the burn-in period.

Table 2.10 demonstrates the true value, estimated posterior mean and 95% credible interval for each unknown parameter from the first study of scenario 1.

	true value	posterior mean	95% CI
r_0	0.2	0.25	(0.14,0.40)
r_1	0.25	0.32	(0.18, 0.41)
p_{00}	0.5	0.53	(0.46, 0.60)
p_{01}	0.3	0.31	(0.27, 0.35)
p_{02}	0.2	0.16	(0.09, 0.23)
p_{10}	0.1	0.1	(0.05, 0.16)
p_{11}	0.3	0.25	(0.18, 0.35)
p_{12}	0.6	0.64	(0.54, 0.73)

Table 2.10: *True values, posterior means, 95% credible intervals of* r_0, r_1 *and* p_{ij}*s. These are results from the first simulation study (one dataset simulation) for scenario 1 in case 1.*

Though there are discrepancies between the estimated posterior means of r_0 and r_1 and their true values, 95% credible intervals of estimated means do cover the true values.

Figure 2.6 and Table 2.11 display the results from the second study (100 datasets simulation) of the first scenario in this case .

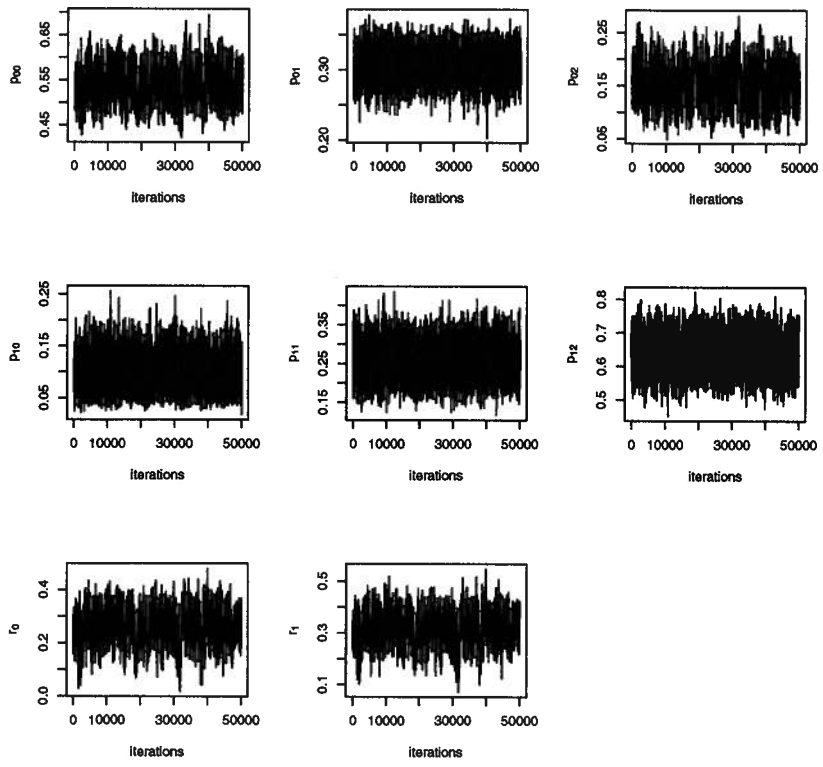

Figure 2.5: *Traceplots of* r_0, r_1 *and* p_{ij}*s from MCMC algorithm in scenario 1 case 2. The traceplots show the 50000 iterations after 1000 burn-in period. The "true" values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

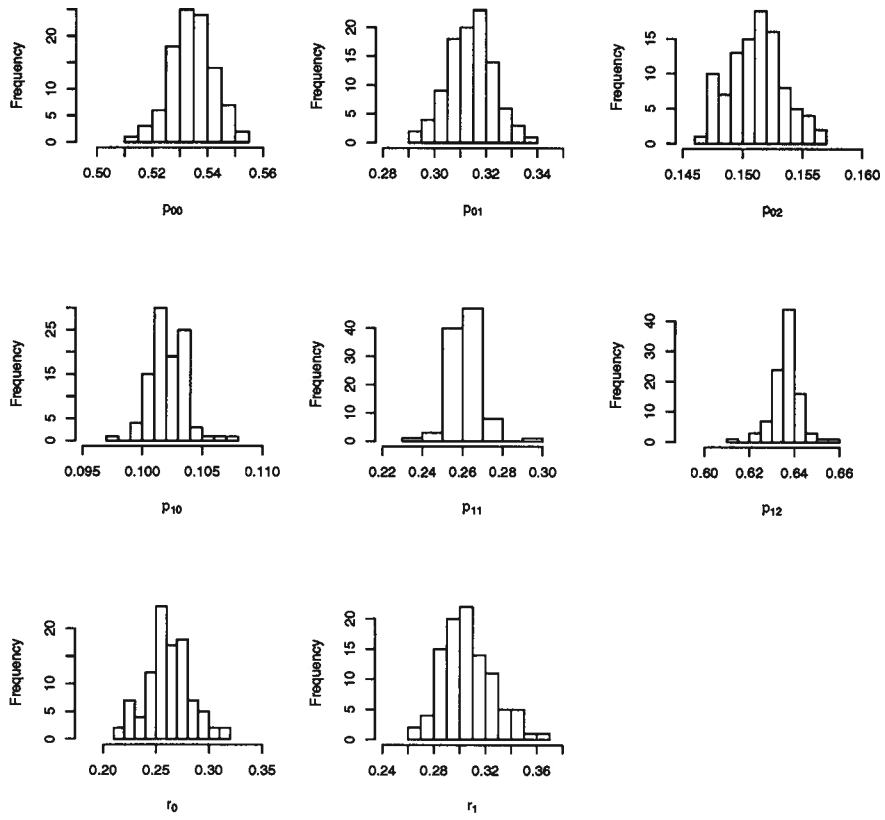

Figure 2.6: *Histogram of 100 posterior means of* r_0, r_1 *and* p_{ij}*s in the second simulation study for scenario 1 case 2. The "true" values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

From Figure 2.6, we observe that the histograms of 100 posterior means of $\hat{r_0}, \hat{r_1}, \hat{p_{00}}, \hat{p_{01}}, \hat{p_{10}}$ and $\hat{p_{02}}$ are shifted to the right from their true values, which indicates an overestimation of the parameters of interest. Though the graph suggests a possible unpleasant result, we still need to study those parameters' corresponding sampling distributions to find out the true performance of our model. Table 2.11 shows the estimated bias, mean square error, 95% credible intervals coverages (of the true value) and the average length of 100 95% credible intervals for each parameter.

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.062	0.0021	100	0.25
r_1	0.056	0.0019	100	0.24
p_{00}	0.035	0.00079	100	0.14
p_{01}	0.014	0.00088	99	0.074
p_{02}	-0.049	0.00023	100	0.14
p_{10}	0.0022	0.00015	100	0.12
p_{11}	-0.039	0.00072	100	0.16
p_{12}	0.0366	0.00061	100	0.18

Table 2.11: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* r_0 *and* r_1 *for scenario 1 case 2. All results are based on 100 datasets, and their true values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

It indicates that the even though the graph shows a potential unpleasant result, the parameters' posterior distributions still cover their true values most of the time. As we discussed earlier, this case is very sensitive to the choice of prior distributions, and the observation obtained from Figure 2.6 could just imply a not strong enough prior for the particular dataset. Thus, it's still reasonable to conclude that the algorithm works well for this case.

Figure 2.7 and Table 2.12 are histogram and statistics results from the second simulation study for scenario 2 in case 2. We omit the results of the first study here due to its limitation of interpretation. From both the figure and table, we notice that no matter whether the probability of exposure is large or small, the prior distributions are always

very important. Weak priors may lead to poor results, and the stronger the prior is the better the results would be. This is dissimilar with most Bayesian cases, where when the sample size is large, the choice of the prior becomes insignificant. Future studies could be conducted on this case.

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.015	0.0027	100	0.23
r_1	0.048	0.0025	99	0.23
p_{00}	0.036	0.0017	100	0.15
p_{01}	0.012	0.0020	96	0.10
p_{02}	-0.04	0.00043	100	0.15
p_{10}	-0.0036	0.00022	100	0.11
p_{11}	-0.011	0.0017	97	0.079
p_{12}	0.015	0.0015	100	0.14

Table 2.12: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* r_0 *and* r_1 *for scenario 1 case 2. All results are based on 100 datasets, and their true values are* $r_0 = 0.7, r_1 = 0.4, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

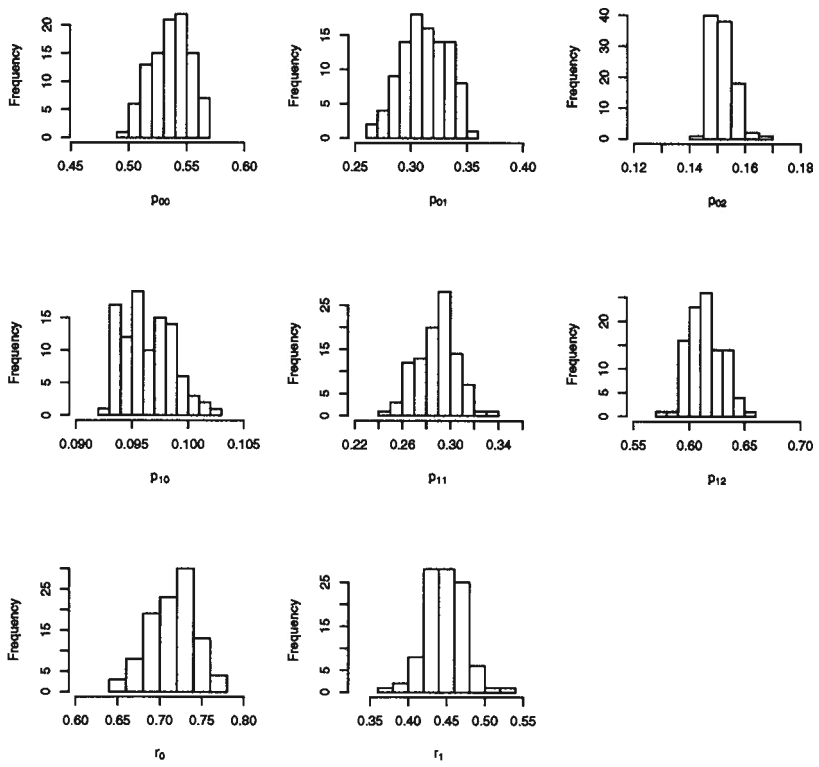

Figure 2.7: *Histogram of 100 posterior means of* r_0, r_1 *and* p_{ij}*s in the second simulation study for scenario 1 case 2. The "true" values are* $r_0 = 0.7, r_1 = 0.4, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$.

2.9 Results for Case 3

Under this cases, we will have four scenarios as follows:

- Scenario 1: $(r_0, r_1) = (0.2, 0.25)$, true odds ratio=1.33, (p_{00}, p_{01}, p_{02}) $= (0.5, 0.3, 0.2)$,$(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$, $(c_{00}, c_{01}, c_{02}) = (1, 1, 1)$, $(c_{10}, c_{11}, c_{12}) = (1, 1, 1)$, validation size=200;

- Scenario 2: $(r_0, r_1) = (0.2.0.25)$, true odds ratio=1.33, (p_{00}, p_{01}, p_{02}) $= (0.5, 0.3, 0.2)$,$(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$, $(c_{00}, c_{01}, c_{02}) = (1, 1, 1)$, $(c_{10}, c_{11}, c_{12}) = (1, 1, 1)$, validation size=100;

- Scenario 3: $(r_0, r_1) = (0.7, 0.4)$, true odds ratio=0.28, (p_{00}, p_{01}, p_{02}) $= (0.5, 0.3, 0.2)$,$(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$, $(c_{00}, c_{01}, c_{02}) = (1, 1, 1)$, $(c_{10}, c_{11}, c_{12}) = (1, 1, 1)$, validation size=200;

- Scenario 4: $(r_0, r_1) = (0.7.0.4)$, true odds ratio=0.28, (p_{00}, p_{01}, p_{02}) $= (0.5, 0.3, 0.2)$, $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$, $(c_{00}, c_{01}, c_{02}) = (1, 1, 1)$, $(c_{10}, c_{11}, c_{12}) = (1, 1, 1)$,validation size=100.

Notice that in scenario 1 and scenario 3, the validation size is 200 compared to scenario 2 and scenario 4 (validation size is 100). As in case 2, let's plot the true values of p_{ij}s along with their prior density functions. From Figure 2.8, we can see that now the range of the x-axis (possible generated values) becomes wider and the "true" values are not that close the to the center of the density functions. This directly implies that these priors are flatter than the one we chose in case 2. We assign flat priors in this case because we believe valuable information could be obtained from the validation data.

Once more, two simulation studies are carried for each of the scenario, and only the first of the two scenario will be talked about in details. Figure 2.9 and Figure 2.10 show the traceplots of MCMC algorithm for r_0, r_1 and p_{ij}s after the first 1000 burn-in period for scenario 1 and 2.

