Dependent Error Misclassification in both the Response Variable and Covariate

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Department of School of Public Health
University of Saskatchewan
Saskatoon

By
Annshirley Aba Afful

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the School of Public Health
104 Clinic Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 2Z4
Canada

Or

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Errors in Variables (EIV) are a long-standing issue in many fields, including medical and epidemiological studies. Ignoring these errors can produce misleading inferential results. In discrete responses, EIV are commonly termed as misclassification errors. Studies on misclassification have mostly focussed on misclassification in only one variable. Joint misclassification in both the response variable and the covariate has been less explored. Some literature on joint misclassification assumes the misclassification process of the response variable is independent of the misclassification process of the covariate. However, in practice, the dependence of misclassification errors can occur. For example both, the response variable and covariate are obtained from a similar source as in the case of self-reported responses from a questionnaire. The objective is to investigate (1) modeling for error-prone response variable and error-prone covariate and (2) consequences of using an incorrect misclassification model.

In this thesis, we first introduce a model that accounts for dependent misclassification error in a binary response variable and a binary covariate. The dependence of error is captured through covariance-like parameters. Simulation studies are conducted to assess the consequences of fitting an independent misclassification model to data generated from a dependent misclassification model. The simulation experiments have several key factors to manipulate: the amount of misclassification error (sensitivity and specificity), the dependence between the misclassification process of the response variable, and the misclassification process of the covariate, and the proportion of internal validation data. Further, the model is extended to a multi-category setting and simulation study is conducted on a trinary response variable and a trinary covariate. Results from the simulation studies indicate that ignoring dependence of the error in misclassification can be worse than ignoring misclassification.

The proposed model is illustrated through a real data example by establishing the true association between Trichomoniasis and Bacterial Vaginosis, using data from the HIV Epidemiology Research Study (HERS). A likelihood-ratio test is proposed to test the independent misclassification assumption. The test concluded that the dependent misclassification error model fits the HERS data significantly better than the model that ignored dependence misclassification.

Acknowledgements

First and foremost, I thank the Almighty God for his divine grace and favour throughout this study. My deepest gratitude goes to my supervisor Dr. Juxin Liu for her mentorship, patience, and financial support. Her commitment to student welfare is very much appreciated.

I am grateful to Dr. Punam Pahwa and the members of my advisory committee Dr. Cindy Feng, Dr. Holly Mansell, and Dr. Bonnie Janzen, whose insightful suggestions and comments help shape the final version of the thesis. Special thanks to Dr. Longhai Li for permitting me to use his cluster for simulation studies. I want to thank the faculty members, students, and administrative staff of the collaborative Biostatistics program.

I am deeply indebted to my family, especially my husband and best friend Raymond Benjamin Afful, for staying with me and for his unconditional love, sacrifice, support, and encouragement throughout this period. I also thank my daughters Aba Baahwa Afful and Ewuradwoa Mbeah Afful. Last but not least, I am grateful to my siblings Araba, Ato and Maame Nyarkowa.

I dedicate my dissertation work to my family. A special gratitude goes to my parents Esther Quansah and Stephen Mensah Appiatse who have been there for me throughout my life. This also goes to my in-laws Jacob Benjamin Afful and Emma Afful for their encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Ideally, variables included in statistical modeling and inference should be accurate and exact; however, this ideal situation is often not attainable due to errors in variables. Errors in variables (EIV) are a long-standing issue in many fields, including medical and epidemiological studies. These errors occur as a result of inaccurate measurement, incorrect diagnostic criteria, and unreliable data sources, and other inadequacies in obtaining data [25]. Different authors have referred to these errors as measurement error, misclassification, mis-measurement, error-prone data, error-contaminated data, or errors-in-variables. They represent the difference between a measured value of a variable and its true value. One may be tempted to proceed with analysis assuming the observed variables are perfectly measured; however, simply ignoring the presence of errors can have a strong impact on the results of statistical analysis that involve such variable(s) [24].

The observed variables are referred to as surrogates or proxy variables [20, 9]. Proceeding with parameter estimation without accounting for errors may produce misleading inference results. For instance, in regression analysis employing surrogates in estimation is called the *naive method*, and this tends to flatten or attenuate the associated regression coefficient.[20]. The term *misclassification error* is used when EIV occurs in a discrete or categorical variable; thus, classifying an individual or an attribute to a value other than that to which it should be assigned [22]. For example, the relationship between COVID-19 infection status and complying with the protective measures (e.g., wearing masks). It is well known that none of the current test methods are perfect. If the information on the compliance with the recommended protective measure is collected by self report data, such information can be

subject to errors. Several reasons may account for such misclassification errors; these include the reluctance of respondents to provide correct information for fear it may be used against them later, despite the assurance of confidentiality before data collection. Another reason is recall bias, where respondents do not accurately remember or omit details of previous events or experiences.

Research Aims and Objectives

The aim of this thesis is to investigate dependence for joint misclassification error in both the response variable and the covariate.

The objectives of the study are to:

- Describe the framework for a joint dependent misclassification error model.
- Perform systematic simulation studies to assess the consequences of ignoring the dependence in joint misclassification.
- Extend the framework of a binary joint misclassification model to a multi-category response variable and a multi-category covariate.
- Illustrate the proposed model for joint misclassification in both the response variable and covariate through a real data example by establishing the true association between Trichomoniasis and Bacterial Vaginosis.

Misclassification Mechanism

Two methods are employed in the characterization of a misclassification process. The methods are differentiated by the choice of conditioning variables in the modeling processes. Let Y and X be the true response variable and covariate respectively, and let Y^* and X^* be their observed or surrogate version, respectively. In the first method, the error-prone variable is conditioned on the error-free variable through conditional probabilities, i.e., $P(Y^*|Y)$ for the response variable and $P(X^*|X)$ for the covariate. This method of characterizing misclassification error was first employed by Bross (1954) [5]. In a binary setting, correct classification

$P(Y^* = 1|Y = 1)$ and $P(Y^* = 0|Y = 0)$ are known as the sensitivity and specificity in the response variable and $P(X^* = 1|X = 1)$ and $P(X^* = 0|X = 0)$ are sensitivity and specificity in the covariate. Sensitivity is the probability that an individual with a condition is classified as having the condition, and specificity is the probability that an individual without a condition is classified as not having the condition. In a perfect situation where there is no misclassification error, both sensitivity and specificity are equal to one.

In the other characterization procedure, which was first considered by Marshall(1990)[34], misclassification errors are characterized differently. That is, for the binary setting, the correct classification $P(Y = 1|Y^* = 1)$ and $P(Y = 0|Y^* = 0)$ are known as the Positive Predictive Value (PPV) and Negative Predictive Value (NPV), respectively, for the response variable. Also, $P(X = 1|X^* = 1)$ and $P(X = 0|X^* = 0)$ are Positive Predictive Value and Negative Predictive Value, respectively, for the covariate. The PPV is the probability that an individual who was classified as having a condition had the condition, and NPV is the probability that an individual who was not classified as having a condition did not have the condition. In the response variable, when the true prevalence (i.e., P(X=1)), sensitivity, and specificity are known, the NPV and the PPV can be derived using Bayes' theorem.

$$\begin{aligned} PPV_X &= \frac{SN_X \cdot P(X=1)}{SN_X \cdot P(X=1) + (1-SP_X) \cdot (1-P(X=1))} \\ NPV_X &= \frac{SP_X \cdot (1-P(X=1))}{(1-SN_X) \cdot P(X=1) + SP_X \cdot (1-P(X=1))} \end{aligned}$$

In terms of covariate classification, this also holds when the true prevalence of the covariate P(Y=1) is known. It is worth noting that, although sensitivity and specificity may be high, translating it to predictive values may produce low values. As an example, say $SN_X = 0.95$ and $SP_X = 0.85$, when $P(X = 1) = 0.5$, the following predictive values are obtained; $PPV_X = 0.86$ and $NPV_X = 0.94$. These predictive values are generally high and can adequately characterize the misclassification process. However, when $P(X = 1) = 0.1$, which is the case for a rare event, the positive predictive values are quite low, $PPV_X = 0.41$ and the negative predictive values are very high, $NPV_X = 0.993$.

At this point, a distinction is made between **non-differential** and **differential** misclassification. Non-differential and differential misclassification can occur in either the covariate

or the response variable. Covariate misclassification is called non-differential if the true value of the response variable cannot provide additional information in the misclassification process; that is, $P(X^*|Y,X) = P(X^*|X)$, otherwise called differential misclassification of the covariate. Similarly, response variable misclassification is called non-differential if the true value of the covariate cannot provide additional information on the misclassification process; that is, $P(Y^*|Y,X) = P(Y^*|Y)$, otherwise called a differential misclassification of response. For example, differential misclassification can occur in a case-control study where a woman diagnosed with breast cancer may improve her diet; hence, her reported diet intake after diagnosis is correlated with cancer, even when the long-term diet has been taken into account. In the non-differential misclassification, the correct classification for the response variable is characterized by one probability instead of the two, i.e.,

$$\begin{aligned} P(Y^* = 1|Y = 1, X = 1) &= P(Y^* = 1|Y = 1, X = 0) = P(Y^* = 1|Y = 1), \\ P(X^* = 1|X = 1, Y = 1) &= P(X^* = 1|X = 1, Y = 0) = P(X^* = 1|X = 1). \end{aligned}$$

In joint misclassification, where both the response variable and covariate are subject to error, misclassification can be either *independent* or *dependent*. Independent misclassification occurs when the probability of the joint occurrence of any classification outcome concerning response status with any classification outcome concerning the covariate status given the true response and true covariate status, is equal to the product of corresponding classification probabilities for response and covariate separately [25]. Simple, classification errors do not correlate. That is,

$$P(Y^*, X^*|Y, X) = P(Y^*|Y) \cdot P(X^*|X), \tag{1.1}$$

otherwise, it is dependent.

1.2 Literature Review

It has long been recognized that misclassification errors can produce biased parameter estimates, and efforts have been made to examine the impact of ignoring these errors. Issues of misclassification can be categorized into three groups: (1) covariate only is subject to misclassification; (2) response variable only is subject to misclassification; and (3) the response

variable and covariate are subject to misclassification [56]. There is an extensive literature on covariate misclassification. The works of Carroll et al. (2006)[9], Gustafson (2003) [20], and references therein comprehensively discussed the impact of misclassification in covariates, as well as various correctional approaches. More recent literature is by Yi [56]; this book focuses on misclassification in covariates, and a portion covered misclassification in response variables applied in survival analysis and longitudinal data. Misclassification in response variables has received less attention compared to misclassification in covariates. Some studies have discussed and proposed approaches to correcting for misclassification error in response variables, to mention a few Magder et al. (1993) [33], Lyles et al. (2011) [32], Jurek et al.(2013) [23], Pekkanen et al.(2006)[40], Hausman et al.(1994)[21]. Relatively, there is minimal works on joint misclassification errors in both the response variable and covariate. The scope of this thesis will be misclassification errors in both the response variable and covariate.

Carroll et al. (2006) discussed the effects of misclassification as causing bias in parameter estimations, loss of power, and masking the features of data [9]. The pioneering work regarding misclassification can be attributed to Bross (1954) [5]. He posited that when two proportions are compared, misclassification tends to reduce the power of the significance test.

Subsequent development on misclassification after the classical paper of Bross (1954) [5] shows, misclassification has the potential of underestimating or overestimating the effects measures, thereby reducing or increasing the apparent strength of association. The magnitude and direction of the bias resulting from misclassification are dependent on the classification parameters (Positive Predictive Value and Negative Predictive Value or sensitivity and specificity [11], as well as the type of misclassification (Non-differential and Differential). In discussing the bias in the estimation of relative risk caused by misclassification, Coperland(1977) considered the case of both differential and non-differential misclassification for two types of epidemiological studies: cohort and case-control studies[14]. In both types of studies, non-differential misclassification produces a bias in the estimates towards the null; this is consistent with other researchers [25, 18] who made the same assertion. This assertion has been proven not always to hold, as it is applicable only when both response variable and covariate are binary, only one variable is subject to misclassification, and misclassification is non-differential and independent [17, 26, 35]. Differential misclassification of response vari-

able or covariate may bias estimation in either direction [14, 35, 24]. Gustafson and Greenand (2014) illustrated that non-differential misclassification in a situation where a trinary covariate is misclassified does not always induce bias towards the null [1]. The impacts of dependent non-differential misclassification were discussed by Kristensen[26] and Chavance[10]. In the work of Vogel et al. (2005) [53], the dependence structure of the joint misclassification was shown by a matrix composed of various dependence parameters. The works of Brenner et al. (1993)[4] and Vogel et al. (2005)[53] concluded that positive correlation of errors in the response variable and covariate might bias the exposure-disease association in any direction in the case of non-differential misclassification.