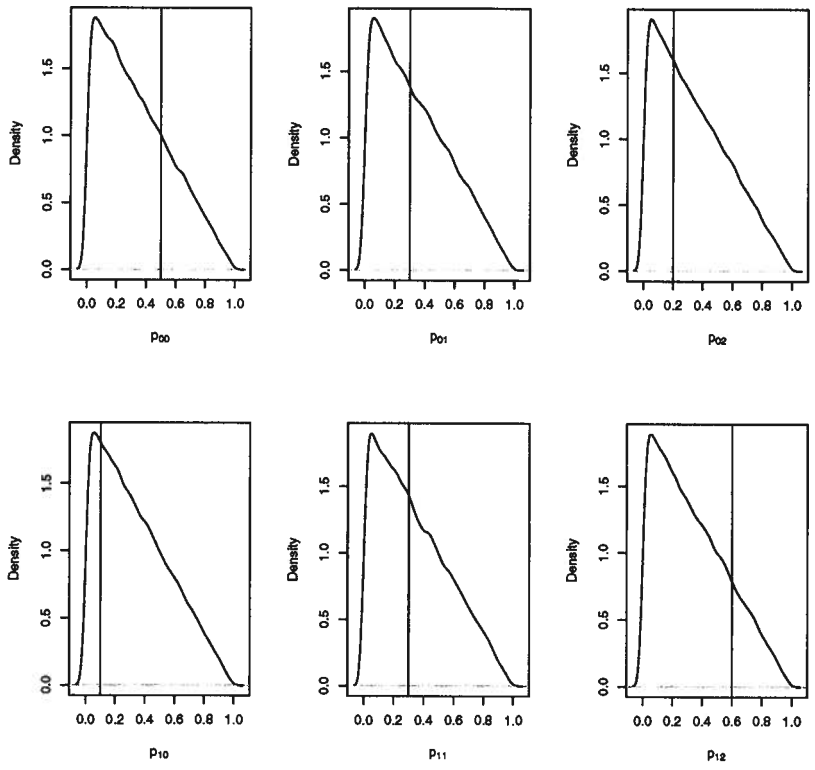

Figure 2.8: *Density plots of "true" p_{ij} values with its corresponding Beta density function. The vertical lines in the graph indicate the "true" values. The Beta density functions (from left to right) are: Beta (1,2), Beta(1,2), Beta(1, 2); Beta(1, 2), Beta(1, 2) and Beta(1,2).*

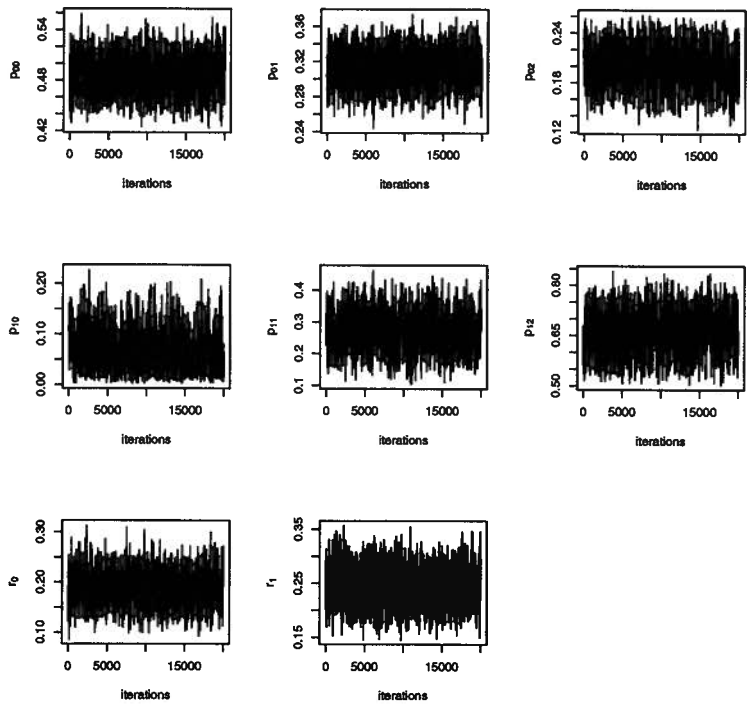

Figure 2.9: *Traceplots of r_0, r_1 and p_{ij}s from MCMC algorithm in scenario 1 case 3. The traceplots show the 20000 iterations after 1000 burn-in period. The "true" values are $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ and $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. Validation size =200.*

We can see that for the generated sample Markov Chain are somehow more stable in this case than case 2. Table 2.13 and 2.14 demonstrates the true value, estimated posterior mean and 95% credible interval for each unknown parameter from the first

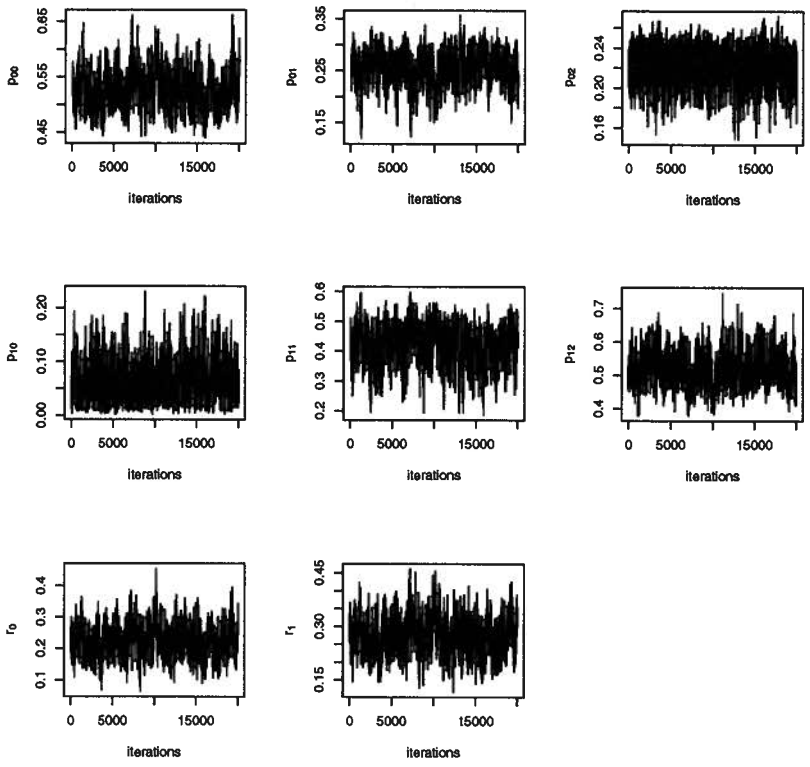

Figure 2.10: *Traceplots of r_0, r_1 and p_{ij}s from MCMC algorithm in scenario 1 case 3. The traceplots show the 20000 iterations after 1000 burn-in period. The "true" values are $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ and $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. Validation size =100.*

study of scenario 1 and 2.

	true value	posterior mean	95% CI
r_0	0.2	0.18	(0.13,0.24)
r_1	0.25	0.24	(0.19, 0.30)
p_{00}	0.5	0.49	(0.45, 0.52)
p_{01}	0.3	0.31	(0.28, 0.34)
p_{02}	0.2	0.20	(0.16, 0.24)
p_{10}	0.1	0.07	(0.00 0.13)
p_{11}	0.3	0.27	(0.16, 0.37)
p_{12}	0.6	0.66	(0.56, 0.77)

Table 2.13: *True values, posterior means, 95% credible intervals of* r_0, r_1 *and* p_{ij}*s. These are results from the first simulation study (one dataset simulation) for scenario 1 in case 3. Validation size =200.*

	true value	posterior mean	95% CI
r_0	0.2	0.22	(0.13,0.31)
r_1	0.25	0.27	(0.18, 0.37)
p_{00}	0.5	0.53	(0.46, 0.59)
p_{01}	0.3	0.25	(0.19, 0.31)
p_{02}	0.2	0.22	(0.19, 0.25)
p_{10}	0.1	0.07	(0, 0.13)
p_{11}	0.3	0.42	(0.29, 0.54)
p_{12}	0.6	0.51	(0.42, 0.61)

Table 2.14: *True values, posterior means, 95% credible intervals of* r_0, r_1 *and* p_{ij}*s. These are results from the first simulation study (one dataset simulation) for scenario 2 in case 3. Validation size =100.*

From these one sample studies, we observe that the approach works very well with flat priors. Though the observation now is only based on one sample study result, it is verified by Tables 2.15 and 2.16 from the second simulation study (sampling distribution study based on 100 datasets).

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.0047	0.0032	93	0.12
r_1	0.0024	0.0029	98	0.12
p_{00}	0.0017	0.0020	95	0.07
p_{01}	-0.00045	0.0015	96	0.067
p_{02}	-0.0013	0.0015	98	0.071
p_{10}	0.0090	0.0036	97	0.15
p_{11}	-0.01	0.0045	97	0.20
p_{12}	0.001	0.0039	99	0.19

Table 2.15: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* r_0 *and* r_1 *for scenario 1 case 3. All results are based on 100 datasets, and their true values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation data=200.*

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.0066	0.0039	96	0.16
r_1	0.0027	0.0038	96	0.16
p_{00}	0.00050	0.0022	94	0.093
p_{01}	0.0049	0.0023	94	0.086
p_{02}	-0.0054	0.0021	99	0.09
p_{10}	0.0090	0.0043	96	0.18
p_{11}	-0.021	0.0068	93	0.25
p_{12}	0.012	0.0049	100	0.23

Table 2.16: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* r_0 *and* r_1 *for scenario 2 case 3. All results are based on 100 datasets, and their true values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation data=100.*

Figure 2.11 and Figure 2.12 show the histograms of sampling distributions for each unknown parameter in scenario 1 and 2. Results from the above two tables and figures indicate that when there are some validation data, the formal approach performs equally well (compare with the case 2) though the prior information is weak. Even though intuitively, we may think that the larger validation size would have better results than the smaller size, by comparing Table 2.15 and Table 2.16, it is hard to conclude that there is significant difference in the results obtained from the scenario 1 and scenario 2. One possible explanation is that by increase the validation size from 100 to 200, the algorithm is only able to gain limited "valuable" information. This immediately rises a question that whether there is a cut off point that we are able to get the maximum benefit, i.e. fewer validation data and enough information to obtain a good estimation. This could be an interesting point to study later on.

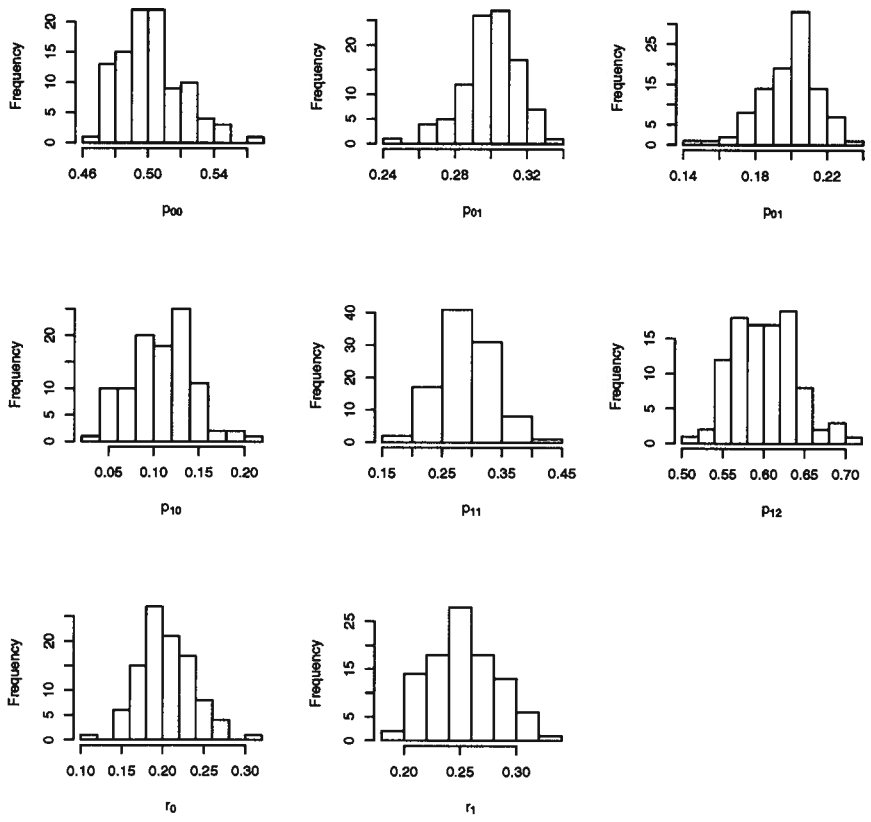

Figure 2.11: *Histogram of 100 posterior means of* r_0, r_1 *and* p_{ij}*s in the second simulation study for scenario 1 case 3. The "true" values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation size =200.*

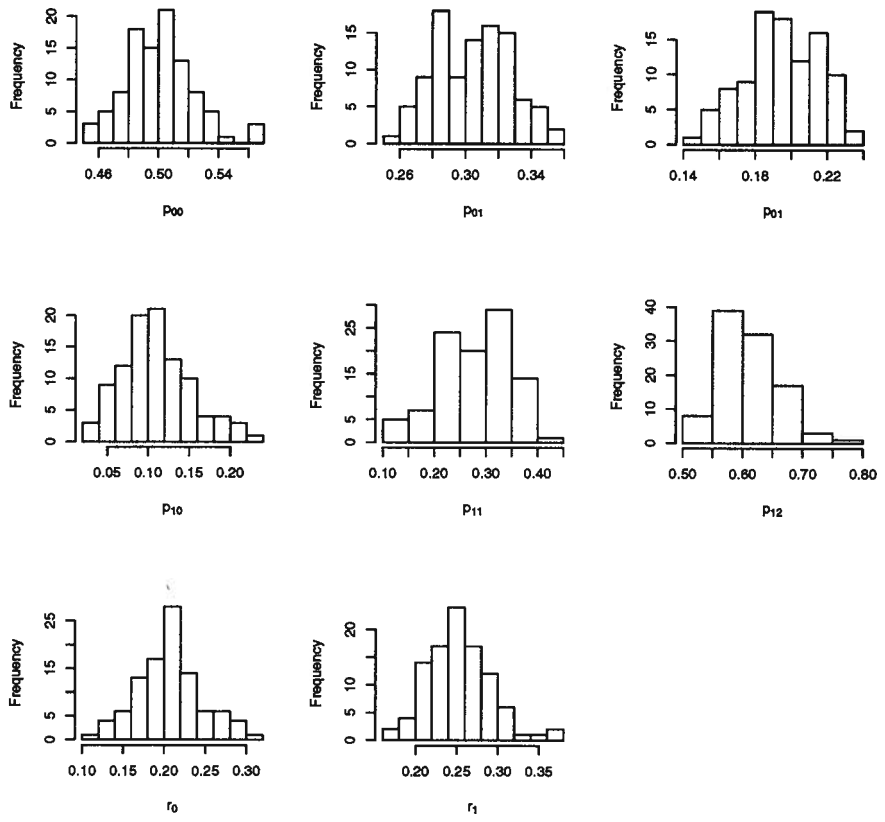

Figure 2.12: *Histogram of 100 posterior means of* r_0, r_1 *and* p_{ij}*s in the second simulation study for scenario 2 case 3. The "true" values are* $r_0 = 0.2, r_1 = 0.25, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation size =100.*

The following tables and figures show the results of scenario 3 and scenario 4 where the prevalence rates are relatively larger ($r_0 = 0.7$ and $r_1 = 0.4$).