Barron(1977)[2] introduced the *matrix method* to intuitively correct for misclassification when both the response variable and covariate are binary. Variants and extensions of the matrix method have been proposed for correcting misclassification errors. Marshall (1990) [34] employed positive and negative predictive values as the correction identity for a covariate misclassification, instead of specificity and sensitivity used by Barron. Marshall's approach was later referred to as the "inverse matrix method" by Morrissey and Spiegelman (1999)[36] in a model efficiency study. Brenner et al.(1993) [4] and Vogel et al. (2005) [53] further extended the matrix approach of Barron (1977) [2] to a non-differential and dependent misclassification situation to correct for misclassification in the estimation of cumulative incidence and attributable risk, respectively. Barron's [2] matrix method is computationally straightforward; since there are no assumptions about the distribution of the true parameters, his approach can be considered a functional modeling approach. Carrol et al.(2006) [9] made a distinction between a *structural modeling approach* and a *functional modeling approach*, he asserted that for the former case, some parametric distribution assumptions are made about the true response and true covariate. A major limitation of the matrix method is that estimated probabilities may be invalid, that is, falls outside the constraint of 0 and 1, and this is as a result of matrix inversion [37]. Other functional approaches have been identified in the literature; the most common is the Simulation Extrapolation. This correction method was initially intended for measurement error in continuous covariate [13] but was later extended to misclassification error by Kuchenhoff (2006) [27].

Most structural approaches for estimating parameters of misclassified data are likelihood-

based methods. The likelihood-based methods are the Maximum Likelihood and Bayesian Method. The maximum likelihood method is the most widely used approach. Tang et al. (2013) [50] formulated generalizations of assumptions underlying the matrix and inverse matrix methods into a framework of maximum likelihood when internal validation data are available. Tang et al. (2015) [49] employed a maximum likelihood approach framework for a common misclassification in both the response variable and the covariate of interest while adjusting for other covariates through a logistic regression model. Morrissey and Spiegelman (1999) [36] compared the matrix method and the inverse matrix method with a maximum likelihood estimator (MLE) in a covariate misclassification. They asserted that, although the MLE is computationally intense, MLE was more efficient than both the direct matrix and direct inverse matrix method. However, the direct inverse method was more efficient than the direct matrix method. For a more comprehensive discussion on Maximum Likelihood Estimates, refer to Lyles (2002), Greenland (2008), and Carrol et al. (2006) [9, 19, 31]. Bayesian computation is mostly implemented with Markov chain Monte Carlo (MCMC) sampling. Bayesian method to adjust for covariate misclassification has been discussed in detail by Gustafson (2003) [20].

A common occurrence of joint misclassification is when both the response variable and covariate are obtained from unreliable sources. Self-reported survey data is commonly encountered with misclassification[28]. Clinical or laboratory criterion is more accurate in getting data than self- reported sources; however, obtaining clinical or laboratory criterion is more costly and time-consuming. Studies have shown a discordance between estimates obtained from self-reported data sources and a criterion standard such as clinical or laboratory examination [39, 3]. For example, if both the response variable and the covariate are obtained from self-reported responses, and respondents do not accurately remember or omit details of previous events or experiences, misclassification errors may occur in both variables. Joint misclassification can occur when information on both the response variable and covariate are obtained from proxy respondents.

Liu et al.[30] and Tarafder et al.[51] employed a Bayesian approach to correct for joint misclassification in both the response variable and covariate. However, they both assumed joint misclassification to be independent. Another instance to consider is when information

on both the response variable and covariate status are obtained from two sources; say, one subgroup's data is obtained from a personal interview, and another subgroup's information can only be obtained by proxy. Information from a proxy is usually less accurate, hence misclassification in both the response variable and covariate will be common in this subgroup. Thus, the dependence of classification errors is likely; therefore, the assumption of conditional independence is not always guaranteed. Brenner at al.(1993)[4] and Vogel et al.(1993) [53] considered the matrix framework for joint dependent misclassification errors in both the response variable and covariate. An example stated in Brenner et al. [4] obtained from Dales et al. (1991) [16] discusses home dampness and molds and its impacts on respiratory health in children. In this study, both the covariate, which is the presence of molds and the response variable, respiratory symptoms, were self-reported. Here the authors emphasized the possibility of under or over-reporting both the respiratory symptoms and exposure to molds depending on the health conciousness of the participants. The above example justifies a situation where dependence error misclassification can occur. This thesis seeks to address the consequences of violating the independence assumption in a joint misclassification error model.

1.3 Effect of Misclassification

Following, the effect of (i) covariate misclassification, and (ii) response variable misclassification are explored.

Covariate Misclassification

For a two binary classifier, Y and X, if misclassification occurs in only the covariate X, X^* is observed, instead of X. The relationship between X and X^* is described by sensitivity: $SN_X = P(X^* = 1|X = 1, Y)$ and specificity: $SP_X = P(X^* = 0|X = 0, Y)$ respectively. Consider a common occurrence in epidemiological studies in which $Y = 1$ indicates the presence of a disease of interest, and X represents an exposure under consideration. Let the prevalence of exposure amongst participants with disease and disease-free participants be; $r_1 = P(X = 1|Y = 1)$ and $r_0 = P(X = 1|Y = 0)$ respectively.

Usually, the inferential interest is to apply odd-ratio to ascertain the association between the covariate and response.

$$\Phi = \frac{r_1(1-r_0)}{r_0(1-r_1)}$$

However, due to misclassification in X, the naive odds ratio is estimated to be;

$$\Phi^* = \frac{r_1^*(1-r_0^*)}{r_0^*(1-r_1^*)}.$$

The attentuation factor arising as a result is [20],

$$\frac{\Phi^*}{\Phi} = \frac{\{1-s/r_1\}/\{1+t/(1-r_1)\}}{\{1-s/r_0\}/\{1+t/(1-r_0)\}}$$

where,

$$s = \frac{1-SN_X}{SN_X+SP_X-1} \qquad t = \frac{1-SP_X}{SN_X+SP_X-1}$$

The naive estimator is biased towards the null value of unity for non-differential misclassification conditional on $SP_X + SN_X - 1 > 0$, in that either $1 \leq \Phi^* \leq \Phi$ or $\Phi \leq \Phi^* \leq 1$[20]. However, in differential misclassification odds ratio may either be under or overestimated, bias may go in either direction[7].

Response Variable Misclassification

Before misclassification in response variable is considered, lets briefly discuss the situation where the response variable is continuous. When the response variable is a continuous variable, measurement error in Y is mostly ignored. Consider a linear regression model,

$$Y = \beta_0 + \beta_1 X + E. \tag{1.2}$$

Let X be an error-free covariate and E be the inter-subject variability having a variance of σ_E^2. Let the estimate of β_1 be $\hat{\beta}_1$ when Y is regressed on X. If Y^* is the surrogate of Y, that is, $Y^* = Y + \xi$, then $Y = Y^* - \xi$, where ξ has a mean of zero and a variance of σ_ξ^2. This gives,

$$Y^* = \beta_0 + \beta_1 X + (E + \xi) \tag{1.3}$$

The structure of the regression does not change, and ξ does not depend on (Y, X). The main difference is the variability, hence $var(Y^*) = \sigma_E^2 + \sigma_\xi^2$. This may be the reason why generally, the emphasis is placed on errors in covariate rather than errors in the response variable.

Unlike error in a response variable that is continuous, misclassification in a response variable has a different mechanism. Consider two binary classifiers Y and X, if we do not observe Y but rather an error-prone version Y^*. The misclassification probabilities can be expressed as, $P(Y^* = y^*|Y = y, X = x)$, where $y^*, x, y = 0, 1$. To distinguish between the probabilities, $SN_Y = 1 - P(Y^* = 0^*|Y = 1, X = x) = P(Y^* = 1^*|Y = 1, X = x)$ is the sensitivity of the response variable while, $SP_Y = 1 - P(Y^* = 1^*|Y = 0, X = x) = P(Y^* = 0^*|X = x, Y = 0)$ is the specificity. The observed response variable is modelled as:

$$\begin{aligned} P(Y^* = 1|X = j) &= \sum_{k=0}^{1} P(Y^* = 1|Y = k, X = j)P(Y = k|X = j) \\ &= (1 - SP_Y)P(Y = 0|X) + SN_Y P(Y = 1|X) \\ &= (SN_Y - (1 - SP_Y))P(Y = 1|X) + (1 - SP_Y) \end{aligned} \tag{1.4}$$

Given the sensitivity and specificity, one can straightforwardly obtain the effect estimates of the covariate. The model relating a response variable Y to a covariate X is usually a logistic regression,

$$\text{logit } P(Y = 1|X) = \beta_0 + \beta_1 X \tag{1.5}$$

but if misclassified model below if fitted,

$$\text{logit } P(Y^* = 1|X) = \beta_0^* + \beta_1^* X. \tag{1.6}$$

The relationship between β_1^* and β_1 are dependent on the type of misclassification. For differential misclassification in Y with respect to X, the bias can be in either direction , towards or away from the null value 0. On the hand, for nondifferential misclassification, the log odds ratio of β_1^* is attenuated relative to β_1^*[20].

1.4 Data Structure

Estimates of the parameters of interest may not always be reliable when inference is based on only information on Y^* and/or X^*. The joint distribution of (Y^*, Y) and/or (X^*, X) must be available or estimated to get precise parameter estimates; this leads to the notion of validation studies. Validation studies can be categorized into two types based on the data source: internal validation study and external validation study. For an internal validation

study, the study subjects who contribute to the validation data are a subsample of the main study. Typically, when perfect measures of both the response variable and covariates (Y, X) are available, then the error-prone variables (Y^*, X^*) becomes redundant in the model. However, measures for (Y, X) may be labor-intensive or costly, making it difficult to be obtained for every study participant. An example of an internal validation data is discussed by Carrol et al. (1993) [8]. The study sought to establish an association between exposure to simplex virus type 2 (HSV-2) and invasive cervical cancer. Invasive cervical cancer was considered error-free, while a refined western blot procedure was used to assess the exposure of some participants to HSV-2, and a less accurate western blot procedure was assessed for all participants. Data obtained from the refined western blot was considered the validation data, and the less accurate western blot procedure was the main study data. The structure of the data is given in Table (1.1). Validation data were available for about 6% of the study participants.

Table 1.1: HSV Data (Carrol et al.(1993))

Study	Y	X	X^*	Count
Validation Data	1	0	0	13
	1	0	1	3
	1	1	0	5
	1	1	1	18
	0	0	0	33
	0	0	1	11
	0	1	0	16
	0	1	1	16
Main Study Data	1	-	0	318
	1	-	1	375
	0	-	0	701
	0	-	0	535

In the absence of internal validation data, additional information can be obtained from

an external source. External validation studies employ samples from a different population. External validation data has the advantage of cost-effectiveness because mostly data is readily available from past studies. However, external validation misclassification error model are not necessarily transportable. Transportability always assumes the error structure in the main data and the external data are the same.

1.5 Estimation Methods

In drawing inference about the relationship between a true response and a covariate, two common approaches for data subject to misclassification error are likelihood-based: the Maximum Likelihood Estimation approach and the Bayesian Estimation approach. Bayesian methods allow the incorporation of historical information that is outside of the observed data. When there is little prior information (that is, very diffuse prior distribution), the Bayesian and MLE estimation are almost the same. Also, the Bayesian MCMC method provides transparent interval estimates for model parameters, especially when the model is very complex and subject to nonlinear constraints on parameters. In this dissertation, the Bayesian method is used to draw inference about the parameters of interest. Below is a brief review of Bayesian inference.

Bayesian Inference

In the Bayesian approach the parameter $\theta \in \Theta$ to be estimated is thought of as a random variable with an assigned probability distribution termed the prior distribution, denoted by $\pi(\theta)$. The prior distribution represents the information about the parameters which may be available before data are observed. The posterior distribution $\pi(\theta|x)$ which is the target of Bayesian inference is composed of prior distribution and likelihood function, $f_x(X|\theta)$. The Likelihood function contains information about the parameter from the available data $X = x$. Updating of the prior is achieved by employing the Bayes' rule giving the relation below:

$$\pi(\theta|\mathtt{x}) = \frac{f_{\mathtt{x}}(\mathtt{x}|\theta)\pi(\theta)}{f_{\mathtt{x}}(\mathtt{x})} \tag{1.7}$$

where $f_{\mathtt{x}}(\mathtt{x}) = \int\int f(\mathtt{x}|\theta)\pi(\theta)d\theta$ is the normalizing constant.

An essential part of Bayesian analysis is the prior specification for the unknown parameters in a model. Specification of priors is entirely subjective, and it is normally based on the nature of the problem or the researcher's view of the problem. Bayesian employs a computer-driven sampling method known as Markov Chain Monte Carlo (MCMC). MCMC has the advantage of drawing samples from some distribution without knowing all the properties of the distribution. It is especially helpful in the setting of Bayesian analysis when posterior distributions are complex. MCMC methods require constructing suitable chains; these chains are the simulated samples from the posterior distribution of the target population.