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.0033	0.0031	97	0.13
r_1	0.0064	0.0039	92	0.14
p_{00}	0.0042	0.0033	94	0.11
p_{01}	0.00019	0.0025	94	0.09
p_{02}	-0.0044	0.0029	96	0.11
p_{10}	0.0001	0.002	95	0.078
p_{11}	-0.0012	0.0020	93	0.075
p_{12}	0.0011	0.0023	96	0.093

Table 2.17: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* r_0 *and* r_1 *for scenario 3 case 3. All results are based on 100 datasets, and their true values are* $r_0 = 0.7, r_1 = 0.4, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation data=200.*

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
r_0	0.0041	0.0038	97	0.19
r_1	0.0062	0.0047	96	0.19
p_{00}	0.0097	0.0037	96	0.14
p_{01}	-0.0030	0.0026	96	0.10
p_{02}	-0.0067	0.0034	96	0.14
p_{10}	-0.0006	0.0021	99	0.10
p_{11}	-0.0011	0.0025	90	0.082
p_{12}	0.0017	0.0028	97	0.11

Table 2.18: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* r_0 *and* r_1 *for scenario 4 case 3. All results are based on 100 datasets, and their true values are* $r_0 = 0.7, r_1 = 0.4, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation data=100*

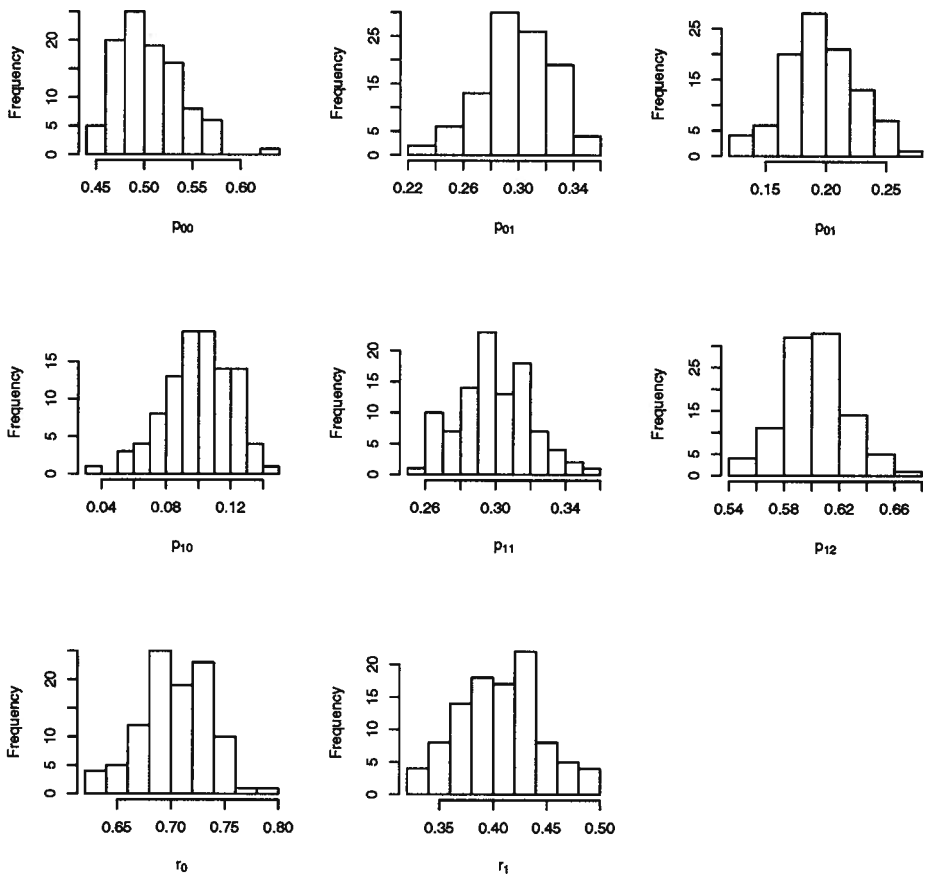

Figure 2.13: *Histogram of 100 posterior means of* r_0, r_1 *and* p_{ij}*s in the second simulation study for scenario 3 case 3. The "true" values are* $r_0 = 0.7, r_1 = 0.4, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation size =200.*

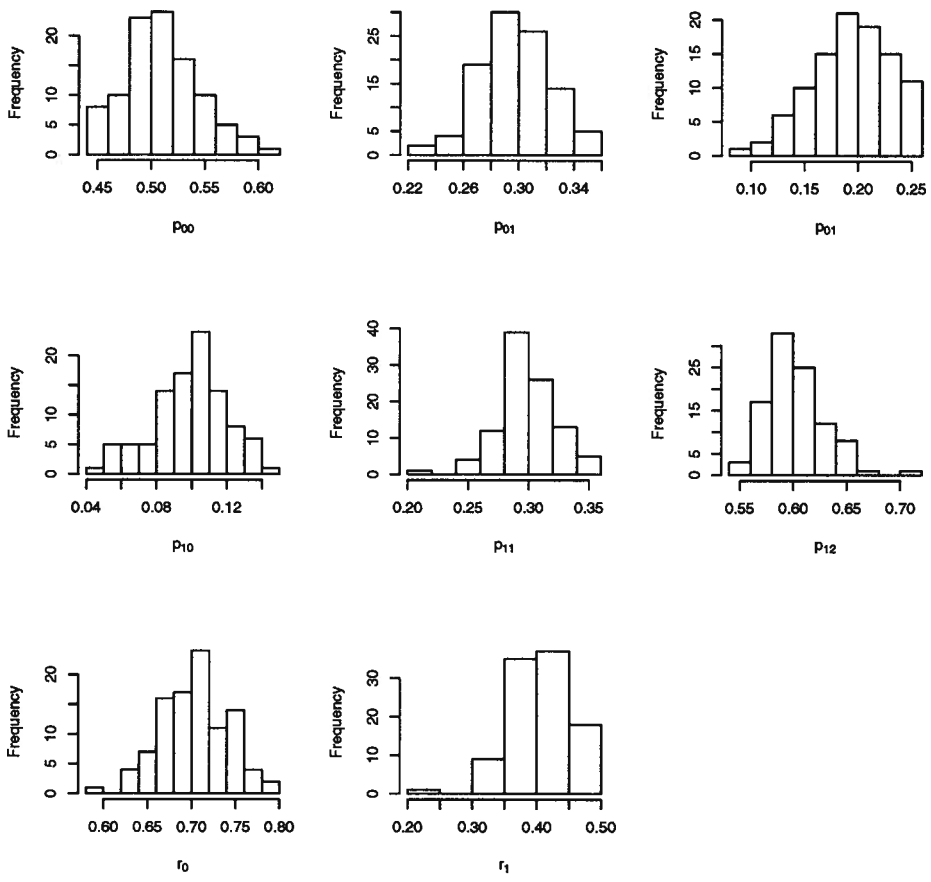

Figure 2.14: *Histogram of 100 posterior means of* r_0, r_1 *and* p_{ij}*s in the second simulation study for scenario 4 case 3. The "true" values are* $r_0 = 0.7, r_1 = 0.4, (p_{00}, p_{01}, p_{02}) = (0.5, 0.3, 0.2)$ *and* $(p_{10}, p_{11}, p_{12}) = (0.1, 0.3, 0.6)$. *Validation size =100*

From the above results, we are able to conclude that our formal approach works well when the prevalence rate is relatively high, and again it is hard to conclude that more validation data will help in getting more precise results.

2.10 Comparison of Odds Ratios

As we talked about previously, it is interesting to compare odds ratios estimated by formal and informal approaches, to see which one tends to be closer to the true vales. Since in the informal approach, we tend to group two categories from the exposure "Unlikely" and "Maybe" together, only when the probability of exposure is rare, we would only compare odds ratios in scenarios that $r_0 = 0.2$ and $r_1 = 0.25$ in each case, i.e. scenario 1 in case 1 and 2, scenario 1, 2 in case 3.

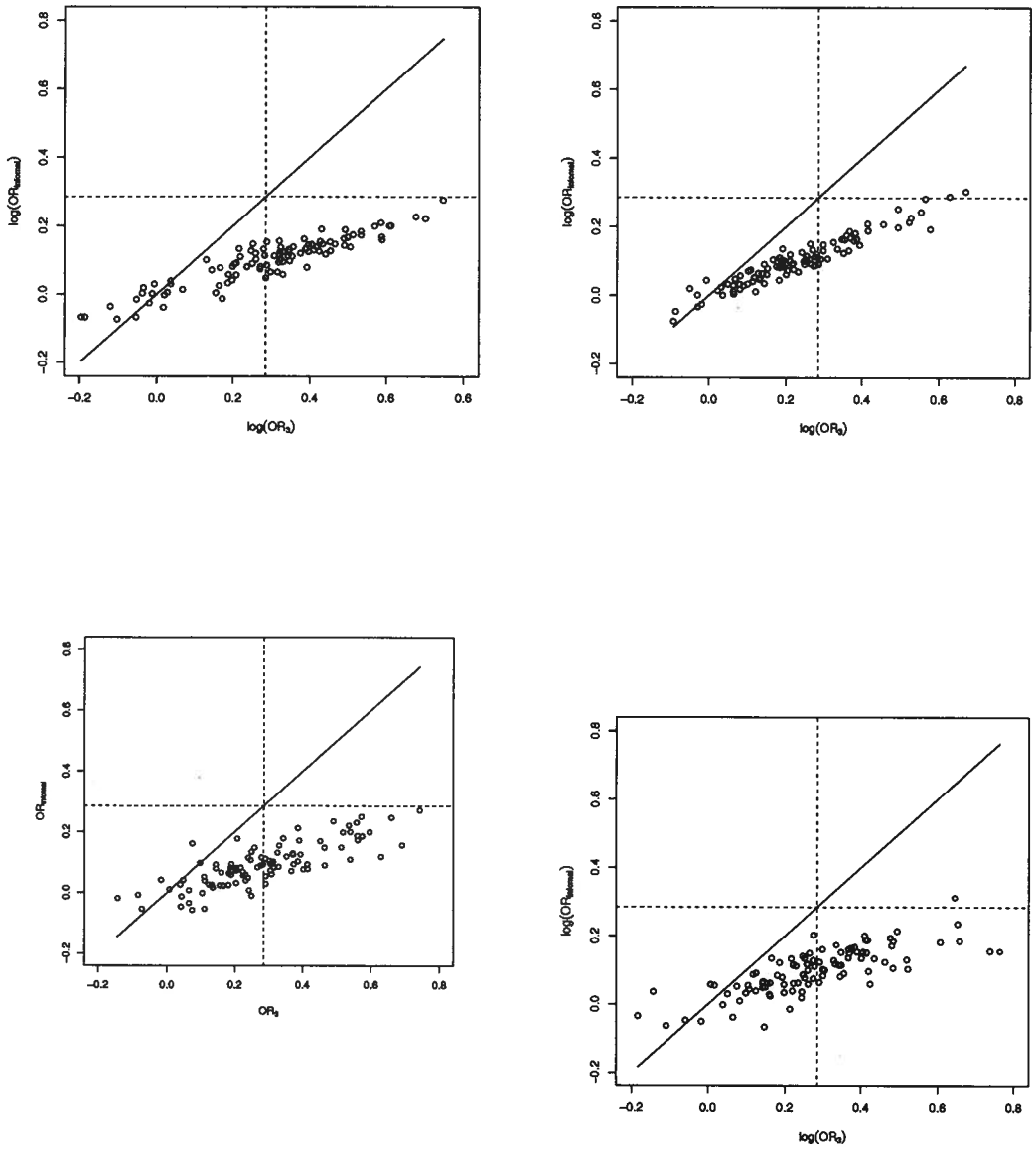

Figure 2.15: *Comparisons of log informal odds ratios with log formal odds ratios for scenario 1 in case 1(upper left), scenario 1 in case 2 (upper right), scenario 1 & 2 in case 3 (lower left and right) . The true log odds ratio value is 0.285. The line in the each graph plots if two odds ratios are the same (y=x line). The dash lines represent the "true" value.*

Figure 2.15 suggests that the informal odds ratios tend to underestimate the true values, where the formal odds ratios sometimes overestimates the true value. For most of generated datasets, the informal odds ratio is always less than the formal odds ratios, since the majority of the plots are below the "y=x" line. The following table illustrates the comparisons in details.

		Bias	MSE	Cov. of 95%CI	Ave. 95%CI Width
Scenario 1 Case 1	$OR_{informal}$	-0.19	0.0078	21	0.27
	OR_3	0.008	0.020	93	0.75
Scenario 1 Case 2	$OR_{informal}$	-0.19	0.0072	23	0.30
	OR_3	-0.055	0.016	93	0.66
Scenario 1 Case 3	$OR_{informal}$	-0.19	0.0072	23	0.27
	OR_3	0.00027	0.018	93	0.70
Scenario 2 Case 3	$OR_{informal}$	-0.18	0.0068	27	0.27
	OR_3	-0.0043	0.018	97	0.76

Table 2.19: *Estimated bias, mean square error (MSE), coverage of 95 % confidence intervals and average 95%CI width of informal and formal log odds ratios for scenario 1 (or 2) in three cases. True log odds ratio is 0.28.*

Table 2.19 indicates that the formal analysis produces more precise estimation of log odds ratio. It has a small bias, great coverage rate, and a reasonable average of 95% CI width, where, on the other hand, the informal approach gives larger bias and unexpected small coverage rate.

Thus, we are able to conclude that the formal approach generally does a better job than the informal approach in estimating the odds ratio. Researchers who apply the informal approach may need to take serious consideration of the measurement error, otherwise, the results might be very biased.

Chapter 3

Simulation Study for Continuous Exposure Variable

3.1 Introduction

Suppose a researcher is interested in investigating the relationship between a health outcome, Y, and a continuous variable, X, and then concluding that $E(Y|X) = \beta_0 + \beta_1 X$ for unknown parameters β_0 and β_1. Nevertheless, when the health outcome is measured precisely, a noisy measurement X^* is often obtained instead of X. If the researcher does not realize the existence of measurement error, or decide to ignore it, then his conclusion about Y and X could be biased. Thus, it is useful to study the impact of the mismeasured covariate X.

Let's assume a precisely measured predictor X has a normal distribution with mean μ and variance λ^2, while its mismeasured surrogate X^* arises from an additive measurement error, which is non-differential, unbiased and normally distributed. Thus, X^* is normally distributed with mean X and variance σ^2. Moreover, since the measurement error is non-differential, which means X^* and Y are conditional independent of Y, the distribution of $(X^*|X,Y)$ and the distribution of $(X^*|X)$ are identical. The joint distribution of X, X^* and Y can then be viewed as:

$$\begin{aligned} f(X^*,Y,X) &= f(X^*|X,Y) \times f(Y|X) \times f(X) \\ &= f(X^*|X) \times f(Y|X) \times f(X). \end{aligned} \tag{3.1}$$

The first term in equation (3.1) is called the measurement model, which is the conditional density of X^* given X and Y. This defines that under the influence of Y, the surrogate X^* arises from the true variable X in a particular way. The second term is called the response model, which explains the relationship between the true explanatory variable$X$ and the response Y. The last term is called the exposure model in epidemiological applications (Gustafson, 2004).

Usually, specific distributions, which involve some unknown parameters will be assumed in equation 3.1, and to make inferences about the X^*, X and Y relationship. Since in reality, the true explanatory variable X is unobserved, the likelihood function of X^* and Y is formed as

$$\begin{aligned} f(X^*, Y) &= \int f(X^*, X, Y) dX \\ &= \int f(X^*|X, Y) \times f(Y|X) \times f(X) dX. \end{aligned} \tag{3.2}$$

Though in some cases, equation 3.2 is easy to evaluate, in other cases, big problems could arise when the integral does not have a closed form. Often, a Bayesian MCMC analysis will be used in such condition since one advantage of Bayesian approach is that the likelihood function is not necessary expressed in explicit form. Dempster, Lairdd, and Rubin (1977) proposed the EM algorithm to solve the implicitly problem in a non-Bayesian way, however, in this paper, we will stay with the Bayesian MCMC methods.