1.6 Contribution and Outline of Thesis

The impact of dependence on misclassification errors has been least explored; this work should be the first to assess the consequences of imposing a wrong misclassification model when both the response variable and covariate are misclassified. Chapter 2 addresses joint misclassification in both the response variable and covariate when both variables are binary. The model for joint misclassification is specified while assuming misclassification errors are non-differential. The specified model considers the dependence of error by including covariance-like parameters. Bayesian method for parameter estimations is proposed for model estimation. Simulation studies are conducted to assess the consequences of fitting the wrong misclassification model.

In Chapter 3, the binary variables are extended to a multi-category situation. Similar to the binary setting, the dependence of error is also captured by dependent parameters, and the Bayesian method is proposed for parameter estimation. Further, simulation studies are conducted for a trinary response variable and a trinary covariate to establish the impact of ignoring dependence of error or misclassification in a multi-category setting. Chapter 4 illustrates the proposed models discussed through a real data example by considering the association between a Trichomoniasis and Bacterial Vaginosis, using data from the HIV Epidemiology Research (HERS). Chapter 5 concludes with the main findings, limitations, and discussion of future work.

Chapter 2

Binary Misclassification in both the Response Variable and Covariate

This chapter examines the impact of mis-specifying joint misclassification errors on the regression co-efficients estimates when both the binary response variable and the binary covariate are subject to misclassification errors. Dependence of misclassification error is characterized by covariance-like parameters[53]. Bayesian MCMC method is used for model estimation. Section 2.1 presents the notations and preliminary concepts as well as the specification of the model for misclassification in both the response variable and covariate. In section 2.2, the Bayesian method for the estimation of the model parameters are discussed by specifying the likelihood functions and priors for the parameters. In Section 2.3, comprehensive simulation study to assess the consequences of ignoring the dependence of the joint misclassification errors in both the response variable and the covariate are conducted. The chapter concludes with a discussion in Section 2.4.

2.1 Model Specification

Let Y and Y^* denote the actual response variable of interest and its surrogate (i.e., error-prone) variable respectively. Let X and X^* denote the actual covariate of interest and its surrogate (i.e., error-prone) variable respectively. Here it is assumed that misclassification is non-differential, consequently, sensitivity in the response variable and covariate are given as $SN_Y = P(Y^* = 1|Y = 1)$ and $SN_X = P(X^* = 1|X = 1)$. Similarly, specificity in response variable and covariate are given as $SP_Y = P(Y^* = 0|Y = 0)$ and $SP_X = P(X^* = 0|X = 0)$.

If information on both the response variable and covariate status are obtained from the same source (e.g., questionnaire, biological specimen) their errors are likely to be dependent.

Following the notion of dependence of misclassification error defined by Vogel et al.(2005) [53], let D_{ij} , for $i, j = 0, 1$, represent the dependence parameters,

$$D_{ij} = P(Y^* = i, X^* = j|Y = i, X = j) - P(Y^* = i|Y = i)P(X^* = j|X = j). \tag{2.1}$$

Please note that in the binary case, the following are obtained:

$$P(Y^* = 1 - i, X^* = 1 - j|Y = i, X = j) - P(Y^* = 1 - i|Y = i)P(X^* = 1 - j|X = j) = D_{ij} \tag{2.2}$$

$$P(Y^* = i, X^* = 1 - j|Y = i, X = j) - P(Y^* = i|Y = i)P(X^* = 1 - j|X = j) = -D_{ij} \tag{2.3}$$

$$P(Y^* = 1 - i, X^* = j|Y = i, X = j) - P(Y^* = 1 - i|Y = i)P(X^* = j|X = j) = -D_{ij} \tag{2.4}$$

The dependence parameters are bounded, see Appendix A for the proof. The specific boundaries for the dependence parameters are given below [4],

$$D_{11} \in \Big[Max\big\{-SN_Y SN_X; -(1 - SN_Y)(1 - SN_X)\big\}, Min\big\{(1 - SN_Y)SN_X; SN_Y(1 - SN_X)\big\}\Big]; \tag{2.5}$$

$$D_{10} \in \Big[Max\big\{-SN_Y SP_X; -(1 - SN_Y)(1 - SP_X)\big\}, Min\big\{(1 - SN_Y)SP_X; SN_Y(1 - SP_X)\big\}\Big]; \tag{2.6}$$

$$D_{01} \in \Big[Max\big\{-SP_Y SN_X; -(1 - SP_Y)(1 - SN_X)\big\}, Min\big\{(1 - SP_Y)SN_X; SP_Y(1 - SN_X)\big\}\Big]; \tag{2.7}$$

$$D_{00} \in \Big[Max\big\{-SP_Y SP_X; -(1 - SP_Y)(1 - SP_X)\big\}, Min\big\{(1 - SP_Y)SP_X; SP_Y(1 - SP_X)\big\}\Big]. \tag{2.8}$$

Let $p^*_{ij} = P(Y^* = i, X^* = j)$ and $p_{kl} = P(Y = k, X = l)$, thanks to the total probability rule, one can easily derive the relationship between the two distribution:

$$p^*_{ij} = \sum_{k=0}^{1}\sum_{l=0}^{1} P(Y^* = i, X^* = j|Y = k, X = l)p_{kl}. \tag{2.9}$$

Each p^*_{ij} can be expressed as a function of SN_X, SP_X, SN_Y, SP_Y and dependence parameters D_{ij}. Below the equation for p^*_{11} is shown:

$$\begin{aligned} p^*_{11} &= \sum_{k=0}^{1}\sum_{l=0}^{1} P(Y^*=1, X^*=1|Y=k, X=l)p_{kl}. \qquad (2.10)\\ &= \Big[SN_Y SN_X + D_{11}\Big]p_{11} + \Big[SN_Y(1-SP_X) - D_{11}\Big]p_{10} \\ &\quad + \Big[(1-SP_Y)SN_X - D_{11}\Big]p_{01} + \Big[(1-SP_Y)(1-SP_X) - D_{11}\Big]p_{00}. \end{aligned}$$

Please see Appendix B for details of p^*_{10}, p^*_{01} and p^*_{00}. The misclassified joint probabilities described above can be expressed in matrix form as derived in Liu et al. (2020)[29]:

$$\boldsymbol{p}^* = (\boldsymbol{M_Y} \otimes \boldsymbol{M_X} + \boldsymbol{D})\boldsymbol{p}, \qquad (2.11)$$

where $\boldsymbol{p}^* = (p^*_{11}, p^*_{10}, p^*_{01}, p^*_{00})'$ and $\boldsymbol{p} = (p_{11}, p_{10}, p_{01}, p_{00})'$. Please note that the operator $\otimes$ is the Kronecker product. The matrices $\boldsymbol{M_Y}$ and $\boldsymbol{M_X}$ are as follows:

$$\boldsymbol{M_Y} = \begin{bmatrix} SN_Y & 1-SP_Y \\ 1-SN_Y & SP_Y \end{bmatrix}, \ \boldsymbol{M_X} = \begin{bmatrix} SN_X & 1-SP_X \\ 1-SN_X & SP_X \end{bmatrix}.$$

The dependence matrix $\boldsymbol{D}$, is composed of the dependence parameters D_{ij} defined in Eqs. (2.1) - (2.4) with the structure;

$$\boldsymbol{D} = \begin{bmatrix} D_{11} & -D_{10} & -D_{01} & -D_{00} \\ -D_{11} & D_{10} & D_{01} & D_{00} \\ -D_{11} & D_{10} & D_{01} & -D_{00} \\ D_{11} & -D_{10} & -D_{01} & D_{00} \end{bmatrix}.$$

When misclassification errors are independent, all entries of the dependence matrix are zero and (2.11) becomes,

$$\boldsymbol{p}^* = (\boldsymbol{M_X} \otimes \boldsymbol{M_Y})\boldsymbol{p}. \qquad (2.12)$$

In this thesis, data structure with validation data is considered. Let's consider a study of N participants (sampling objects), with both the response variable and the covariate subject to misclassification errors. Assuming n_v of the N participants have observations

made with an error-free measure in addition to the error-prone measure (where, $N \geq n_v$), the n_v participants constitutes the validation study subjects. The remaining $n_m = N - n_v$ participants are the main study subjects. For the validation dataset, all four assessments $(Y^* = i, X^* = j, Y = k, X = l)$ are observed for each sampling unit. These assessments are binary and takes on 0 and 1, hence 16 distinct patterns of the validation data are derived;

$$\begin{aligned} p_{ijkl} &= P(Y^* = i, X^* = j, Y = k, X = l), \\ &= P(Y^* = i, X^* = j | Y = k, X = l) P(Y = k, X = l), \\ &= (P(Y^* = i | Y = k) P(X^* = j | X = l) + D_{ijkl}) P(Y = k, X = l). \end{aligned} \tag{2.13}$$

On the other hand, for the main data set, $(Y^* = i, X^* = j)$ are observed for each sampling unit, and 4 distinct patterns are derived;

$$p^*_{ij} = P(Y^* = i, X^* = j). \tag{2.14}$$

The data layout for joint misclassification in both the response variable and covariate when validation data is available is given in Table (2.1).

Table 2.1: Data layout for joint misclassification in both the response variable and covariate when validation data is available

Study	Sampling unit	X	Y	X^*	Y^*
Validation	1	✓	✓	✓	✓
	·	✓	✓	✓	✓
	·	✓	✓	✓	✓
	·	✓	✓	✓	✓
	n_v	✓	✓	✓	✓
Main	n_{v+1}	-	-	✓	✓
	·	-	-	✓	✓
	·	-	-	✓	✓
	·	-	-	✓	✓
	N	-	-	✓	✓

2.2 Bayesian method of adjustment for misclassification error

2.2.1 Likelihood functions

The joint probabilities p_{ij} are parameterized through a logistic regression model with a response variable and a covariate. Let us assume the underlying logistic regression model of interest is:

$$logit(P(Y = 1|X = x)) = \beta_0 + \beta_1 X. \tag{2.15}$$

Employing the relationship below;

$$P(Y = i, X = j) = \frac{exp(i(\beta_0 + \beta_1 j))}{1 + exp(\beta_0 + \beta_1 j)}(p_X)^i(1 - p_X)^{1-i}, \tag{2.16}$$

where, $p_X = P(X = 1)$, the p_{ij} are reparameterized into regression parameters $\beta_k, k = 0, 1$. Let n_{ijkl} be the total number of individuals having $(Y^* = i, X^* = j, Y = k, X = l)$ and n_{ij} be the total number of individuals having $(Y^* = i, X^* = j)$.

Let $\boldsymbol{\theta}$ represent the parameters of interest for a model that considers dependence of misclassification errors (Eq. 2.11) , that is, $\boldsymbol{\theta} = (SN_Y, SN_X, SP_Y, SP_X, D_{11}, D_{10}, D_{01}, D_{00}, \beta_0, \beta_1)$. The likelihood function for the validation data, L_v is given by:

$$L_v(\boldsymbol{\theta}|Y, X, Y^*, X^*) = \prod_{ijkl} P(Y^* = i, X^* = j, Y = k, X = l)^{n_{ijkl}}. \tag{2.17}$$

Let $\boldsymbol{\eta}$ represent the parameters of interest for a model that ignores dependence of misclassification errors (Eq.2.12) , that is, $\boldsymbol{\eta} = (SN_Y, SN_X, SP_Y, SP_X, \beta_0, \beta_1)$. The likelihood function for the validation data, L_v is given by:

$$L_v(\boldsymbol{\eta}|Y, X, Y^*, X^*) = \prod_{ijkl} P(Y^* = i, X^* = j, Y = k, X = l)^{n_{ijkl}}. \tag{2.18}$$

Further, the main data's likelihood function which is explicitly based on (Y^*, X^*) is given by:

$$L_m(\boldsymbol{\theta}|Y^*, X^*) = \prod_{ij}(P(Y^* = i, X^* = j|Y = k, X = l)P(Y = k, X = l))^{n_{ij}}, \tag{2.19}$$

and

$$L_m(\boldsymbol{\eta}|Y^*, X^*) = \prod_{ij}(P(Y^* = i, X^* = j|Y = k, X = l)P(Y = k, X = l))^{n_{ij}} \tag{2.20}$$

respectively for when dependence of misclassification errors is considered and when dependence of misclassification errors is ignored. The overall likelihood function used in parameter estimation is proportional to the product of the validation data's likelihood and the main data's likelihood, that is, $L_v \times L_m$.