Researchers often want to compare the health outcome Y within two groups, thus, they often dichotomize the continuous variable X into two or more categories. Though, Royston, Altman, and Sauerbrei (2006) pointed out a considerable disadvantage of the dichotomization, it is still very common in the literature (MacCallum, Zhang, Preacher, and Rucker, 2002). In this paper, we dichotomize X into two groups with the rule if $X > c$, the subject is truly exposed, otherwise, the subject is not exposed. Note that in reality, the value c is often decided from a previous study or chosen by a health expert.

The relationship of the health outcome and predictor variable is often estimated by obtaining the coefficients from a linear regression model. Since the health outcome, Y, is a binary variable here, logistic regression appeals as a suitable model. General speaking, there are three approaches to estimate the coefficients. An "naive" approach would dichotomize X^* with respect to c, where an "informal" approach dichotomize the surrogate variable X^* according to c^* (a threshold not necessarily the same as c). In reality, the true predictor variable X is unobserved, however, in the "formal" approach (discuss in the paper), it is pretended to be known and be dichotomized with respect to c and fit the model afterward. The choice of the threshold, c^*, is somehow arbitrary, however, in order to keep a high specificity (as in discrete case), some epidemiologists would intend to choose c^* to be bigger than the true c value, such that $Pr(X^* > c^* | X < c)$ is very small.

Notice that researchers, who use the "naive" approach are often not aware of the measurement error or intend to ignore it, while people who use "informal" or "formal" approach do acknowledge existence of the measurement error and try to find out a solution to the problem. Results from all three approach will be compared later in the Chapter.

3.2 Posterior and Prior Distributions

In this paper, the constituent models are studied based on normal distributions. Specifically speaking, we assume the measurement model, $(X^*|X, Y)$, to be a normal distribution with mean X and variance σ^2. Since the measurement error is non-differential, we have

$$X^*|X \sim N(X, \sigma^2).$$

The exposure model is also assumed as a normal distribution with unknown parameters μ and λ^2, i.e.

$$X \sim N(\mu, \lambda^2).$$

Prentice and Pyke (1979) pointed out that the odds ratios are equivalent when both prospective and retrospective logistic model are applied to the case-control data, thus we would like to assume the response model $Y|X$ follows a logistic regression, which is $logitPr(Y=1|X) = \beta_0 + \beta_1 I(X > c)$. By easy transformation, the response model turn to:

$$Pr(Y=1|X) = \frac{e^{[\beta_0+\beta_1 I(X>c)]}}{1+e^{\beta_0+\beta_1 I(X>c)}}.$$

Note that all parameters $\sigma^2, \mu, \lambda^2, \beta_0$ and β_1 are unknown, and proper prior distributions might be needed in order to proceed. Meanwhile, we would like to assume the independence of all prior distributions, so that

$$f(\sigma^2, \mu, \lambda^2, \beta_0, \beta_1) = f(\sigma^2) \times f(\mu) \times f(\lambda^2) \times f(\beta_0) \times f(\beta_1).$$

Specific prior distributions will be assigned later on.

In this chapter, we focus on studying three cases to demonstrate the performance of the "formal approach" : when we have some knowledge the noise of the true exposure value from previous study, i.e. σ^2 is known; when we only have some prior information about the noise term, i.e. σ^2 is unknown but we have some information there; when we have some validation data, i.e. we have some data on X along with X^* and Y for some subjects.

3.3 Case 1: When We Know σ^2

Sometimes, researchers might have some knowledge about the noise of the true explanatory variable from previous study results, then the posterior density function is :

$$\begin{aligned} &f(X_1, ..., X_n, \beta_0, \beta_1, \mu, \lambda^2 | Y_1, ..., Y_n, X_1^*, ..., X_N^*) \qquad (3.3) \\ &\propto \prod_j f(Y_j|X_j) \times \prod_j f(X_j^*|X_j) \times \prod_j f(X_j) \times f(\beta_0) \times f(\beta_1) \times f(\mu) \times f(\lambda^2). \end{aligned}$$

In order to proceed, we need to specify prior distributions for unknown parameters μ, λ^2, β_0 and β_1. To simplify the MCMC algorithm later on, it is convenient to assign the normal distribution for μ, β_0, β_1 and Inverse Gamma distribution for λ^2 as their prior distributions. Thus, we have

$$\begin{aligned}\beta_0 &\sim N(0, d_1^2);\\ \beta_1 &\sim N(0, d_2^2);\\ \mu &\sim N(0, d_3^2);\\ \lambda^2 &\sim IG(d_4, d_5).\end{aligned}$$

where the choice of hyper-parameters d_1^2, d_2^2, d_3^2, d_4 and d_5 determine how flat or concentrated a prior could be. Here, we choose $d_1^2 = d_2^2 = d_3^2 = 100^2$ and $d_4 = d_5 = 0.01$ so that have flatter priors for unknown parameters μ, λ^2, β_0 and β_1.

The posterior density function in (3.3) now turns to:

$$\begin{aligned}&f(X_1, X_2, X_3, ..., X_n, \beta_0, \beta_1, \mu, \lambda^2 | Y_1, Y_2, ..., Y_n, X_1^*, X_2^*, X_3^*, ..., X_n^*) \qquad (3.4)\\ &\propto \prod_j f(Y_j|X_j) \times \prod_j f(X_j^*|X) \times \prod_j f(X_j) \times f(\beta_0) \times f(\beta_1) \times f(\mu) \times f(\lambda^2)\\ &\propto \frac{e^{\sum_j (Y_j[\beta_0 + \beta_1 I(X_j > c)])}}{\prod_j 1 + e^{\beta_0 + \beta_1 I(X_j > c)}} \times \frac{1}{\sigma^n} e^{\frac{-\sum_j (X_j^* - X_j)^2}{2\sigma^2}} \times \frac{1}{\lambda^n} e^{\frac{-\sum_j (X_j - \mu)^2}{2\lambda^2}}\\ &\times \frac{1}{d_1} e^{\frac{-\beta_0^2}{2d_1^2}} \times \frac{1}{d_2} e^{\frac{-\beta_1^2}{2d_2^2}} \times \frac{1}{d_3} e^{\frac{-\mu^2}{2d_3^2}} \times \frac{d_5^{d_4}}{\Gamma(d_4)} \times \lambda^{2-(d_4+1)} e^{-d_5/\lambda^2}\end{aligned}$$

We chose the normal distribution and Inverse Gamma distribution for μ and λ^2 as prior distributions, since they have the property of being conjugate. A conjugate prior means the posterior distribution of μ and λ^2 would come from the same distribution family as the prior distribution. In particular, a simple Gibbs sampler algorithm can be used to generate a sample of μ from the posterior distribution of μ, which is a normal distribution. Similarly, we can use Gibbs sampler algorithm to generate a sample of λ^2 from its posterior distribution, the Inverse Gamma distribution.

Unlike with μ and λ^2, the posterior distributions of β_0, β_1 and X do not have the same form as any other familiar distributions that we recognize . Thus, the Gibbs sampler does not work for updating them and we need to use the Metropolics - Hasting algorithm instead. As introduced in Chapter 1, this algorithm is based on the accept/ reject rule. We would like to avoid the acceptance rate being extreme, such as 100% or 0%. The key to decide the rate is the jump distribution of the parameter. Let's take β_0 for example. Suppose we are at i^{th} iteration right now and after updating μ and λ^2, we want to update the i^{th} value of β_0. We would calculate the joint density as

$$a = f^{i-1}(X_1, X_2, X_3, ..., X_n, \beta_0^{i-1}, \beta_1, \mu^{(i)}, \lambda^{2(i)} | Y_1, Y_2, ..., Y_n, X_1^*, X_2^*, X_3^*, ..., X_n^*)$$

where f is the joint density as in equation (3.4). Then, we would assign a jump size for β_0, such that $\beta_0^{(cond)} = \beta_0^{i-1} + t$, where t is from the normal distribution with mean 0 and variance k^2 and we would again calculate the new joint density value as:

$$b = f^{i-1}(X_1, X_2, X_3, ..., X_n, \beta_0^{(cond)}, \beta_1, \mu^{(i)}, \lambda^{2(i)} | Y_1, Y_2, ..., Y_n, X_1^*, X_2^*, X_3^*, ..., X_n^*)$$

Now, we will pick the new updated β_0^i as:

$$\beta_0^i = \begin{cases} \beta_0^{(cond)} & \text{with the probability } \min(b/a, 1), \\ \beta_0^{i-1} & \text{otherwise.} \end{cases}$$

A similar procedure is applied for β_1 and X. The individual jump size often follows a normal distribution with mean 0 and variance k^2. Note that the jump size for each estimated parameter may vary in next two cases.

3.4 Case 2: When We Don't Know σ^2

Though, in the previous section we talked about knowing the measurement error variance of the true X from other studies, in most situation, we do not know the exact value of σ^2 but rather have a prior distribution for it. Then, the new posterior density turns to:

$$\begin{aligned} & f(X_1, X_2, X_3, ..., X_n, \beta_0, \beta_1, \mu, \lambda^2, \sigma^2 | Y_1, Y_2, ..., Y_n, X_1^*, X_2^*, X_3^*, ..., X_n^*) \\ & = \prod_j f(Y_j|X_j) \times \prod_j f(X_j^*|X) \times \prod_j f(X_j) \times f(\beta_0) \times f(\beta_1) \times f(\mu) \times f(\lambda^2) \\ & \times f(\sigma^2) \end{aligned} \tag{3.5}$$

Similarly for other parameters, we need specify a prior distribution for σ^2. Again, because of the conjugately property of Inverse Gamma distribution, we would like to assign Inverse Gamma distribution with shape parameter d_6 and scale parameter d_7 as the prior distribution of σ^2. Note that d_6 and d_7 are hyper-parameters. The choice of prior distributions for other unknown parameters are the same as in case 1 and all other hyper-parameters would be assigned the same values.

Now the joint density becomes:

$$\begin{aligned} & f(X_1, X_2, X_3, ..., X_n, \beta_0, \beta_1, \mu, \lambda^2, \sigma^2 | Y_1, Y_2, ..., Y_n, X_1^*, X_2^*, X_3^*, ..., X_n^*) \\ & = \prod_j f(Y_j|X_j) \times \prod_j f(X_j^*|X) \times \prod_j f(X_j) \times f(\beta_0) \times f(\beta_1) \times f(\mu) \\ & \times f(\lambda^2) \times f(\sigma^2) \\ & \propto \frac{e^{\sum_j (Y_j[\beta_0 + \beta_1 I(X_j > c)])}}{\prod_j 1 + e^{\beta_0 + \beta_1 I(X_j > c)}} \times \frac{1}{\sigma^n} e^{\frac{-\sum_j (X_j^* - X_j)^2}{2\sigma^2}} \times \frac{1}{\lambda^n} e^{\frac{-\sum_j (X_j - \mu)^2}{2\lambda^2}} \\ & \times \frac{1}{d_1} e^{\frac{-\beta_0^2}{2d_1^2}} \times \frac{1}{d_2} e^{\frac{-\beta_1^2}{2d_2^2}} \times \frac{1}{d_3} e^{\frac{-\mu^2}{2d_3^2}} \times \frac{d_5^{d_4}}{\Gamma(d_4)} \times \lambda^{2-(d_4+1)} e^{-d_5/\lambda^2} \\ & \times \frac{d_7^{d_6}}{\Gamma(d_6)} \times \sigma^{2-(d_6+1)} e^{-d_7/\sigma^2} \end{aligned}$$

We are able to update σ^2 by Gibbs sampling since its posterior distribution is known as Inverse Gamma distribution, and hyper-parameters d_6 and d_7 are specified later in the result section.

3.5 Case 3: When We Have Validation Data

Similarly as the validation case in the discrete case, here we assume there is a small proportion of data with complete information on X, X^* and Y, whereas the majority of the data do not have the precise measurement of X but have the surrogate variable X^* instead. Thus, unlike the equation (3.3) and (3.5) in case 1 and case 2, the joint density here becomes:

$$\begin{aligned} &f(X_1, X_2, X_3, ..., X_m, \beta_0, \beta_1, \mu, \lambda^2, \sigma^2 | X_{m+1}, X_{m+2}, ..., X_n, Y_1, Y_2, ..., \\ &\quad Y_n, X_1^*, X_2^*, X_3^*, ..., X_n^*) \\ &= \prod_{j=1}^{m} f(Y_j|X_j) \times \prod_j f(X_j^*|X_j) \times \prod_j f(X_j) \prod_{j=m+1}^{n} f(Y_j|X_j) \\ &\quad \times \prod_j f(X_j^*|X_j) \times \prod_j f(X_j) \times f(\beta_0) \times f(\beta_1) \times f(\mu) \times f(\lambda^2) \times f(\sigma^2), \end{aligned}$$

where the first $j = 1, ...m$ are the non-validation data and the rest is the validation data. Though the joint density changes, there are not many changes regarding the simulation process. The only difference is we do not update the "known" X values in the simulations. We used Gibbs sampler to update μ, λ^2 and σ^2, and the Metropolis - Hastings Algorithm is used to update the β_0, β_1 and the "unknown" X values.

3.6 Results

In the following subsections, traceplots of MCMC algorithms in all three cases are checked along with some statistics for each unknown parameter, such as the posterior mean, 95%

equal-tail credible interval, bias, and estimated mean square error. In this way, we are able to gain some information about how well the MCMC algorithms are working for particular models and randomly generated datasets. Moreover, comparisons of estimated logistic regression coefficients are made among all three approaches, "formal", "naive" and "informal".

In all three cases, the true unknown (common) parameters are set to be: $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$, and $\beta_1 = 1.5$, with the dichotomization value $c = 1$. Other choice of the hyper-parameters as well as the jump size are subject to different cases. Moreover, two simulation studies are performed for each case. The first concentrates on gathering information on estimates from one sample. The second focuses on studying the sampling distributions of each estimator across 100 simulated datasets.

3.6.1 Results for Case 1

In the first study, a dataset of size 500 was generated based on a choice of given parameter values. Then, we use 6000 iterations (the first 1000 iterations are burn-in period) of the MCMC algorithm to estimate unknown parameters. Note that the jump sizes for β_0, β_1 and X in the Metropolis-Hasting algorithm are chosen to be 0.15, 0.75, and 1 respectively. Figure 3.1 and Figure 3.2 show the traceplots of the MCMC algorithm outputs for unknown parameters in case 1. These traceplots only plots the iterations after the burn-in period. We can see that the Markov Chain is somehow stabilized after the burn-in period, and the Markov chain move thoroughly within the target range.