2.2.2 Prior Construction

1. Priors for SN_Y, SP_Y, SN_X, SP_X: Truncated beta distributions are assigned to the priors of the misclassification parameters. The beta distribution is suitable because the misclassification parameters are probabilities defined on the interval [0,1]. In practice, however, it is rare to encounter sensitivities and specificities less than 0.5 [30]. Sensitivity and Specificity are critical in making clinical decisions. Clinicians are interested in knowing how well a test distinguishes between patients who have a disease and those who do not have. A low sensitivity test misses a lot of positives whiles giving high false negative rate (type 2 errors) and a low specficity test misses a lot of negatives whiles giving high false positive rate (type 1 error). For this reason the distribution is truncated to lie within [0.5,1], that is;

$$\begin{aligned}
SN_Y &\sim Beta(\alpha_{_{SN_Y}}, \beta_{_{SN_Y}})I(SN_Y > 0.5);\\
SP_Y &\sim Beta(\alpha_{_{SP_Y}}, \beta_{_{SP_Y}})I(SP_Y > 0.5);\\
SN_X &\sim Beta(\alpha_{_{SN_X}}, \beta_{_{SN_X}})I(SN_X > 0.5);\\
SP_X &\sim Beta(\alpha_{_{SP_X}}, \beta_{_{SP_X}})I(SP_X > 0.5);
\end{aligned}$$

   where, $I(SN_Y > 0.5), I(SP_Y > 0.5), I(SN_X > 0.5)$ and $I(SP_X > 0.5)$ are indicator functions with value equal to 1 if the input is greater than 0.5 and 0 otherwise. An equal-tailed 95% CI (0.55,0.95) is used to obtain the priors for the misclassification parameters. Numerical methods are employed to estimate the hyperparameters.

2. Priors for β_k, where $k = 0, 1$: The priors of the regression parameter are weakly

informative priors,

$$\beta_k \sim N(0, 1000), k = 0, 1.$$

3. Priors for $D_{11}, D_{10}, D_{01}, D_{00}$: The dependence parameters are constrained to lie within an interval determined by the misclassification parameters (2.6) - (2.8), thus, uniform distributions are assigned to the dependence parameters which is also constrained within (2.6) - (2.8).

Given the complexity of the model with multi-parameters, it is not feasible to obtain the posterior estimates of the parameters analytically; hence Markov Chain Monte Carlo (MCMC) sampling, implemented in R via "Just Another Gibbs Sampler" (JAGS), is used. JAGS is a statistical program that implements MCMC methods [42].

2.3 Simulation Study

In this section, simulation studies are conducted with the aim of checking the consequences of fitting an independent misclassification error model and a naive model to data generated from a dependent misclassification error model. The fitted models are,

1. The model for **dependent misclassification errors**. In this model, the misclassification errors in the response variable depends on the misclassification errors in the covariate and vice versa.

2. The model for **independent misclassification errors** in the response variable and the covariate.

3. The **naive model** which assumes no misclassification error.

The scenarios for the simulation studies are selected based on:

- varying the misclassification parameters. Large values of the SN_X, SP_X, SP_Y, SP_X corresponds to less amount of misclassification.

- varying the extent of dependence. Here, the function $\delta = \sum_{ij}(-1)^{i+j} D_{ij} P_{ij}$ is used to control the dependence in the model. Please refer to section(2.3.2) for details of the δ function.

- varying the proportion of the validation data $\frac{n_v}{N}$; 10% proportion of validation data and 50% proportion of validation data.

2.3.1 Simulation Setup

Based on the logistic regression model,

$$\text{logit } P(Y = 1|x) = \beta_0 + \beta_1 X,$$

one can derive the joint distribution for (Y, X), p_{ij}. I set the values of β_0 and β_1 to be 1 and $P(X = 1) = 0.1$. To introduce non-differential and dependent misclassification errors in Y and X, I set up the true value of M_Y,M_X, and D. In binary misclassification with validation, the 16 distinct patterns of the validation data are derived from $P(Y^* = i, X^* = j, Y = k, X = l)$ where $i, j, k, l = 0, 1$.

In the simulation studies, two different proportions of the validation data are considered, and each has a sample size 10,000 including both main data and validation data. (a) 10% validation data: 9000 are main data observations, that is, there are observations for only $Y^*$and X^*, and 1000 are validation data observations, that is observations for all $Y^*, X^*, Y$and X. (b) 50% of the sampling unit as validation data; both the main data and the validation data have 5000 observations. There are eight scenarios for each proportion of the validation data, and repeatedly 1000 data sets are generated for each scenario for the simulation study. The average of each of the parameter estimates out of the 1000 datasets for each fitted model are calculated.

2.3.2 Choice of Dependence

High and low dependence are chosen based on the optimization of the function below:

$$\delta = \sum_{ij} (-1)^{i+j} D_{ij} P_{ij}. \tag{2.21}$$

The above function is equal to $\delta = E(Cov(Y^*, X^*|Y, X))$[30] and can be extracted from Eq. (2.1). Please see Appendix C for proof details. Table (2.2) gives the Low and high values for the dependence parameters of various misclassification scenarios considered. Please note

that a scaled version of δ, is defined by:

$$\delta_r = E(corr((Y^*, X^*|Y, X))) = \frac{E(Cov(Y^*, X^*|Y, X)}{\sqrt{Var(Y^*|Y, X)Var(X^*|Y, X)}}.$$

Table 2.2: Low and High dependence values for the dependence parameters based on the misclassification scenarios.

Misclassification scenario	**Parameters**	**Low**	**High**
	D_{11}	-0.0086	0.1600
$SNY = 0.8$, $SNX = 0.8$	D_{10}	0.0418	-0.0400
$SPY = 0.8$, $SPX = 0.8$	D_{01}	0.0126	-0.0400
	D_{00}	0.1174	0.1600
	δ_r	1.186e-18	0.0995
	D_{11}	-0.0054	0.1600
$SNY = 0.8$, $SNX = 0.8$	D_{10}	0.0141	-0.0100
$SPY = 0.95$, $SPX = 0.95$	D_{01}	0.0116	-0.0100
	D_{00}	0.0408	0.0475
	δ_r	2.486e-19	0.0607
	D_{11}	-0.0011	0.0475
$SNY = 0.95$, $SNX = 0.95$	D_{10}	0.0241	-0.0100
$SPY = 0.8$, $SPX = 0.8$	D_{01}	0.0155	-0.0100
	D_{00}	0.0667	0.1600
	δ_r	7.7726e-19	0.0370
	D_{11}	-0.0003	0.0475
$SNY = 0.8$, $SNX = 0.8$	D_{10}	0.0149	-0.0025
$SPY = 0.95$, $SPX = 0.95$	D_{01}	0.0115	-0.0025
	D_{00}	0.04111	0.0475
	δ_r	5.7063e-20	0.0199

2.3.3 Simulation Results

The result of the simulation studies that are aimed at checking the consequences of fitting the wrong models (an independent misclassification error model and a naive model) to data generated from a dependent misclassification model is presented below. The results are presented based on: (1) 10% proportion of validation data, and (2) 50% proportion of validation data.

Results for simulation studies employing 10% validation data

Tables (2.3) -(2.6) shows the average posterior means and the 95% credible intervals for the regression parameters (β_0 and β_1), misclassification parameters (SN_Y, SN_X, SP_Y, SN_X) and the D-parameters(D_{11},,$D_{10}D_{01}$,D_{00}) in the three models for each of the 8 simulation scenarios when 10% validation data is employed for a 1000 replicated data. Also included in the tables are the average relative bias for each parameter. Note that, the relative bias for θ is defined as $\left|\frac{\theta_{true}-\hat{\theta}}{\theta_{true}}\right|$. The *naive* model has estimates for only β_0 and β_1. Please note that the primary interest parameter is β_1 because the estimate of this parameter gives the mathematical relationship between the response variable and covariate.

Table (2.3) shows the results of low sensitivity and specificity scenarios (that is, $SN_Y = SN_X = SP_Y = SP_X = 0.8$). It is observed that scenario 2 (low δ) produces estimates that are closer to the true values than scenario 1 (high δ). The estimates of the misclassification parameters for both the *dependent misclassification error model* and the *independent misclassification error model* under scenario 2 are relatively close to the true value. However, in scenario 1 although the misclassification parameters estimates for the *dependent misclassification error model* are close to the true value, notably high mean values for the *independent misclassification error model* (that is, $S\hat{N}_Y = 0.924$, $SN_X = 0.936$, $SP_Y = 0.899$, $SP_X = 0.931$) is observed. The dependence parameters are reasonably close to the true value for both scenarios 1 and 2 when fitting the dependent misclassification error model.

Table 2.3: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive* Model using simulated data (N=10,000) with 10% validation data using settings from scenario 1 (High Dependence) and scenario 2 (Low Dependence).

		Dependent Model			**Independent Model**			***Naive* Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 1 (*High* δ)										
β_0	1	0.993	(0.875 , 1.117)	0.007	0.286	(0.212 , 0.359)	0.714	0.221	(0.177 , 0.264)	0.779
β_1	1.000	1.352	(0.905 , 1.885)	0.352	4.670	(4.107 , 5.321)	3.670	3.229	(3.009 , 3.457)	2.229
SNX	0.800	0.794	(0.759 , 0.822)	0.008	0.936	(0.910 , 0.958)	0.170	-	-	-
SNY	0.800	0.797	(0.779 , 0.816)	0.003	0.924	(0.913 , 0.935)	0.155	-	-	-
SPX	0.800	0.798	(0.785 , 0.811)	0.003	0.931	(0.922 , 0.940)	0.164	-	-	-
SPY	0.800	0.800	(0.783 , 0.815)	0.001	0.899	(0.874 , 0.921)	0.123	-	-	-
D_{11}	0.160	0.153	(0.135 , 0.167)	0.045	-	-	-	-	-	-
D_{10}	-0.040	-0.036	(-0.041 , -0.031)	0.093	-	-	-	-	-	-
D_{01}	-0.040	-0.001	(-0.041 , 0.071)	0.974	-	-	-	-	-	-
D_{00}	0.160	0.161	(0.152 , 0.169)	0.005	-	-	-	-	-	-
Scenario 2 (*Low* δ)										
β_0	1.000	1.000	(0.891 , 1.114)	0.000	1.002	(0.879 , 1.128)	0.002	0.622	(0.576 , 0.667)	0.378
β_1	1.000	1.083	(0.488 , 1.758)	0.083	1.092	(0.543 , 1.729)	0.092	0.107	(0.014 , 0.199)	0.893
SNX	0.800	0.797	(0.721 , 0.864)	0.004	0.799	(0.725 , 0.865)	0.001	-	-	-
SNY	0.800	0.800	(0.781 , 0.819)	0.000	0.799	(0.779 , 0.820)	0.001	-	-	-
SPX	0.800	0.801	(0.786 , 0.816)	0.001	0.801	(0.786 , 0.815)	0.001	-	-	-
SPY	0.800	0.801	(0.766 , 0.833)	0.001	0.800	(0.755 , 0.842)	0.000	-	-	-
D_{11}	-0.009	-0.003	(-0.029 , 0.029)	0.613	-	-	-	-	-	-
D_{10}	0.042	0.041	(0.032 , 0.051)	0.009	-	-	-	-	-	-
D_{01}	0.013	0.027	(-0.031 , 0.100)	1.188	-	-	-	-	-	-
D_{00}	0.117	0.115	(0.095 , 0.133)	0.023	-	-	-	-	-	-

Table 2.4: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 10% validation data using settings from scenario 3 (High Dependence) and scenario 4 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 3 (*High* δ)										
β_0	1.000	1.022	(0.936 , 1.114)	0.022	0.700	(0.620 , 0.778)	0.300	0.783	(0.735 , 0.830)	0.217
β_1	1.000	1.246	(0.844 , 1.734)	0.246	4.007	(3.444 , 4.659)	3.007	2.528	(2.325 , 2.737)	1.528
SNX	0.950	0.927	(0.886 , 0.954)	0.024	0.961	(0.937 , 0.979)	0.011	-	-	-
SNY	0.950	0.942	(0.927 , 0.955)	0.009	0.977	(0.970 , 0.983)	0.028	-	-	-
SPX	0.800	0.797	(0.782 , 0.811)	0.004	0.900	(0.888 , 0.912)	0.125	-	-	-
SPY	0.800	0.796	(0.778 , 0.812)	0.005	0.840	(0.805 , 0.873)	0.050	-	-	-
D_{00}	0.160	0.158	(0.148 , 0.168)	0.01	-	-	-	-	-	-
D_{01}	-0.010	0.013	(-0.015 , 0.054)	2.337	-	-	-	-	-	-
D_{10}	-0.010	-0.011	(-0.014 , -0.007)	0.051	-	-	-	-	-	-
D_{11}	0.048	0.048	(0.032 , 0.062)	0.008	-	-	-	-	-	-
Scenario 4 (*Low* δ)										
β_0	1	1.007	0.918) , 1.097)	0.007	1.011	(0.909 , 1.113)	0.011	1.084	(1.033 , 1.134)	0.084
β_1	1.000	1.107	(0.537 , 1.760)	0.107	1.111	(0.633 , 1.680)	0.111	0.246	(0.142 , 0.350)	0.754
SNX	0.95	0.906	(0.853 , 0.949)	0.046	0.91	(0.857 , 0.952)	0.042	-	-	-
SNY	0.95	0.945	(0.932 , 0.957)	0.005	0.945	(0.930 , 0.958)	0.006	-	-	-
SPX	0.800	0.796	(0.782 , 0.812)	0.004	0.797	(0.783 , 0.812)	0.003	-	-	-
SPY	0.800	0.793	(0.752 , 0.831)	0.009	0.794	(0.750 , 0.835)	0.008	-	-	-
D_{00}	0.067	0.068	(0.048 , 0.089)	0.025	-	-	-	-	-	-
D_{01}	0.016	0.027	(-0.015 , 0.079)	0.744	-	-	-	-	-	-
D_{10}	0.024	0.025	(0.018 , 0.033)	0.053	-	-	-	-	-	-
D_{11}	-0.001	0.009	(-0.004 , 0.029)	8.834	-	-	-	-	-	-

Comparing the three models, for both scenarios considered in the Table (2.3), it is noticed that for the β_1 parameter, estimates from the *dependent misclassification error model* are closer to the true values than the other two models. However, the *naive model* is superior to the *independent misclassification error model* when δ is high. The model that produced the largest bias in Table(2.3) for the β_1 parameter is the *independent misclassification error model* with high δ ($\hat{\beta}_1$=4.670). A look at Table (2.4), where high sensitivity (that is, $SN_Y = SN_X = 0.95$) and low specificity (that is, $SP_Y = SP_X = 0.8$) scenarios are considered, the *dependent misclassification error model* performs better than the *independent misclassification error model* and the *naive model* for both scenarios 3 and scenario 4. Estimates of the misclassification parameters in both scenario 3 and scenario 4 are a bit close to the true values, also the D parameters are close to the true values except D_{11} in scenario 4 which has a rather high relative bias of 8.834. The parameter estimates of β_1 for the *dependent misclassification error model* and *independent misclassification error model*s are somewhat close for scenario 4. It is observed that the *naive model* gives β_1 estimates that are lower than the true mean values for the high δ scenario in Table (2.4), a similar observation was made in scenarios 5 where low δ scenario was also considered. The best performing model for β_1 is the *dependent misclassification error model* with low δ ($\hat{\beta}_1$=1.107) and the model that produce the largest bias for β_1 is the *independent misclassification model* with high δ ($\hat{\beta}_1$=4.007)).

From Table (2.5) where low sensitivity (that is, $SN_Y = SN_X = 0.8$) and high specificity (that is, $SP_Y = SP_X = 0.95$) scenarios are considered, patterns for the misclassification and D-parameters are very similar to what is observed in Table (2.4). Although β_1 estimates reduced in the *independent model* for the high δ scenario, it produce largest bias in estimating the true β_1 value ($\hat{\beta}_1 = 3.512$). When the δ is low, both the dependent and independent misclassification error model provide accurate point estimate for the β_1 parameter.

For high sensitivity and specificity scenarios (that is, $SN_Y = SN_X = SP_Y = SP_X = 0.95$), it is shown from the Table (2.6) that both the dependent misclassification model and the independent misclassification model accurately estimates the true values of the parameters under consideration fro scenario 8 which considers a low δ scenario.

Table 2.5: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 10% validation data using settings from scenario 5 (High Dependence) and scenario 6 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 5 (*High* δ)										
β_0	1	1.001	(0.893 , 1.114)	0.001	0.6	(0.520 , 0.682)	0.4	0.322	(0.281 , 0.362)	0.678
β_1	1	content...1.233	(0.841 , 1.712)	0.233	3.512	(2.957 , 4.154)	2.512	2.197	(1.992 , 2.409)	1.197
SNX	0.800	0.793	(0.757 , 0.823)	0.009	0.847	(0.795 , 0.893)	0.059		-	-
SNY	0.800	0.797	(0.776 , 0.817)	0.004	0.865	(0.848 , 0.882)	0.082		-	-
SPX	0.950	0.944	(0.933 , 0.954)	0.006	0.978	(0.972 , 0.983)	0.029		-	-
SPY	0.95	0.941	(0.925 , 0.955)	0.009	0.953	(0.931 , 0.971)	0.003		-	-
D_{00}	0.048	0.05	(0.040 , 0.060)	0.048	-	-	-		-	-
D_{01}	-0.010	0.011	(-0.011 , 0.041)	2.072	-	-	-		-	-
D_{10}	-0.010	-0.01	(-0.013 , -0.007)	0.008	-	-	-		-	-
D_{11}	0.160	0.153	(0.134 , 0.168)	0.044	-	-	-		-	-
Scenario 6 (*Low* δ)										
β_0	1	0.997	(0.893 , 1.107)	0.003	0.987	(0.879 , 1.099)	0.013	0.46	(0.418 , 0.501)	0.54
β_1	1	1.087	(0.535 , 1.726)	0.087	1.109	(0.665 , 1.644)	0.109	0.323	(0.201 , 0.445)	0.677
SNX	0.8	0.785	(0.714 , 0.849)	0.019	0.789	(0.720 , 0.852)	0.013	-	-	-
SNY	0.8	0.798	(0.777 , 0.818)	0.003	0.797	(0.776 , 0.818)	0.004	-	-	-
SPX	0.95	0.945	(0.934 , 0.956)	0.005	0.947	(0.936 , 0.957)	0.004	-	-	-
SPY	0.95	0.94	(0.918 , 0.957)	0.011	0.932	(0.902 , 0.958)	0.019	-	-	-
D_{00}	0.041	0.043	(0.031 , 0.055)	0.065	-	-	-	-	-	-
D_{01}	0.012	0.016	(-0.010 , 0.047)	0.411	-	-	-	-	-	-
D_{10}	0.014	0.015	(0.008 , 0.022)	0.066	-	-	-	-	-	-
D_{11}	-0.005	-0.001	(-0.029 , 0.032)	0.756	-	-	-	-	-	-

Table 2.6: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 10% validation data using settings from scenario 7 (High Dependence) and scenario 8 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 7 ($High\ \delta$)										
β_0	1	1.021	(0.942 , 1.106)	0.021	0.866	(0.794 , 0.938)	0.134	0.813	(0.769 , 0.857)	0.187
β_1	1.000	1.180	(0.823 , 1.628)	0.18	2.805	(2.285 , 3.420)	1.805	1.646	(1.455 , 1.843)	0.646
SNX	0.950	0.924	(0.881 , 0.953)	0.027	0.925	(0.881 , 0.960)	0.027	-	-	-
SNY	0.950	0.940	(0.925 , 0.955)	0.01	0.958	(0.946 , 0.968)	0.008	-	-	-
SPX	0.950	0.943	(0.931 , 0.954)	0.008	0.966	(0.958 , 0.973)	0.017	-	-	-
SPY	0.950	0.940	(0.922 , 0.954)	0.011	0.938	(0.910 , 0.961)	0.013	-	-	-
D_{00}	0.048	0.051	(0.040 , 0.062)	0.072	-	-	-	-	-	-
D_{01}	-0.003	0.019	(-0.003 , 0.049)	8.543	-	-	-	-	-	-
D_{10}	-0.003	-0.002	(-0.004 , 0.002)	0.3	-	-	-	-	-	-
D_{11}	0.048	0.049	(0.031 , 0.064)	0.023	-	-	-	-	-	-
Scenario 8 ($Low\ \delta$)										
β_0	1.000	1.007	(0.935 , 1.083)	0.007	0.997	(0.915 , 1.081)	0.003	0.899	(0.854 , 0.944)	0.101
β_1	1.000	1.137	(0.637 , 1.733)	0.137	1.138	(0.777 , 1.578)	0.138	0.523	(0.388 , 0.66)	0.477
SNX	0.95	0.903	(0.849 , 0.947)	0.050	0.909	(0.856 , 0.951)	0.043	-	-	-
SNY	0.95	0.943	(0.929 , 0.956)	0.007	0.943	(0.928 , 0.957)	0.007	-	-	-
SPX	0.95	0.942	(0.931 , 0.954)	0.008	0.945	(0.934 , 0.956)	0.005	-	-	-
SPY	0.95	0.937	(0.916 , 0.955)	0.013	0.932	(0.901 , 0.957)	0.019	-	-	-
D_{00}	0.041	0.046	(0.033 , 0.058)	0.117	-	-	-	-	-	-
D_{01}	0.012	0.023	(-0.003 , 0.055)	1.021	-	-	-	-	-	-
D_{10}	0.015	0.017	(0.011 , 0.023)	0.139	-	-	-	-	-	-
D_{11}	-0.000	0.01	(-0.004 , 0.031)	37.042	-	-	-	-	-	-

The point estimates of the misclassification and D-parameters in scenario 7 and scenario 8 remain reasonably close to the true values. The point estimates in scenario 7 (low δ) are better than the point estimates in scenario 8 (high δ). The point estimate of β_1 for the naive model is very low compared to the true value. ($\hat{\beta}_1$=2.805).

For easy visual comparison, Figure (2.1) shows the graph of the relative bias of the estimated misclassification parameters and the regression coefficients. Generally, it is observed that the misclassification parameters have a minimal relative bias, especially in models with low δ (that is, B, D, F, H). Also the relative bias for the β_1 is high for models with high δ (that is, A, C, E, G). However, the value reduces as the misclassification error decreases. For the low misclassification parameters, the graph of the relative bias has almost identical patterns irrespective of the δ value; this is evident in subfigures G and H of Figure (2.1).

Figure 2.1: Graph of the relative bias (in absolute value) of the misclassification parameters and the regression coefficient with 10% validation.