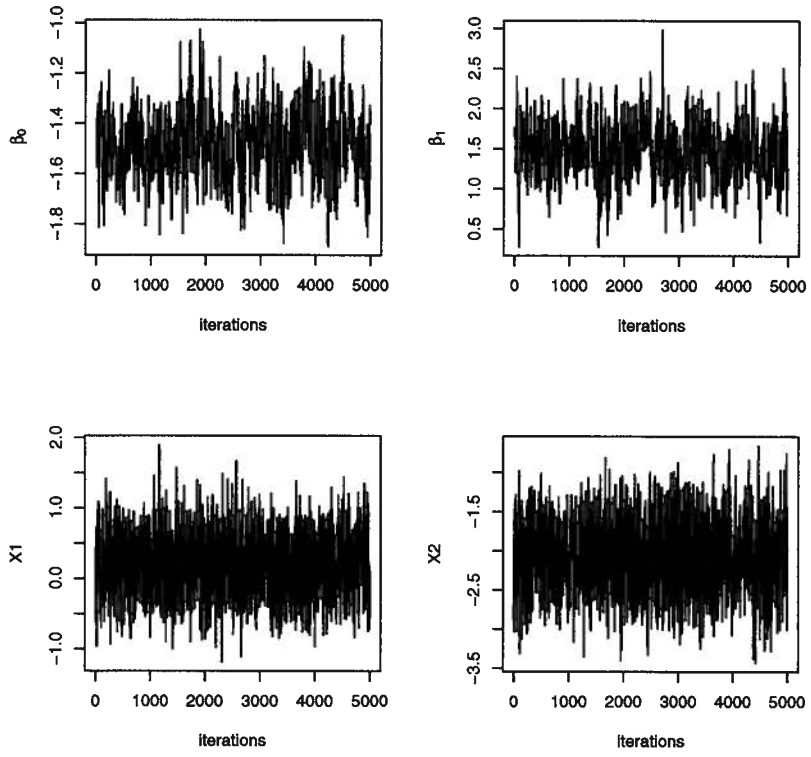

Figure 3.1: *Traceplots of $\beta_0, \beta_1, X_1, X_2$ from MCMC algorithm in case 1. The traceplots show the 5000 iterations after 1000 burn-in period.*

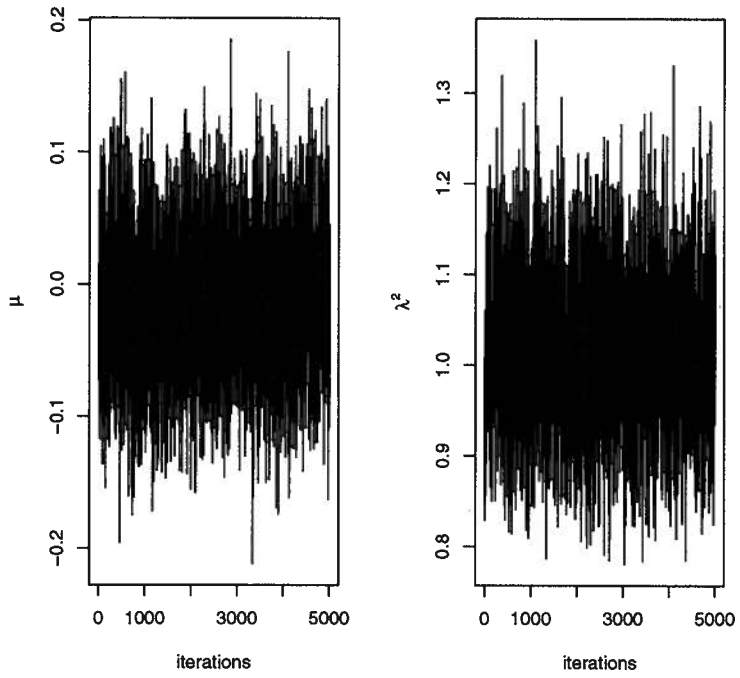

Figure 3.2: *Traceplots of μ, λ^2 from MCMC algorithm in case 1. The traceplots show the 5000 iterations after 1000 burn-in period.*

Table 3.1 shows the true values, posterior means and the 95% credible intervals for each of the unknown parameters estimated from the data.

	true value	posterior mean	95% CI
μ	0	-0.013	(-0.112, 0.086)
λ^2	1	1.014	(0.86, 1.17)
β_0	-1.5	-1.48	(-1.73, -1.21)
β_1	1.5	1.48	(0.79, 2.13)

Table 3.1: *True values, posterior means, 95% credible intervals of* $\mu, \lambda^2, \beta_0, \beta_1$. *These are results from the first study in case 1.*

From the table, we can see that the 95% credible interval of each unknown parameter actually covers the corresponding true value and the posterior mean is very close to the corresponding true value.

The second study just involves repeating the first study 100 times and the sampling distribution of each estimator is studied. Figure 3.3 is the histogram of the posterior means for the 100 samples in case 1.

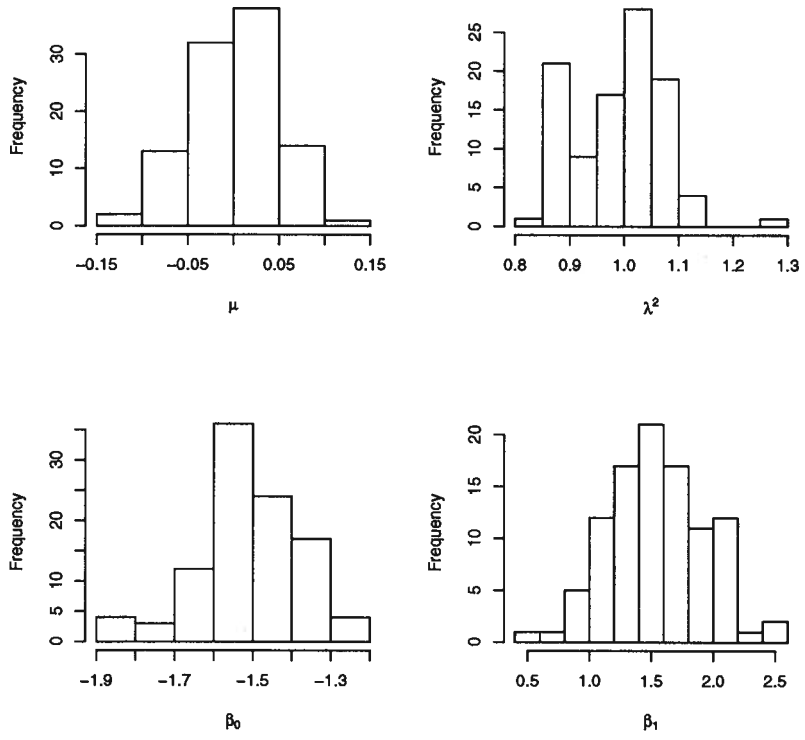

Figure 3.3: *Histograms of 100 posterior means for* $\mu, \lambda^2, \beta_0, \beta_1$ *in the second study in case 1. The true values are* $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$ *and* $\beta_1 = 1.5$

From the figure, we can see that the sampling distributions of $\hat{\mu}$ and $\hat{\beta}_1$ are approximately normally distributed and centered at their true values, whereas the sampling distribution for $\hat{\lambda}^2$ is a little right skewed and $\hat{\beta}_0$ is somehow left skewed.
Table 3.2 summarizes each parameter estimator as:

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
μ	0.0023	0.0050	95	0.19
λ^2	-z0.0096	0.0080	98	0.31
β_0	-0.065	0.013	93	0.54
β_1	0.051	0.038	94	1.42

Table 3.2: *Estimated bias, mean square error (MSE), coverage of 95% CI and the average width of $\mu, \lambda^2, \beta_0, \beta_1$ for case 1. All results are based on 100 datasets.*

The above table shows that the biases and estimated mean square error (MSE) for each unknown parameter are quite small, especially for μ and λ^2. Furthermore, the average widths, out of the 100 runs, of the credible intervals for μ and λ^2 are pretty small. However, the wide average widths for β_0 and β_1 suggest that there is more variation among the 100 estimated β_0 and β_1. Also, out of the 100 times, the 95% credible intervals cover (CI) cover the true μ and λ^2 value 96 times, and cover the true β_0 and β_1 94 times, which suggests a good overall performance of the formal approach.

3.6.2 Results for Case 2

In this case, σ^2, the measurement error variance, is also estimated with other parameters and the "true value" is chosen as 0.25. The hyper-parameters of the Inverse Gamma distribution of σ^2 are specified as 200 and 50 respectively. The specific choice of the hyper-parameters results in a concentrated prior distribution. From Figure 3.4, we can see that this prior has a relatively narrow range and the centre of the distribution is close to the "true" value.

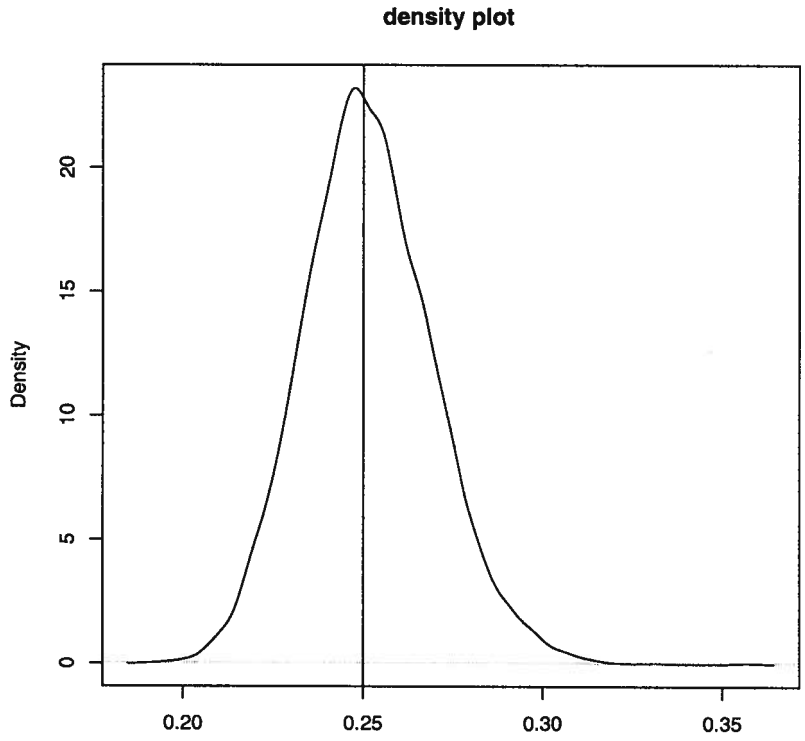

Figure 3.4: *Density plot of Inverse Gamma distribution with hyper-parameters:* $\alpha = 200$ *and* $\beta = 50$. *The vertical line is "true" value of the* $\sigma^2 = 0.25$

Moreover, the jump sizes for updating β_0, β_1 and X in Metropolis - Hastings Algorithm are changed to 0.15, 0.55, 0.8, to avoid extreme acceptance/rejection rates.
Again results from the first study (1 sample study) are displayed first. Figure 3.5 and Figure 3.6 are the traceplots of parameters the MCMC algorithm in case 2.

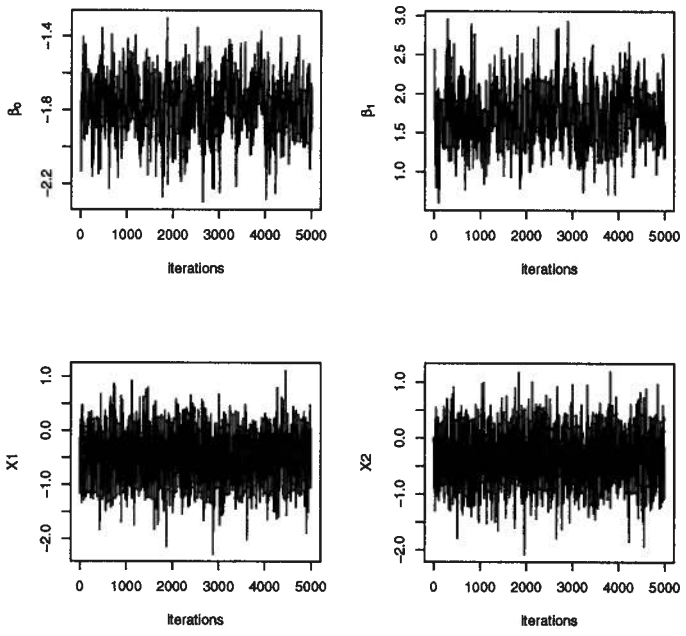

Figure 3.5: *Traceplots of $\beta_0, \beta_1, X_1, X_2$ from MCMC algorithm in case 2. The traceplots show the 5000 iterations after 1000 burn-in period.*

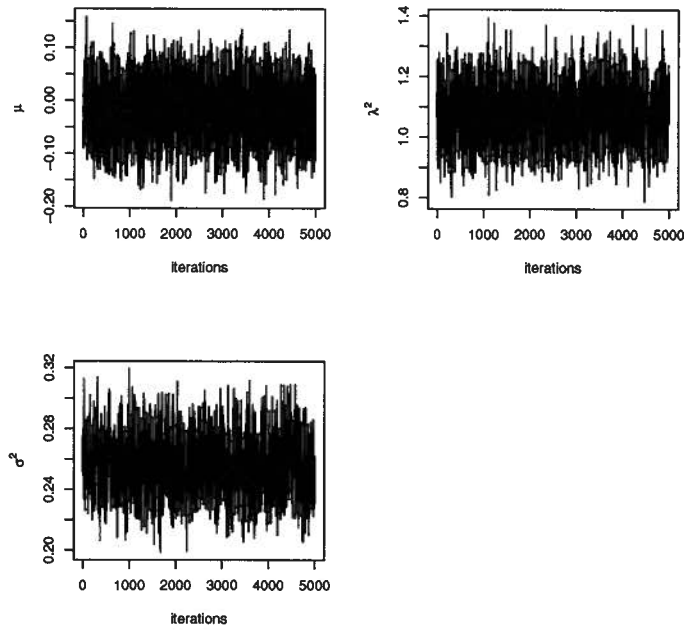

Figure 3.6: *Traceplots of μ, λ^2 and σ^2 from MCMC algorithm in case 2. The traceplots show the 5000 iterations after 1000 burn-in period.*

We can see that the Gibbs sampling algorithm is very stable when updating σ^2 in 5000 iterations, and the chains do not have a mixing or convergence problem.

Table 3.3 shows the posterior means and 95% CI analyzed based on the particular dataset. Again, all the 95% CI covers the true values of the parameters, and it's reasonable to conclude that the approach works well for this particular dataset.

The procedure of the first study is repeated 100 times in the second study. We are able

	true value	posterior mean	95% CI
μ	0	-0.0061	(-0.089, 0.10)
λ^2	1	0.88	(0.73, 1.03)
σ^2	0.25	0.25	(0.22, 0.29)
β_0	-1.5	-1.51	(-1.77, -1.27)
β_1	1.5	1.72	(1.06, 2.39)

Table 3.3: *True values, posterior means, 95% credible intervals of* $\mu, \lambda^2, \beta_0, \beta_1$ *and* σ^2. *These are results from the first study in case 2. The "true" values are:* $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$ *and* $\beta_1 = 1.5$.

to study the sampling distribution of β_1 to get a better understanding of the potential problem (overestimate the parameter) in the first study. The problem could be just happening by chance or it could show that overall the MCMC algorithm underestimates β_1. Figure 3.7 confirms that the problem of underestimating β_1 is just due to chance and overall the estimated β_1 is roughly follows a normal distribution centered at the "true" value. Except for λ^2, other parameter estimators are all approximately following a normal distribution centered at their corresponding "true" value.