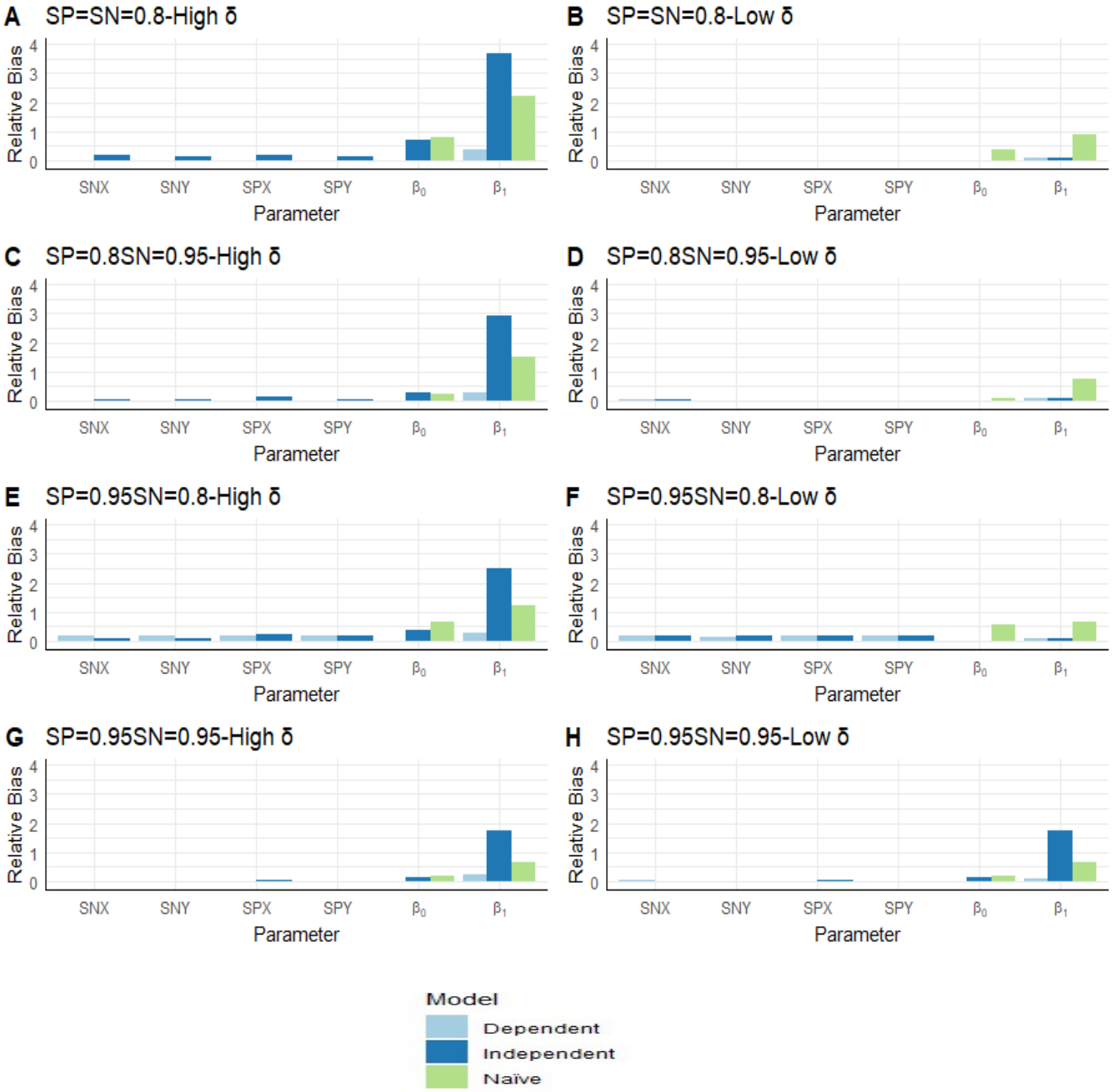


Table 2.7: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 50% validation data The table employs settings from scenario 1 (High Dependence) and scenario 2 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 1 (*High* δ)										
β_0	1.000	0.999	(0.940 , 1.059)	0.0007	0.936	(0.876 , 0.996)	0.064	0.440	(0.177 , 0.477)	0.560
β_1	1.000	1.048	(0.814 , 1.292)	0.0481	1.629	(1.355 , 1.915)	0.629	2.681	(3.009 , 2.863)	1.681
SNX	0.800	0.800	(0.787 , 0.812)	0.000	0.825	(0.794 , 0.855)	0.032	-	-	-
SNY	0.800	0.800	(0.789 , 0.810)	0.000	0.808	(0.797 , 0.819)	0.010	-	-	-
SPX	0.800	0.800	(0.791,0.809)	0.000	0.809	(0.800 , 0.819)	0.012	-	-	-
SPY	0.800	0.800	(0.790,0.809)	0.000	0.811	(0.791,0.830)	0.013	-	-	-
D_{11}	0.160	0.158	(0.150,0.165)	0.014	-	-	-	-	-	-
D_{10}	-0.040	-0.040	(-0.042 ,-0.037)	0.006	-	-	-	-	-	-
D_{01}	-0.040	-0.026	(-0.040 ,0.004)	0.339	-	-	-	-	-	-
D_{00}	0.160	0.159	(0.154, 0.165)	0.004	-	-	-	-	-	-
Scenario 2 (*Low* δ)										
β_0	1.000	1.000	(0.943,1.059)	0.0004	1.001	(0.941,1.062)	0.001	0.732	(0.694,0.771)	0.268
β_1	1	1.014	(0.746, 1.297)	0.0143	1.014	(0.744 , 1.297)	0.014	0.121	(0.035 , 0.207)	0.879
SNX	0.800	0.800	(0.765 , 0.833)	0.000	0.801	(0.766,0.833)	0.001	-	-	-
SNY	0.800	0.800	(0.789 , 0.811)	0.000	0.800	(0.789, 0.811)	0.000	-	-	-
SPX	0.800	0.800	(0.791 , 0.809)	0.000	0.800	(0.791,0.809)	0.000	-	-	-
SPY	0.8	0.800	(0.783,0.816)	0.000	0.800	(0.779, 0.820)	0.000	-	-	-
D_{11}	-0.008	-0.007	(-0.020 , 0.007)	0.165	-	-	-	-	-	-
D_{10}	0.042	0.042	(0.036 , 0.047)	0.000	-	-	-	-	-	-
D_{01}	0.013	0.016	(-0.019 , 0.058)	0.285	-	-	-	-	-	-
D_{00}	0.117	0.117	(0.108 , 0.126)	0.000	-	-	-	-	-	-

Results for simulation studies employing 50% validation data

Tables (2.7) -(2.10) show the posterior means, 95% credible intervals, and the relative bias of the parameters in the three models for each of the eight simulation scenarios for the case where 50% validation data is used.

In Table (2.7) low sensitivity and specificity scenarios (that is, $SN_Y = SN_X = SP_Y = SP_X = 0.8$) are shown. It is observed that in scenario 2 (low δ), both the dependent and independent misclassification error model accurately estimate the true values of the misclassification parameters. In scenario 1 (high δ), the dependent misclassification error accurately estimates the true values of the misclassification parameters, although estimates from the independent misclassification model are quite close to the true value. For the β_1 parameter, the point estimates of the misclassification parameters in both scenario 1 and scenario 2 are relatively close to the true parameters. The dependent and independent misclassification error model gave accurate point estimates for β_1 parameter in scenario 1. The model that produced the largest bias is the *naive model* with high δ ($\hat{\beta}_1$=2.681).

For scenarios that considers high sensitivity (that is, $SN_Y = SN_X = 0.95$), and low specificity (that is $SP_Y = SP_X = 0.8$) scenarios, as is the situation in the Table (2.8), it is observed that the *dependent model* remains the best model. However, there is minimal distinction between the *dependent model* and the *independent model* for low dependence error. The model that produce largest bias in estimating the β_1 parameter is the *naive model* with high dependence error ($\hat{\beta}_1$=2.196). Patterns observed in Table (2.9) are very similar to (2.8) which considers low sensitivity (that is $SN_Y = SN_X = 0.8$) and high specificity (that is $SP_Y = SP_X = 0.95$) scenarios. Although the $\hat{\beta}_1$ estimates have reduced compared to preceding models the *naive model* with high δ remains the model that produces the largest bias in estimating the $beta_1$ parameter ($\hat{\beta}_1$=1.889). Considering high misclassification scenarios (that is, $SN_Y = SN_X = SP_Y = SP_X = 0.95$) as recorded in the Table (2.10), generally the there is a reduction in the relative bias values. The model produce largest bias in estimating the β_1 parameter is the *independent misclassification error model* with high δ ($\hat{\beta}_1$=1.516).

Table 2.8: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 50% validation data The table employs settings from scenario 3 (High Dependence) and scenario 4 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 3 (*High* δ)										
β_0	1	1.001	(0.948 ,1.054)	0.001	0.953	(0.898 , 1.008)	0.047	0.848	(0.808 , 0.888)	0.152
β_1	1.000	1.045	(0.822 , 1.277)	0.045	1.679	(1.411 , 1.961)	0.679	2.196	(2.028 , 2.369)	1.196
SNX	0.950	0.949	(0.940 , 0.957)	0.001	0.953	(0.935 , 0.968)	0.003	-	-	-
SNY	0.950	0.949	(0.942 , 0.956)	0.001	0.953	(0.946 , 0.959)	0.003	-	-	-
SPX	0.800	0.800	(0.791 , 0.809)	0.000	0.810	(0.800 , 0.819)	0.012	-	-	-
SPY	0.800	0.800	(0.791 , 0.809)	0.000	0.805	(0.785 , 0.825)	0.007	-	-	-
D_{11}	0.048	0.046	(0.039 , 0.053)	0.028	-	-	-	-	-	-
D_{10}	-0.010	-0.010	(-0.011 , -0.008)	0.012	-	-	-	-	-	-
D_{01}	-0.010	0.001	(-0.010 , 0.021)	1.090	-	-	-	-	-	-
D_{00}	0.160	0.159	(0.153 , 0.165)	0.006	-	-	-	-	-	-
Scenario 4 (*Low* δ)										
β_0	1	1.001	(0.948 , 1.055)	0.001	1.001	(0.945 , 1.057)	0.001	1.058	(1.017 , 1.100)	0.058
β_1	1.000	1.016	(0.752 , 1.293)	0.0157	1.017	(0.754 , 1.293)	0.017	0.340	(0.244 , 0.436)	0.660
SNX	0.950	0.945	(0.924 , 0.963)	0.005	0.946	(0.925 , 0.964)	0.004	-	-	-
SNY	0.950	0.950	(0.943 , 0.956)	0.000	0.950	(0.943 , 0.956)	0.000	-	-	-
SPX	0.800	0.800	(0.790 , 0.809)	0.000	0.800	(0.791 , 0.809)	0.000	-	-	-
SPY	0.800	0.799	(0.780 , 0.818)	0.001	0.799	(0.778 , 0.819)	0.001	-	-	-
D_{11}	-0.001	0.001	(-0.003 , 0.007)	1.685	-	-	-	-	-	-
D_{10}	0.024	0.024	(0.020 , 0.028)	0.005	-	-	-	-	-	-
D_{01}	0.016	0.017	(-0.003 , 0.040)	0.090	-	-	-	-	-	-
D_{00}	0.160	0.067	(0.057 , 0.077)	0.583	-	-	-	-	-	-

Table 2.9: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 50% validation data The table employs settings from scenario 5 (High Dependence) and scenario 6 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 5 (*High* δ)										
β_0	1.000	0.999	(0.941 , 1.056)	0.248	0.937	(0.880 , 0.995)	0.172	0.492	(0.457 , 0.527)	0.508
β_1	1	1.041	(0.820 , 1.271)	0.302	1.614	(1.347 , 1.894)	1.017	1.889	(1.718 , 2.064)	0.889
SNX	0.800	0.800	(0.787 , 0.812)	0.200	0.811	(0.779 , 0.842)	0.189	-	-	-
SNY	0.800	0.800	(0.789 , 0.811)	0.200	0.806	(0.795 , 0.817)	0.194	-	-	-
SPX	0.950	0.949	(0.944 , 0.955)	0.001	0.955	(0.949 , 0.960)	0.005	-	-	-
SPY	0.950	0.949	(0.943 , 0.955)	0.001	0.951	(0.939 , 0.961)	0.001	-	-	-
D_{11}	0.048	0.158	(0.150 , 0.165)	2.325	-	-	-	-	-	-
D_{10}	-0.010	-0.010	(-0.011 , -0.009)	0.013	-	-	-	-	-	-
D_{01}	-0.010	0.001	(-0.010 , 0.021)	1.089	-	-	-	-	-	-
D_{00}	0.160	0.047	(0.042 , 0.053)	0.704	-	-	-	-	-	-
Scenario 6 (*Low* δ)										
β_0	1	1.001	(0.945 , 1.058)	0.2514	1.000	(0.943 , 1.058)	0.250	0.598	(0.562 , 0.634)	0.402
β_1	1	1.013	(0.753 , 1.286)	0.2656	1.011	(0.754 , 1.283)	0.264	0.399	(0.288 , 0.511)	0.601
SNX	0.800	0.800	(0.765 , 0.832)	0.201	0.800	(0.766 , 0.832)	0.200	-	-	-
SNY	0.800	0.800	(0.789 , 0.811)	0.200	0.800	(0.789 , 0.811)	0.200	-	-	-
SPX	0.950	0.950	(0.944 , 0.955)	0.001	0.950	(0.944 , 0.955)	0.000	-	-	-
SPY	0.950	0.949	(0.941 , 0.957)	0.001	0.948	(0.936 , 0.960)	0.002	-	-	-
D_{11}	0.041	-0.004	(-0.017 , 0.010)	1.101	-	-	-	-	-	-
D_{10}	0.012	0.014	(0.011 , 0.018)	0.223	-	-	-	-	-	-
D_{01}	0.014	0.013	(-0.004 , 0.033)	0.053	-	-	-	-	-	-
D_{00}	-0.005	0.041	(0.035 , 0.047)	8.558	-	-	-	-	-	-

Table 2.10: Average posterior summaries and relative bias (in absolute value) of the 1000 replicated data set for the Model Parmeters under the *Dependent Model*, *Independent Model* and *Naive Model* using simulated data (N=10,000) with 50% validation data The table employs settings from scenario 7 (High Dependence) and scenario 8 (Low Dependence).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
Scenario 7 ($High\ \delta$)										
β_0	1	1.002	(0.951 , 1.054)	0.055	0.966	(0.915 , 1.017)	0.017	0.864	(0.826 , 0.902)	0.136
β_1	1	1.033	(0.823 , 1.252)	0.088	1.516	(1.260 , 1.785)	0.595	1.482	(1.319 , 1.649)	0.482
SNX	0.950	0.949	(0.939 , 0.957)	0.051	0.948	(0.929 , 0.965)	0.052	-	-	-
SNY	0.950	0.949	(0.942 , 0.955)	0.051	0.952	(0.945 , 0.958)	0.048	-	-	-
SPX	0.950	0.949	(0.943 , 0.955)	0.001	0.953	(0.947 , 0.958)	0.003	-	-	-
SPY	0.950	0.949	(0.943 , 0.955)	0.001	0.949	(0.936 , 0.960)	0.001	-	-	-
D_{11}	0.048	0.046	(0.039 , 0.053)	0.025	-	-	-	-	-	-
D_{10}	-0.003	-0.002	(-0.003 , -0.001)	0.083	-	-	-	-	-	-
D_{01}	-0.003	0.007	(-0.002 , 0.025)	3.966	-	-	-	-	-	-
D_{00}	0.048	0.047	(0.042 , 0.053)	0.001	-	-	-	-	-	-
Scenario 1 ($Low\ \delta$)										
β_0	1.000	1.001	(0.952 , 1.052)	0.054	1.001	(0.949 , 1.053)	0.054	0.928	(0.889 , 0.966)	0.072
β_1	1.000	1.016	(0.772 , 1.275)	0.069	1.014	(0.776 , 1.267)	0.068	0.596	(0.473 , 0.720)	0.404
SNX	0.950	0.944	(0.923 , 0.963)	0.056	0.946	(0.925 , 0.963)	0.054	-	-	-
SNY	0.950	0.949	(0.943 , 0.955)	0.051	0.949	(0.943 , 0.956)	0.051	-	-	-
SPX	0.950	0.949	(0.943 , 0.955)	0.001	0.950	(0.944 , 0.955)	0.000	-	-	-
SPY	0.950	0.949	(0.941 , 0.957)	0.001	0.948	(0.936 , 0.959)	0.002	-	-	-
D_{11}	0.041	0.002	(-0.002 , 0.008)	0.962	-	-	-	-	-	-
D_{10}	0.012	0.015	(0.012 , 0.019)	0.317	-	-	-	-	-	-
D_{01}	0.015	0.015	(0.001 , 0.034)	0.037	-	-	-	-	-	-
D_{00}	0.000	0.041	(0.035 , 0.047)	155.758	-	-	-	-	-	-

Figure (2.2) shows the average relative bias of the estimated misclassification and regression parameters for the scenarios where 50% of validation data is employed. The relative biases for the misclassification and regression parameters are generally small compared to the relative biases obtained when 10% of validation data was used. Here, the model produced the largest bias in estimating the true values of the β_1 parameter for most scenarios is the *naive model*.