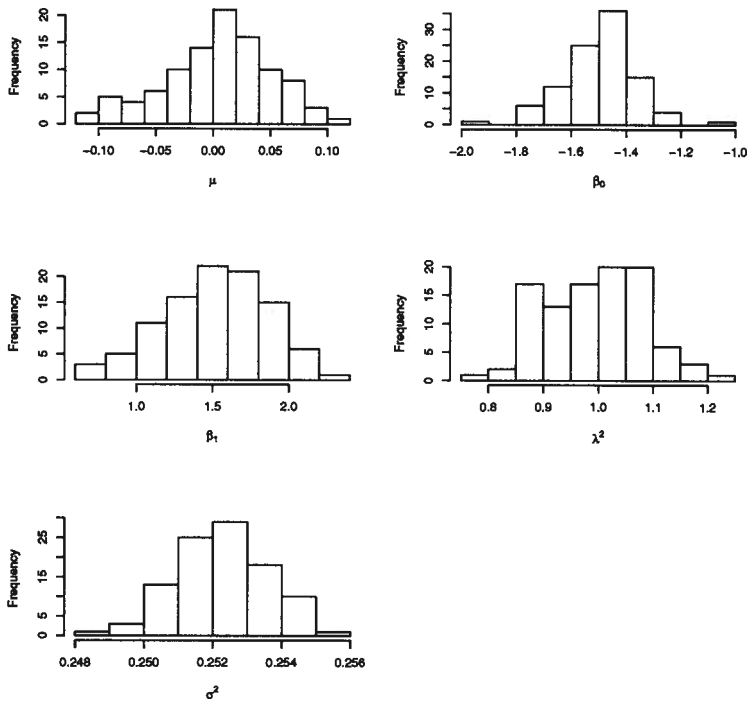

Figure 3.7: *Histograms of 100 posterior means for* $\mu, \lambda^2, \beta_0, \beta_1$ *and* σ^2 *in the second study in case 2. The "true" values are:* $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$ *and* $\beta_1 = 1.5$.

Table 3.4 outlined the bias, MSE, coverage of 95% and average width of the 95% for each estimator. It suggest that our approach produces reliable estimators with small biases, small MSE, satisfactory coverage rates and reasonable average credible interval widths.

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
μ	0.0034	0.0047	95	0.20
λ^2	0.007	0.0087	95	0.32
β_0	-0.054	0.013	97	0.54
β_1	0.019	0.035	95	1.41
σ^2	0.0023	0.00013	100	0.071

Table 3.4: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* $\mu, \lambda^2, \beta_0, \beta_1$ *and* σ^2 *for case 2. All results are based on 100 datasets. The "true" values are:* $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$ *and* $\beta_1 = 1.5$.

3.6.3 Results for Case 3

In this case, we took our validation size to be 50, which means 10% each dataset has precise measurements of X. The new jump sizes for β_0, β_1 and X are 0.1, 0.35 and 0.7. Figure 3.8 and Figure 3.9 are the traceplots of parameters in the MCMC algorithm in the first study .

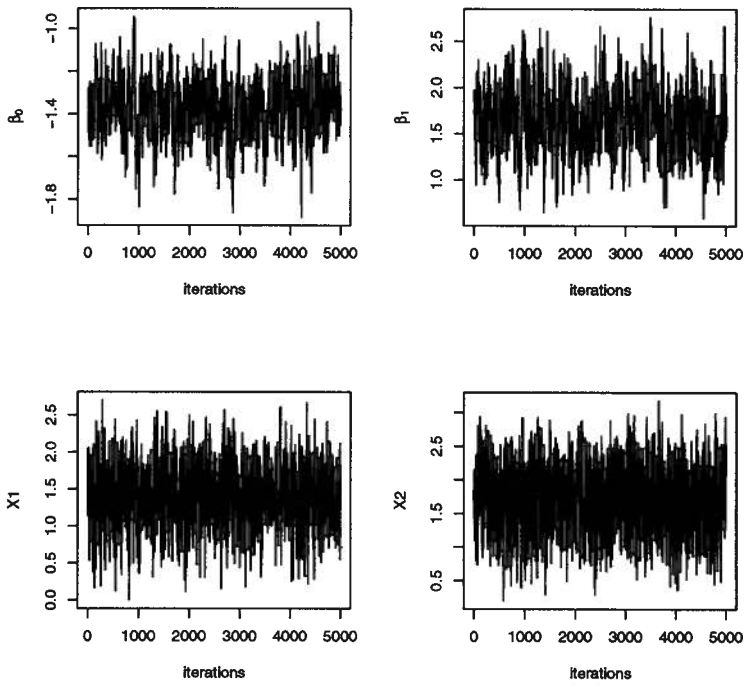

Figure 3.8: *Traceplots of $\beta_0, \beta_1, X_1, X_2$ from MCMC algorithm in case 3. The traceplots show the 5000 iterations after 1000 burn-in period.*

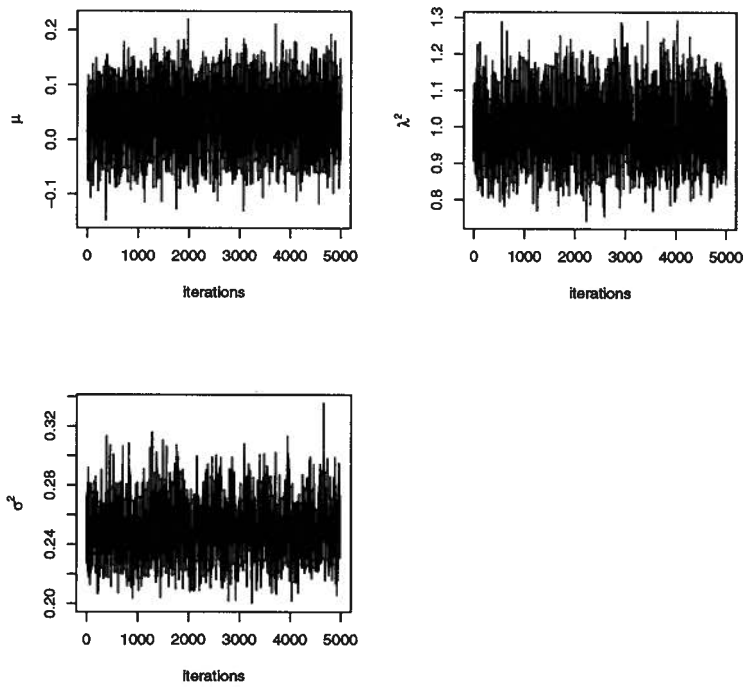

Figure 3.9: *Traceplots of μ, λ^2 and σ^2 from MCMC algorithm in case 3. The traceplots show the 5000 iterations after 1000 burn-in period.*

	true value	posterior mean	95% CI
μ	0	-0.032	(-0.013, 0.066)
λ^2	1	1.03	(0.87, 1.19)
σ^2	0.25	0.24	(0.21, 0.28)
β_0	-1.5	-1.44	(-1.68, -1.19)
β_1	1.5	1.30	(0.63, 1.97)

Table 3.5: *True values, posterior means, 95% credible intervals of* $\mu, \lambda^2, \beta_0, \beta_1$ *and* σ^2. *These are results from the first study in case 3.*

Both figures (Figure 3.8 and Figure 3.9) and statistic values (Table 3.5) indicate that for this particular generated dataset, the approach did a good job. Next, the histogram (Figure 3.10) and summary statistics (Table 3.6) of the sampling distribution for each unknown parameter in the second study are presented.

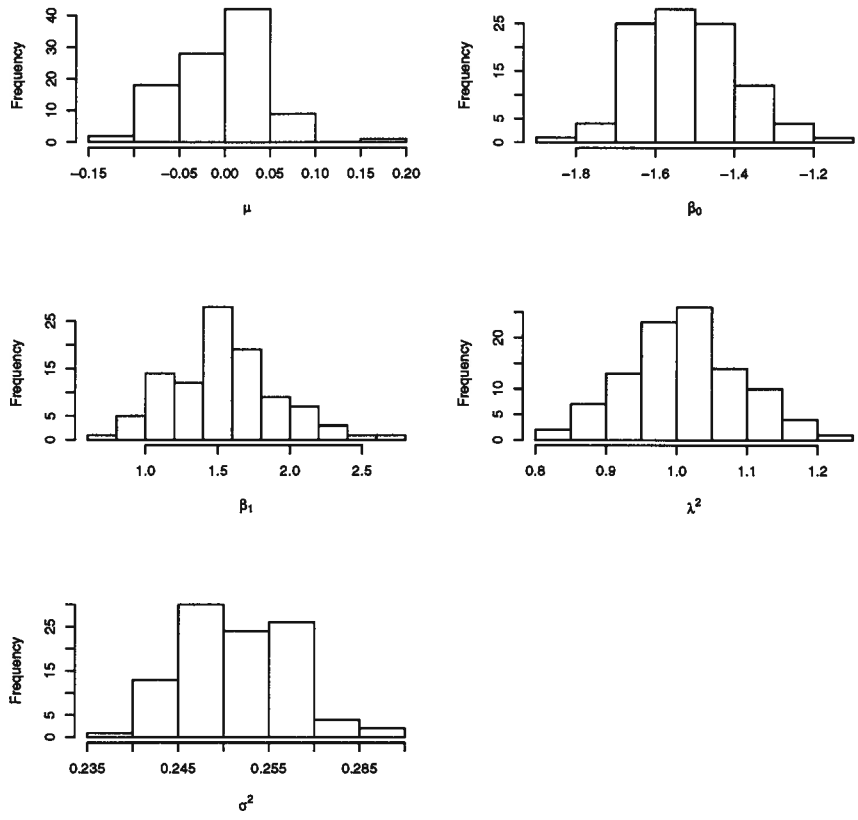

Figure 3.10: *Histograms of 100 posterior means for* $\mu, \lambda^2, \beta_0, \beta_1$ *and* σ^2 *in the second study in case 3. The validation size is 50. The "true" values are:* $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$ *and* $\beta_1 = 1.5$.

From Figure 3.10, we can see that all the sampling distributions of parameter estimators are approximately normally distributed with some skewness involved, expect the histogram for σ^2 looks uniformly distributed at first glance. However, by taking a close look at the figure, we noticed that the scale for the histogram of σ^2 has 3 decimal places, which suggests that most estimated values of σ^2 are very close to the true value, 0.25. This observation is also confirmed in Table 3.6, since the estimation of σ^2 has the smallest bias, MSE and average CI width and highest coverage rate (100%).

	Bias	MSE	Coverage of the 95%CI	Average 95%CI Width
μ	-0.0032	0.0050	97	0.19
λ^2	0.012	0.0080	95	0.31
β_0	-0.024	0.013	95	0.55
β_1	0.051	0.036	95	1.34
σ^2	0.00148	0.00059	100	0.066

Table 3.6: *Estimated bias, mean square error (MSE), coverage of 95 % CI and the average width of* $\mu, \lambda^2, \beta_0, \beta_1$ *and* σ^2 *for case 3. All results are based on 100 datasets with validation size 50. The "true" values are:* $\mu = 0, \lambda^2 = 1, \beta_0 = -1.5$ *and* $\beta_1 = 1.5$.

Evidence, such as small bias, small MSE and hight percentage of true values coverage, in Table 3.6 demonstrate a good performance for the validation model.

Overall, we can conclude that the "formal" approach did an excellent job in estimating unknown parameters for three different models as: knowing the measurement error variance, prior information of the measurement error variance and validation model. Next, we are going to study the comparative performance of the "formal" approach, the "naive" and the "informal" approach.

3.7 Comparison of Three Approaches

As we discussed previously, the "naive", "formal" and "informal" approach would dichotomize either X or X^* with respect to c or c^* to fit a logistic regression model as

following:

$$\begin{aligned} logit(Pr(Y=1|X))_{formal} &= \beta_0 + \beta_1 I(X > c) \\ logit(Pr(Y=1|X^*))_{informal} &= \beta_0 + \beta_1 I(X^* > c^*) \\ logit(Pr(Y=1|X^*))_{naive} &= \beta_0 + \beta_1 I(X^* > c) \end{aligned}$$

The relationship of the health outcome, Y and the exposure variable X is gained from the estimated coefficients, β_0 and β_1. Thus, the comparisons are mainly based on estimating these two parameters. The true values are the same as before: $\beta_0 = -1.5$ and $\beta_1 = 1.5$, and comparisons are constructed in each of the three cases of the "formal" approach. Note that the c^* is chosen as 1.3 so that the specificity and the sensitivity is very high (both are around 95%).

Figure 3.11 to Figure 3.13 are the pairwise plots of results from the three approaches. By observing them, we see there are some linear relationship between estimators, and the linear relationship is somehow weaker when estimating β_1 than estimating β_0. Moreover, both the "naive" approach and the "informal" approach tend to overestimate β_0 but underestimate β_1 , whereas estimations from "formal" approach are located around the true value. Since it is very hard to tell which one is "better" between the "naive" and the "informal" approach from the figures, the summary statistics of estimators' sampling distribution in these two approaches are crucial to know.

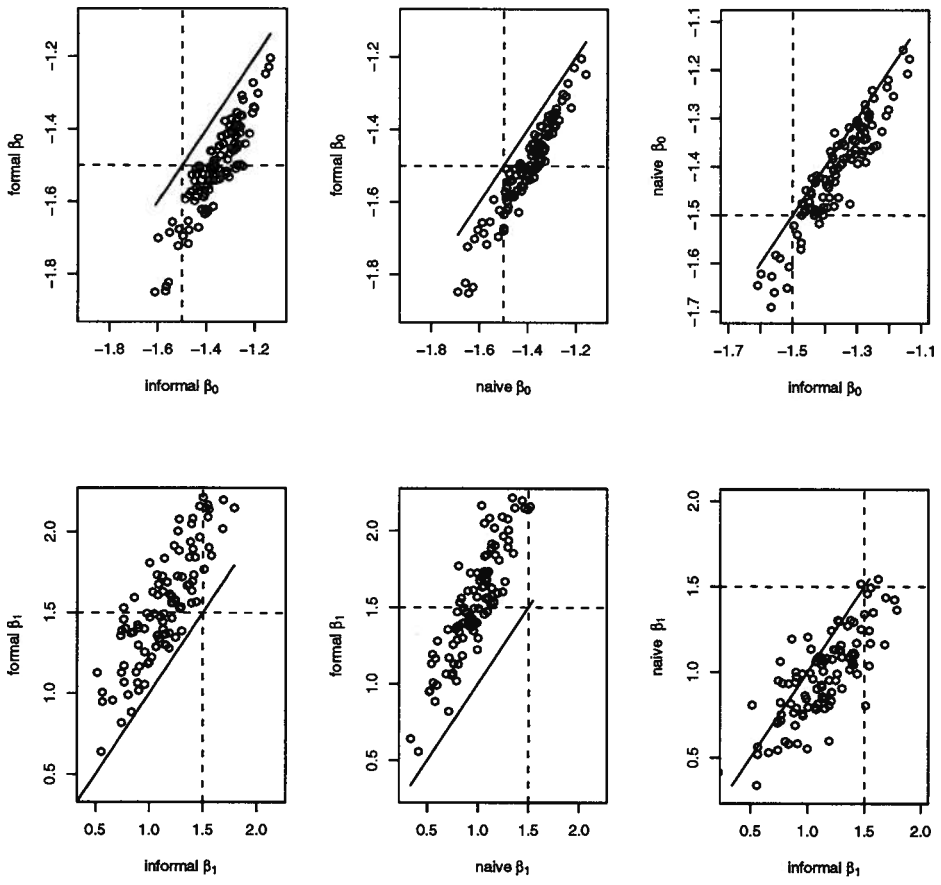

Figure 3.11: *Pairwise plots of three approaches, "naive", "informal" and "formal", in estimating β_0 and β_1 under case 1. The solid line is if "$y = x$", and the dash lines are the corresponding true values.*

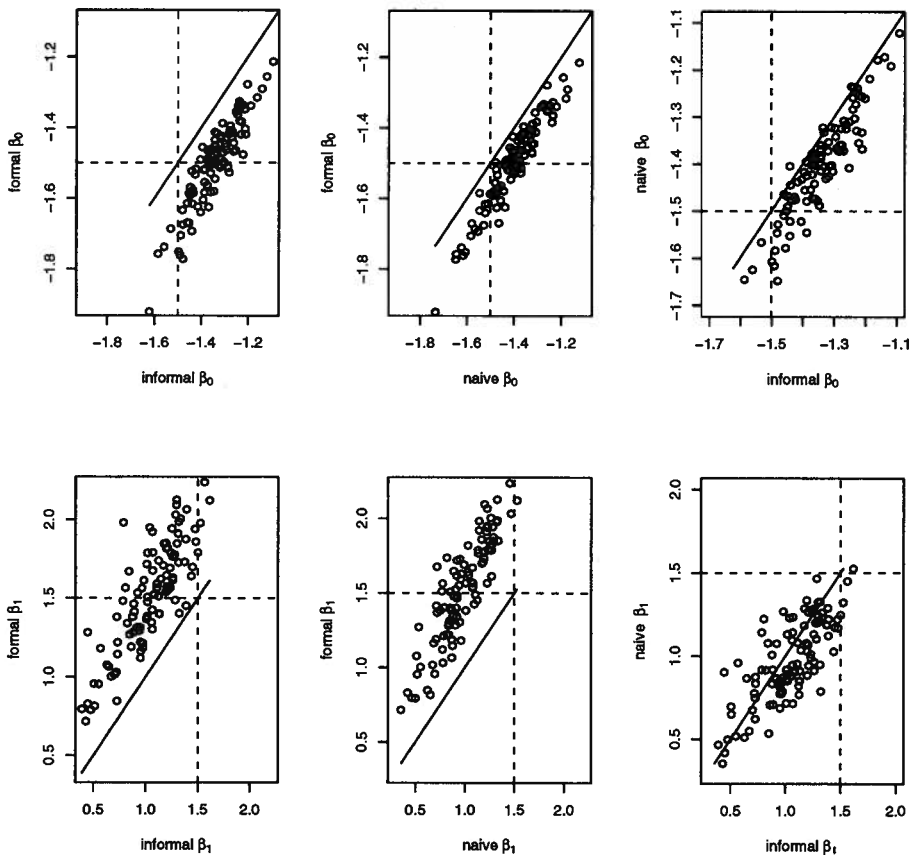

Figure 3.12: *Pairwise plots of three approaches , "naive", "informal" and "formal",in estimating β_1 and β_1 under case 2. The solid line is if "$y = x$", and the dash lines are the corresponding true values.*

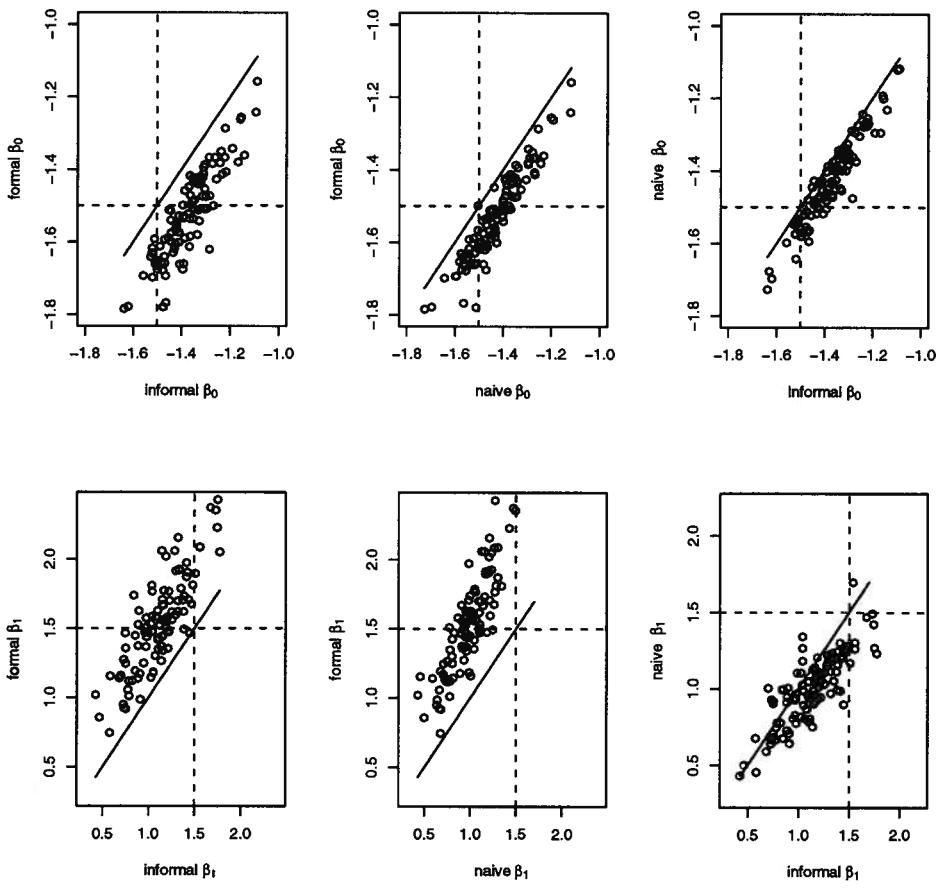

Figure 3.13: *Pairwise plots of three approaches, "naive", "informal" and "formal", in estimating β_0 and β_1 under case 3. The solid line is if "$y = x$", and the dash lines are the corresponding true values.*

Table 3.7 reports the average posterior means of β_0 and β_1, as well as 95% confidence intervals for the average posterior means in all three cases. We are able to conclude that the formal approach is superior to informal and naive approach, since only the confidence interval of "formal" approach cover the true values of β_0 and β_1. Moreover, when estimating β_1, the formal approach produces posterior means and confidence intervals, which are more closer to the true values. The "naive" approach generated the most narrow confidence interval that may also implies over-confidence. These results suggests that it is very dangerous to ignore the measurement errors in the analysis and making proper adjustments for the measurement error is crucial.

		Average Posterior Mean	95%Confidence Interval
Case 1	β_{0naive}	-1.41	(-1.43, -1.39)
	$\beta_{0informal}$	-1.36	(-1.38, -1.34)
	$\beta_{0formal}$	-1.52	(-1.53, -1.48)
	β_{1naive}	0.98	(0.93, 1.03)
	$\beta_{1informal}$	1.14	(1.08, 1.20)
	$\beta_{1formal}$	1.55	(1.47, 1.63)
Case 2	β_{0naive}	-1.40	(-1.42, -1.37)
	$\beta_{0informal}$	-1.34	(-1.36, -1.31)
	$\beta_{0formal}$	-1.49	(-1.52, -1.47)
	β_{1naive}	0.97	(0.92, 1.0201)
	$\beta_{1informal}$	1.04	(0.99, 1.10)
	$\beta_{1formal}$	1.52	(1.45, 1.59)
Case 3	β_{0naive}	-1.43	(-1.45, -1.40)
	$\beta_{0informal}$	-1.37	(-1.39, -1.35)
	$\beta_{0formal}$	-1.49	(-1.52, -1.47)
	β_{1naive}	0.99	(0.95, 1.04)
	$\beta_{1informal}$	1.12	(1.06, 1.18)
	$\beta_{1formal}$	1.55	(1.48, 1.62)

Table 3.7: *Average of posterior means and 95% confidence intervals for the average posterior means of β_0 and β_1 for "naive", "informal" and "formal" approaches. Results are based on 100 samples in case 1, 2 and 3*

Chapter 4

QRS Data Study

To illustrate the ideas and methods that we discussed in the previous chapters in a real world example, we use the QRS dataset. This dataset is provide by Vittinghoff, Glidden, Shiboski, and McCulloch (2004). Heart problems can be diagnosed through the timing of diverse stages in the contraction of the heart. Electrocardiography(EKG) is the device that records the electrical activities of the heart through a duration of time. As the authors indicate the QRS wave is defined as a commonly measured time interval in the contraction of the ventricles. The study dataset contains the QRS times (in milliseconds) for 53 patients, of whom 18 have the inducible ventricular tachycardia (IVT) and 35 of them are without IVT. Note that the sample size is relatively small, since it is very difficult to assemble a large number of subjects to participate in a brain wave study, and the cost of the study is very high. Thus, studies that involve brain waves and electrocardiography devices commonly have small sample sizes.

Though the sample size is considerably small, it is still a good and clean dataset to illustrate our ideas and methods. The response variable Y takes the value of 1 if the subjects has IVT, and 0 otherwise, while the covariates variable X is the QRS time (in milliseconds). Since the QRS time is a continuous variable, we are focusing on the approach introduced in Chapter 3. Even through, in the literature, there are researchers who argue about the accuracy of the QRS duration (Tomlinson, Betts, and Rajappan, 2009), which indicates that measurement error could exist in the measurement of timing in the real world, for the purpose of this thesis, we treat the QRS timing as precisely measured (X) and we simulate the surrogate variable, X^*, in order to compare the results obtained from the true values, naive analysis, informal analysis and our formal analysis.

Nevertheless accuracy of the QRS is questioned by many researchers, there are few articles states the possible magnitude of the measurement error. According to Sahambi, Tandon, and Bhatt (2009), the maximum error rate of the QRS is 6.25% due to the 50 Hz power-line interference. As we lack of detailed information about how the data are collected, we would simply adopt the measurement error rate stated by Sahambi et al. (2009). For our illustrative proposes, we have to assume we know the variance of additive measurement error, σ^2, in order to generate X^*. The error rate we accepted previously is a multiplicative error, and proper transformation is necessary to acquire an additive error (as we defined in Chapter 3). As a result, we choose $X = \log(QRStime)$ instead of QRS time directly. As defined in Chapter 3, under the nondifferential assumption, the measurement model here is

$$X^*|X \sim N(\log QRS, \sigma^2).$$

Mathematically, we can compute σ^2 as:

$$\begin{aligned} X^* &= \log QRS + \sigma Z \\ \Rightarrow e^{X^*} &= (QRS)e^{\sigma Z} \end{aligned}$$

which motivates

$$\Rightarrow e^{\sigma} = 1.0625$$

where Z is a standard normal random variable. We get the variance of additive measurement error as $0.031^2 = 0.00096$, and the surrogate variable, X^*, can be generated afterward.

Since we don't have validation data and we suppose we don't know the variance of the measurement error, the "formal analysis" approach is based on the model that was

introduced as case 2 in Chapter 3. Thus, the response model is

$$\begin{aligned} logit(P(Y=1|X)) &= \log \frac{P(Y=1|I(X>c))}{1-P(Y=1|I(X>c))} \\ &= \beta_0 + \beta_1 I(X>c) \end{aligned} \tag{4.1}$$

Under the assumption of nondifferential measurement error, the measurement model is

$$X^*|X \sim N(X, \sigma^2).$$

The exposure model is:

$$X \sim N(\mu, \lambda^2).$$

We will use this set-up to conduct naive, informal and formal analysis and compare results produced by three approaches with the true values.

4.1 Naive Approach and Informal Approach

For the response model, the value of c is chosen as log(120) as suggested by Tomlinson, Betts, and Rajappan (2009). To refresh the memory, the naive approach would formulate the response model based on X^* and c as:

$$\begin{aligned} logit(P(Y=1|X^*)) &= \log \frac{P(Y=1|I(X^*>c))}{1-P(Y=1|I(X^*>c))} \\ &= \beta_0 + \beta_1 I(X^*>c) \end{aligned}$$

In order to perform the informal approach, we need to pick up a c^* value such that, the specificity is very high (as discussed in Chapter 3), and it is chosen as $c^* = \log(123)$ so

that the $SP = 0.94$. Then the response model of informal approach is:

$$\begin{aligned} logit(P(Y=1|X^*)) &= logit\frac{P(Y=1|I(X^*>c^*))}{1-P(Y=1|I(X^*>c^*))} \\ &= \beta_0 + \beta_1 I(X^*>c^*) \end{aligned}$$

The estimates of β_0 and β_1 for these two approaches can be easily obtained by using the glm function in R. Results are shown in the section 4.3.

4.2 Formal Approach

Since we are assume the QRS timing from the dataset is the true value, we are able to obtain some information about the exposure variable X, such as the the mean of $mean(X) = \mu = 4.64, var(X) = \lambda^2 = 0.094$ and the measurement error variance $\sigma^2 = 0.00091$. Note that those values are not applicable in the real world, and they are available here for this example only (because of our assumption). As discussed early in the simulation study of Chapter 3, we need to assign prior information for the "unknown" parameters. Remember that in section 3.6.2, flatter priors are assigned to unknown parameters since the simulated sample size is considerably large. Though, the sample size is pretty small in this dataset, we are still able to assign flatter priors to some of the unknowns. For example, μ is assigned as $\mu \sim N(0, 100^2)$, λ^2 is assigned as $\lambda^2 \sim IG(0.01, 0.01)$ and β_1 is assigned with $N(0, 100^2)$. Dislike the non-informative priors for μ, λ and β_0, the prior information becomes very crucial for σ^2 and β_1. It is reasonable to assign concentrated priors to these two unknowns, since researchers could easily obtain relative information about these two unknowns from precious study. As a result, concentrated priors are assigned as such that $\beta_1 \sim N(0, 1^2)$ and $\sigma^2 \sim IG(100, 0.1)$. Figure 4.1 displays the prior density plots vs their corresponding "true" values.