2.3.4 MCMC Diagnostics

For MCMC diagnostics, two Markov chains are constructed, each having 10,000 iterations. The initial 5,000 iterations are discarded as burn-in. To assess whether chains from the MCMC algorithm have converged to a stationary distribution, I perform a couple of tests. A visual inspection is performed by examining trace and density plots of each parameter before formal tests are conducted. For 10% validation data, the trace plots and density plots for the first Scenario are shown in Figures (2.3) - (2.8). Also for 50% validation data the trace and density plots for the second Scenario are shown, please refer to Figures (2.12) - (2.17). These few plots are shown because similar plots are obtained for the remaining scenarios considered. The trace plots show no apparent patterns, indicating that the sampler mixed well and stationarity is achieved. The formal test considered is the Gelman-Rubin diagnostics. The Gelman-Rubin R statistics are approximately equal to 1 for all parameters, which is a confirmation of stationarity. Autocorrelation plots are used to check convergence and mixing performance. Autocorrelation plot are shown in Figures (2.9) - (2.11) and Figures (2.18) - (2.20). The quick decay shows a good mix and an evidence of convergence.

Figure 2.2: Graph of the average relative bias of the misclassification parameters and the regression coefficient with 50% validation.

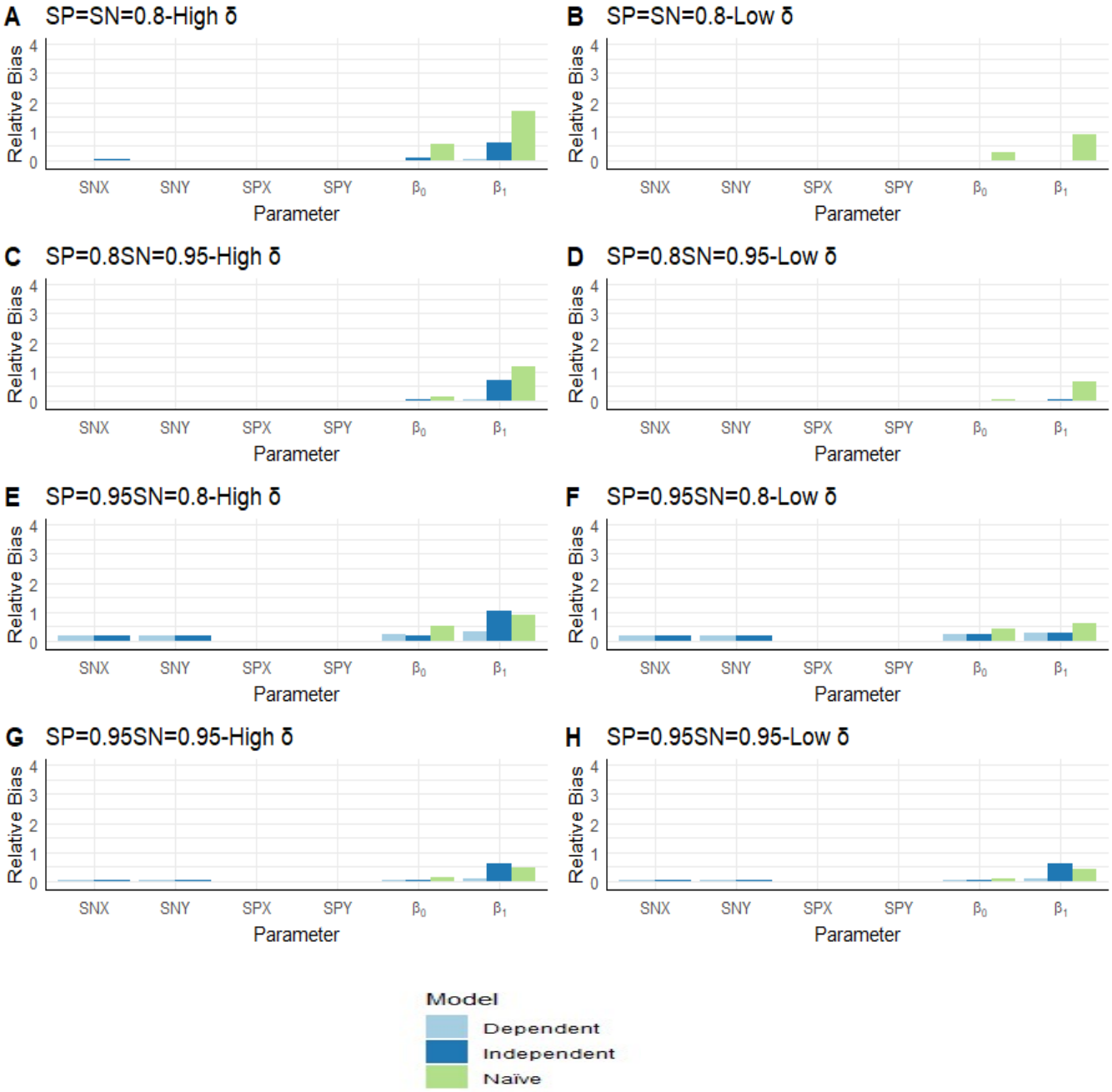


Figure 2.3: Trace plots for the posterior samples under the *dependent misclassification error odel* with 10% validation data for Scenario 1.($SNX = 0.8$,$SNY = 0.8$,$SPX = 0.8$,$SPY = 0.8$, High Dependence)

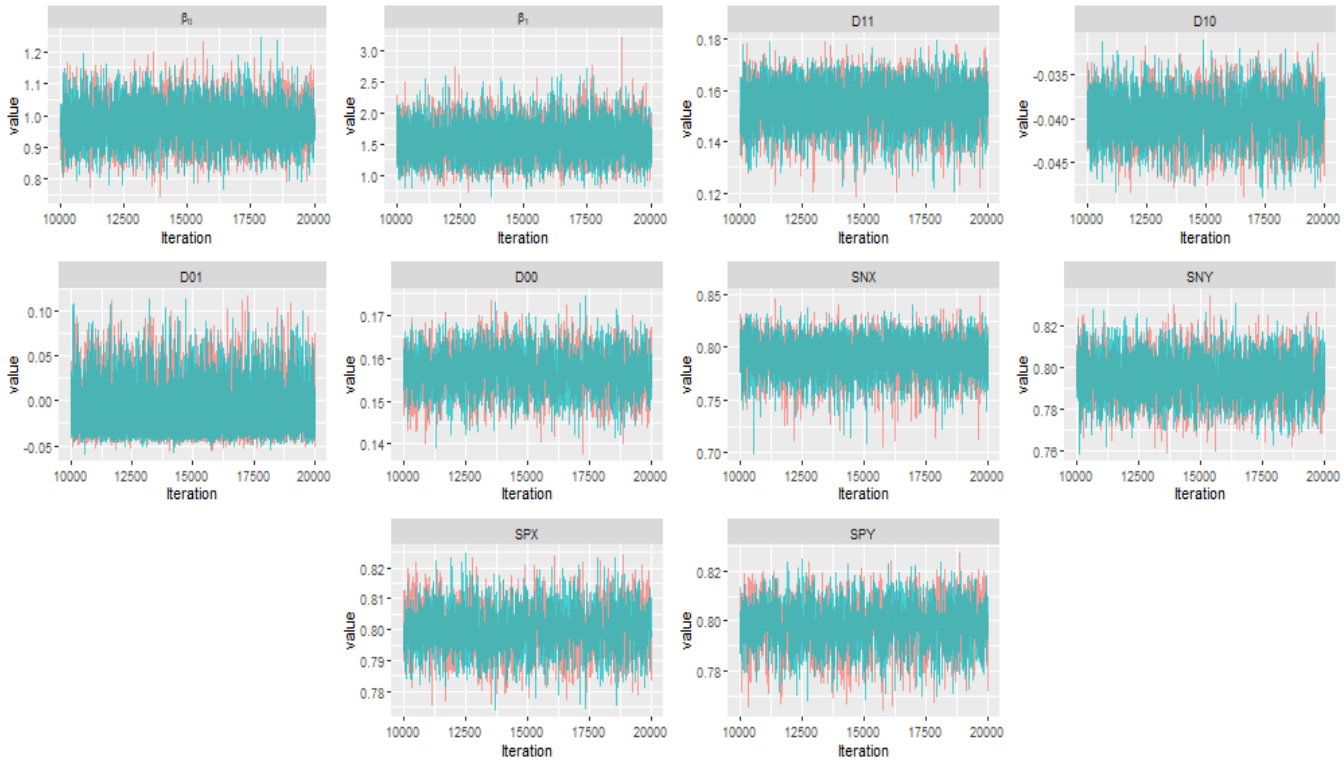


Figure 2.4: Trace plots for the posterior samples under the *independent misclassification error model* with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High Dependence)

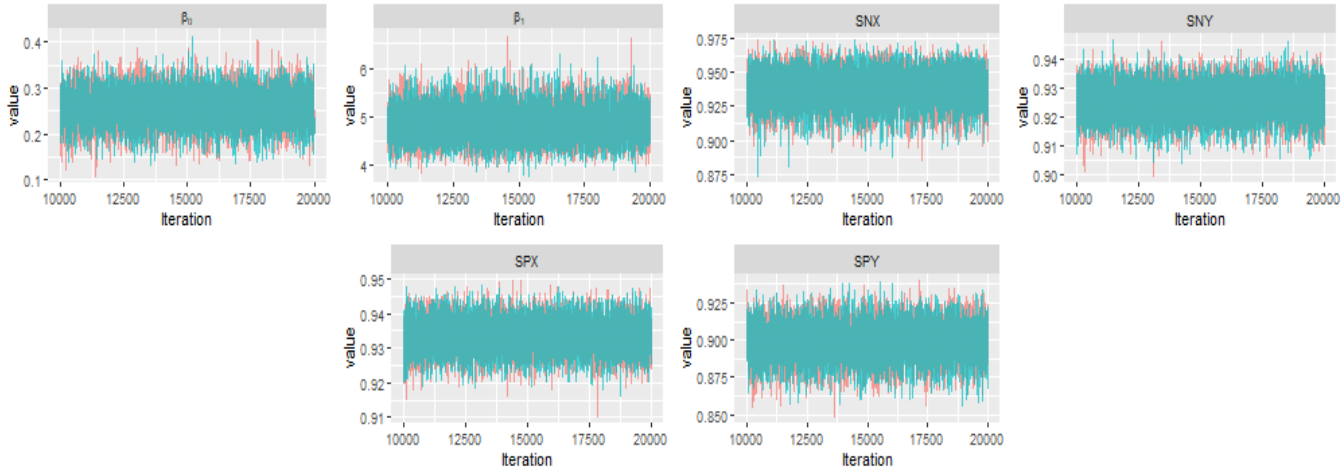


Figure 2.5: Trace plots for the posterior samples under the *Naive* Model with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High Dependence)