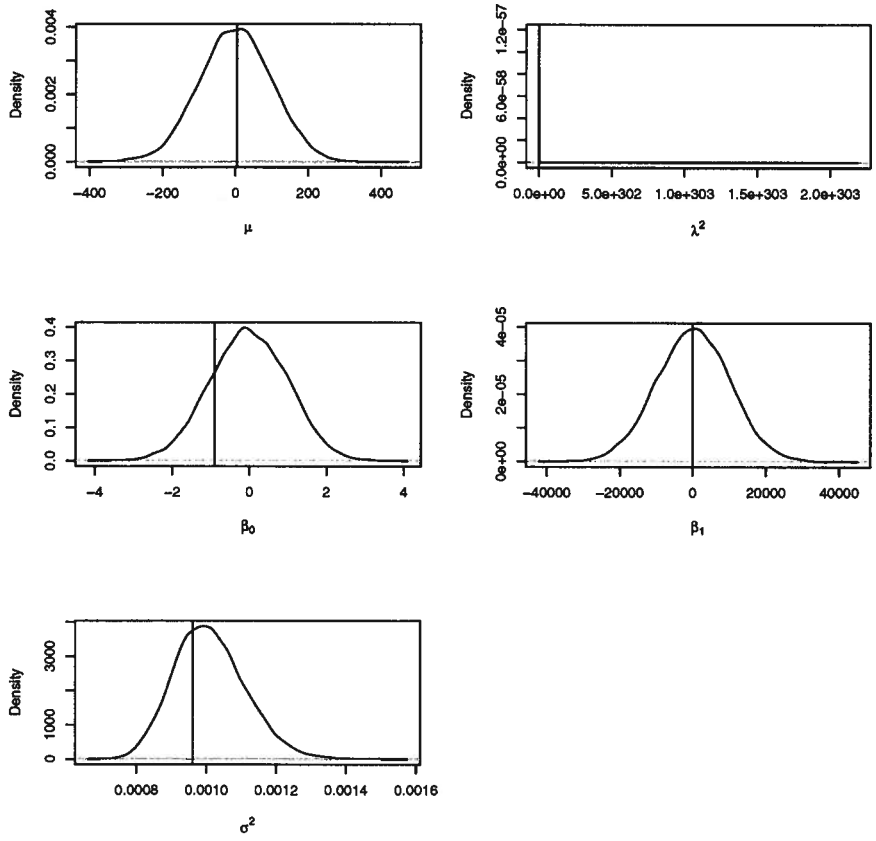

Figure 4.1: *Prior plots of unknown parameters with their hyper-parameters:* $\mu \sim N(0, 100^2), \lambda^2 \sim IG(0.01, 0.01), \sigma^2 \sim IG(100, 0.1), \beta_0 \sim N(0, 100^2)$ *and* $\beta_1 \sim N(0, 1^2)$. *The vertical lines are the corresponding "true" values as:* $\mu = 4.64, \lambda^2 = 0.094, \beta_0 = -0.90, \beta_1 = 0.76$ *and* $\sigma^2 = 0.00096$.

We observe that, regardless of the strength of the prior, the center of prior density for each unknown parameter is most likely located around the corresponding true value. The plot for λ^2 looks abnormal, since the range of its density function goes from 0 to infinity so that it is quite difficult to display on a limited scale. σ^2 has the most concentrated prior, since approximately 95% of its data are enclosed by 0.0008 and 0.0014, a pretty small range. x

Similarly as in the simulation study, we are unable to obtain the full conditional distributions for β_0 and β_1, so we are going to use the Metropolis- Hastings Algorithm to obtain their estimators. All other unknown parameters are updated as in the simulation study of Chapter 3 case 2. The jump size for updating X, β_0 and β_1 in MH consist with the simulation study, which are 0.15, 0.35, 0.8 respectively. Figure 4.2 shows the traceplot of 200000 iterations after 2000 burn-in period, and there are no apparent mixing problems to be noticed. The numeric results are presented and compared in the next section.

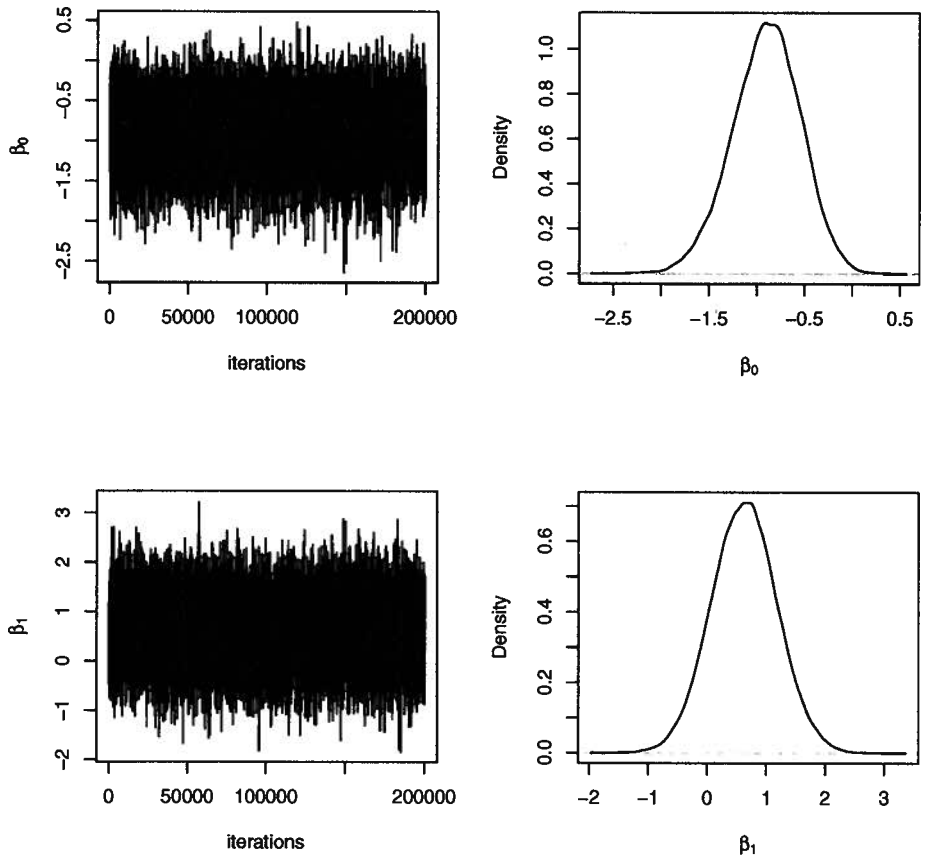

Figure 4.2: *Traceplot and posterior density plots of 20000 iterations after 1000 burn-in period of β_0 and β_1 when applying the MH sampling method.*

4.3 Results

Before discussing the results estimated from the naive, informal and formal approaches, we would like to find out the supposed true result first. It is very easily obtained from the glm function in R and the true model for explaining whether or not a subject has the IVT is estimated as follows:

$$\log \frac{P(Y=1|I(X>c))}{1-P(Y=1|I(X>c))} = -0.90 + 0.76I(X>c).$$

Table 4.1 records results of β_0 and β_1 from the naive, informal and formal approaches.

	Estimate	95%CI	CI width
β_{0naive}	-0.86	(-1.58, -0.14)	1.44
$\beta_{0informal}$	-0.86	(-1.58, -0.14)	1.44
$\beta_{0formal}$	-0.90	(-1.63, -0.22)	1.35
β_{1naive}	0.61	(-0.63, 1.84)	2.48
$\beta_{1informal}$	0.61	(-0.63, 1.84)	2.48
$\beta_{1formal}$	0.67	(-0.46, 1.73)	2.19

Table 4.1: *Estimators, 95% confidence, or credible, intervals of β_0 and β_1 by using "naive", "informal" and "formal" approaches.*

In light of the study performed in Chapter 3, the results for analyzing this dataset behave as we would expect. Though the results are close, the formal approach performed the best when strong priors for β_1 and σ^2 are provided. As the data size gets larger, we believe that the formal approach will keep doing a good job, i.e. estimated values close to the "true" values and the less variability of estimated parameters, even when flatter priors are assigned. Surprisingly, the naive and informal approaches produce the same results, and one possible explanation is that the data size is very small, and there is no significant difference in modeling X^* with threshold c or c^* in this special case. Note that when the data size increases, the chance that naive and informal approaches produce the same results will become slim.

Chapter 5

Conclusion and Future Work

In this thesis, we propose a formal approach to adjust mismeasurement in case-control studies. Ignoring potential mismeasurement on exposure variables could lead to serious problems, such as loss of power, biased estimation and misleading conclusions. In the literature, many methods were proposed to deal with misclassification and measurement error, such as matrix method, inverse matrix method, regression calibration, SIMEX, Expectation-Maximization algorithm in frequentist perspective. Lots of methods are ready to use, nevertheless, they all have their limitations. For example, Carroll, Ruppert, Stefanski and Crainiceanu (2006) stated that though the SIMEX and regression calibration are simple methods to implement, they have limited contributions in reducing the bias caused by the measurement error. The Bayesian approach, on the other hand, is able to correct the bias more precisely and generally. Though, a potentially misspecified exposure model, too complex posterior and intensive computational requirements are occasionally drawbacks in the approach, it has the great advantage that the uncertainties of parameters can be fully incorporated.

A formal approach in the Bayesian perspective is introduced in this dissertation to account for both categorical and continuous exposure variables under the non-differential assumption. Fundamental techniques and concepts are introduced in Chapter 1. Ideas of the proposed formal approach that deals with a categorical exposure variable is introduced and studied through investigating three cases in Chapter 2. The underlying theme of the formal approach that adjusts the measurement error (continuous exposure variable) is presented and examined by studying its performance on three cases again in

Chapter 3. In Chapter 4, a real world dataset is used to evaluate the proposed model. Gibbs sampler and Metropolis-Hasting algorithm are mainly used to sample the parameters of interest from their corresponding posterior distributions.

In Chapter 2, we investigate three cases where we have different levels of knowledge about misclassified probabilities. In each case, the approach is implemented with both low and high prevalence rate, as well as a different validation sample size when we assume validation data are available. Stabilized traceplots suggest that the overall convergence rate is adequate for Markov chain simulation in our proposed model. When the sampling distribution of each unknown parameter is studied, statistical assessments such as, small estimator bias, small mean square error, high coverage rate of the true value and reasonable average 95% credible interval length, all indicate that overall the model is efficient and accurate. When only the prior information about the misclassified probabilities is known, strong and concentrated priors are required to get good estimation. One possible explanation is that, a strong prior is able to reduce the variability of estimators and improve the efficiency of the approach. However, when a small proportion of the validation data is available, it is found that the strong priors become unnecessary, which indicates that the model is able to capture enough information to make good estimation. Moreover, it seems like the size of the validation data does not significantly affect the estimation, and this would be an interesting point to study later on. When the results obtained from low prevalence rate and high prevalence in each case is compared, it is delightful to observe that the approach could work for any prevalence rate. In the end of Chapter 2, estimated log odds ratios are compared for the proposed formal approach and informal approach. It is found, as expected, that the informal approach tends to underestimate the association between the exposure variable and response variable most of the time, and that less than 25% of the 95% confidence intervals actually cover the true value. Even though, the formal approach sometimes overestimate the log odds ratio, the majority of its 95% confidence intervals include the true values and the estimator

bias is much smaller than it is in the informal approach.

In Chapter 5, the proposed approach is implemented with a continuous exposure variable. A logistic regression model is specified for the binary exposure variable and dichotomized continuous exposure variable so that their association is measured according to the coefficients of logistic regression. Two simulation studies for three cases studies are conducted based on our knowledge of the magnitude of the measurement error. As expected, we find that our proposed approach is both efficient and accurate. As in the discrete case, proper priors are crucial when we only have the prior information in hand but not so important when the validation data are accessible. Coefficients obtained from the "naive" approach and informal approach (both of them use the association estimated from response variable Y and X^* to estimate the true relationship between Y and X) are compared with our proposed formal approach at the end of the chapter. Overall, the performance of those approaches decline as we move from formal approach to informal approach and then naive approach. A real world example is used to illustrate our idea and approach. Due to a very small sample size of the dataset, adjusted proper priors are again critical to gain valuable estimations. Fortunately, it is proved that our suggested approach can work practically after strong priors are assigned.

Our proposed Bayesian adjustment for mismeasurement can be extended to a variety of research areas. One straightforward extension would be having more precisely measured covariates in addition to the model that we have right now. Further investigation can be conducted to understand the weakness of our approach, which is that strong priors are needed when we only have prior information. A relevant and interesting study would be to find out whether there is a "cut-off" validation size so that the researchers are able to gain the "maximum" information while spending minimum time or money.

Bibliography

B. A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33:414–418, 1977.

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

G. Casella and E.I. George. Explaining the gibbs sampler. *The American Statistician*, 46:167–174, 1992.

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

A. P. Dempster, N. M. Lairdd, and D. B. Rubin. Maximum likelihoodd from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B 39(1):1–38, 1977.

E. Greenberg and Siddhartha Chib. Understanding the metropolis-hastings algoithm. *The American Statistician*, 49(4):167–174, 1995.

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

H. Kuchenhoff, S. M. Mwalili, and E. Lesaffre. A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics*, 62:85–96, 2006.

R. H. Lyles. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics*, 58:1034–1037, 2002.

R.C. MacCallum, S. Zhang, K.J. Preacher, and D.D. Rucker. On the practice of dichotimization of quantitative variables. *Psychological Methods,*, 7:19–40, 2002.

R. J. Marshall. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43:941–947, 1990.

M. J. Morrissey and D. Spiegelman. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, 55:338–344, 1999.

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

R.L. Prentice and R. Pyke. Quantitative analysis of errors due to power-line interference and base-line drift in detection of onsets and offsets in ECG using wavelets. *Medical and Biological Engineering and Computing*, 35(6):747–751, 1979.

P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

J.S. Sahambi, S.N. Tandon, and R.K.P. Bhatt. Quantitative analysis of errors due to power-line interference and base-line drift in detection of onsets and offsets in ECG using wavelets. *Medical and Biological Engineering and Computing*, 35(6):747–751, 2009.

A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: multi-level, longitudinal and strucutral equation models.* Chapman &Hall/CRC, Boca Raton., 2004.

D.R. Tomlinson, T.R. Y. Betts, and K. Rajappan. Accuracy of manual qrs duration assessment: its importance in patient selection for cardiac resynchronization and implantable cardioverter defibrillator therapy. *Europace*, 11(5):638–642, 2009.

E. Vittinghoff, David V. Glidden, S.C. Shiboski, and C. E McCulloch. *Regression methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models.* Springer, 2004.