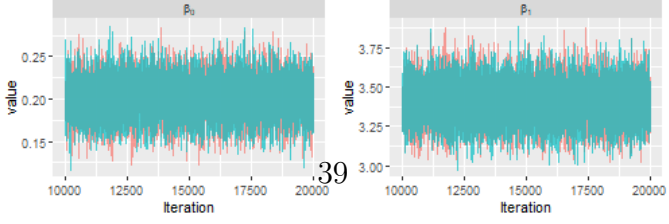


Figure 2.6: Density plots for the posterior samples under the *dependent misclassification error model* with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High δ)

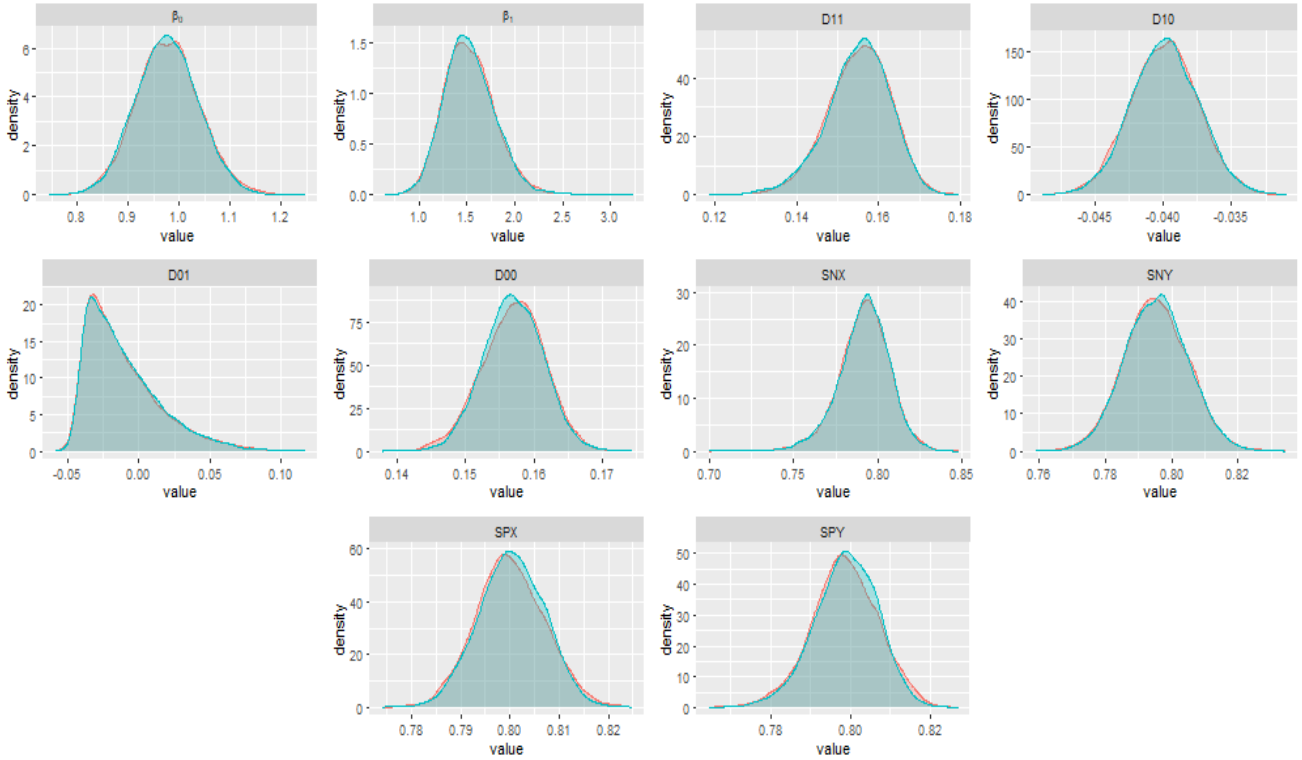


Figure 2.7: Density plots for the posterior samples under the *independent misclassification error model* with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High δ)

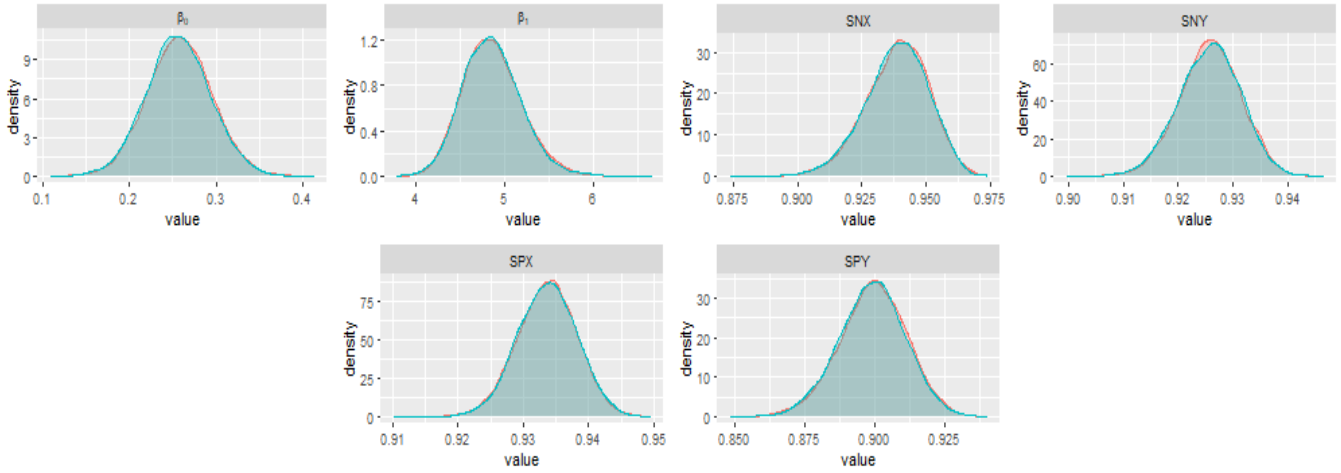


Figure 2.8: Density plots for the posterior samples under the *Naive* Model with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High δ)

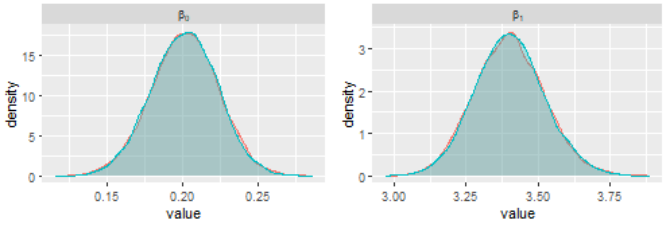


Figure 2.9: Autocorrelation plots for the posterior samples under the *Dependent misclassification error model* with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High δ)

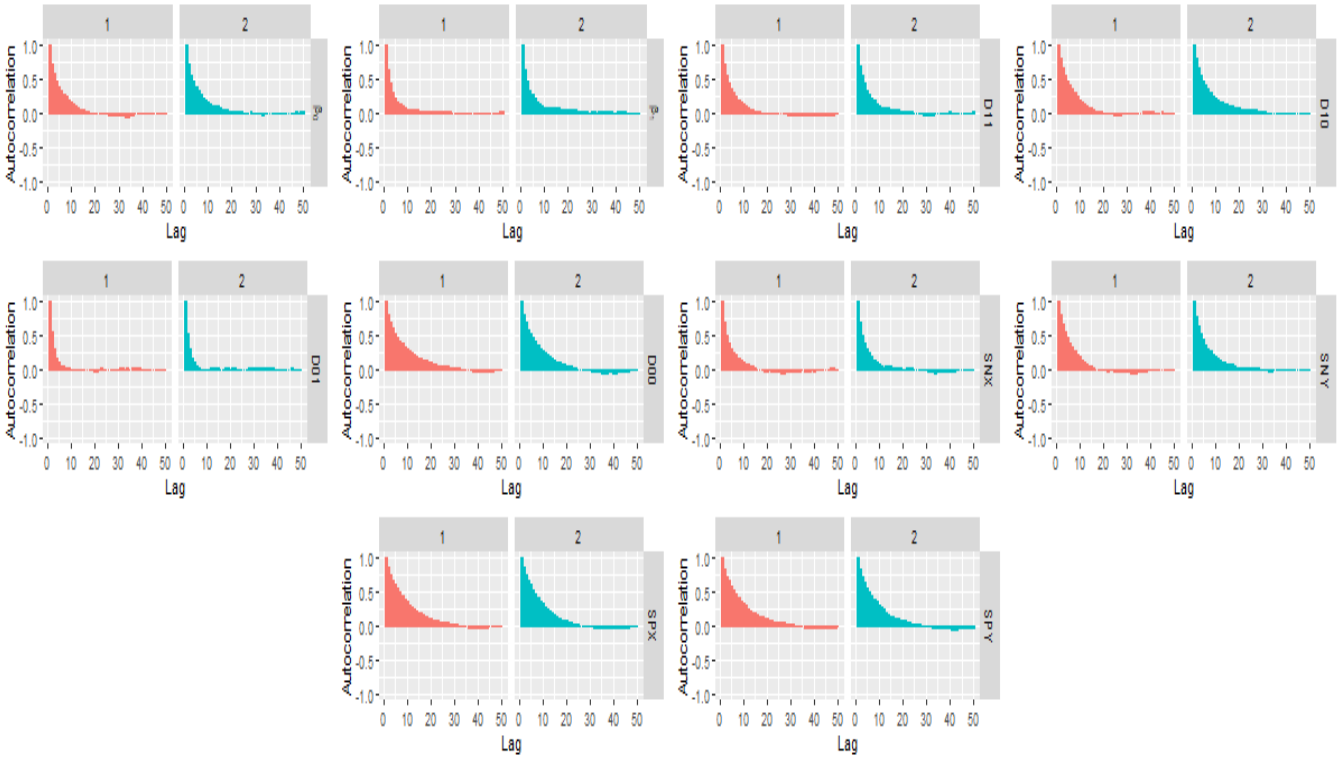


Figure 2.10: Autocorrelation plots for the posterior samples under the *Independent misclassification error model* with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High δ)

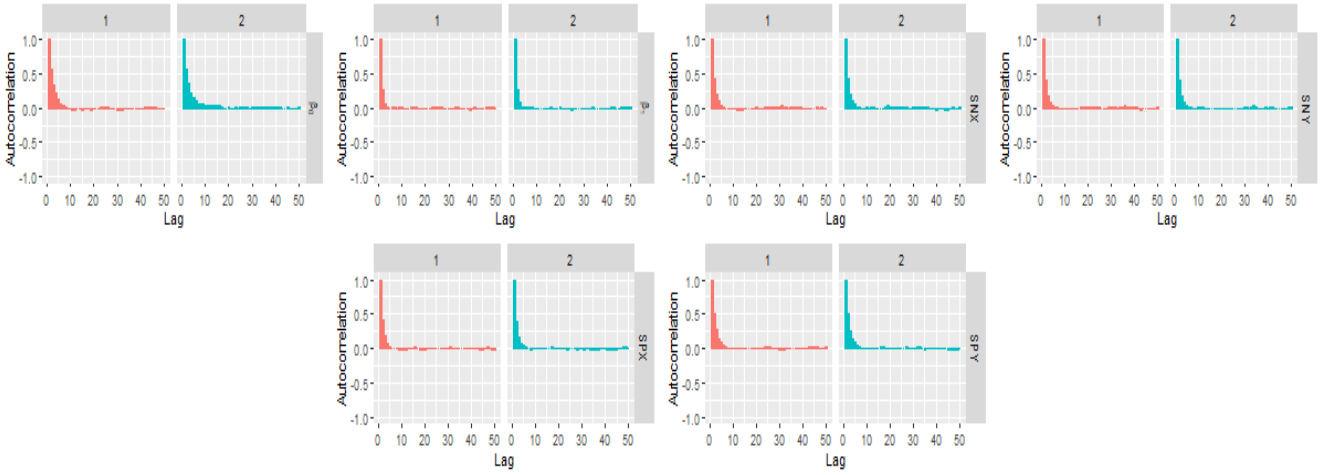


Figure 2.11: Autocorrelation plots for the posterior samples under the *Naive* Model with 10% validation data for Scenario 1.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, High δ)

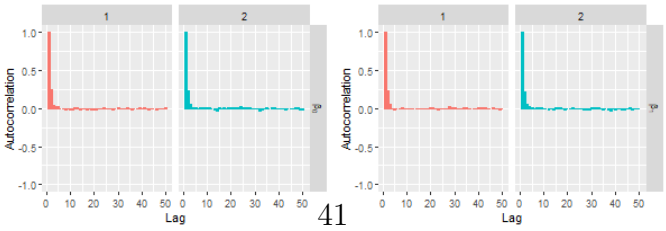


Figure 2.12: Trace plots for the posterior samples under the *Dependent misclassification model* with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

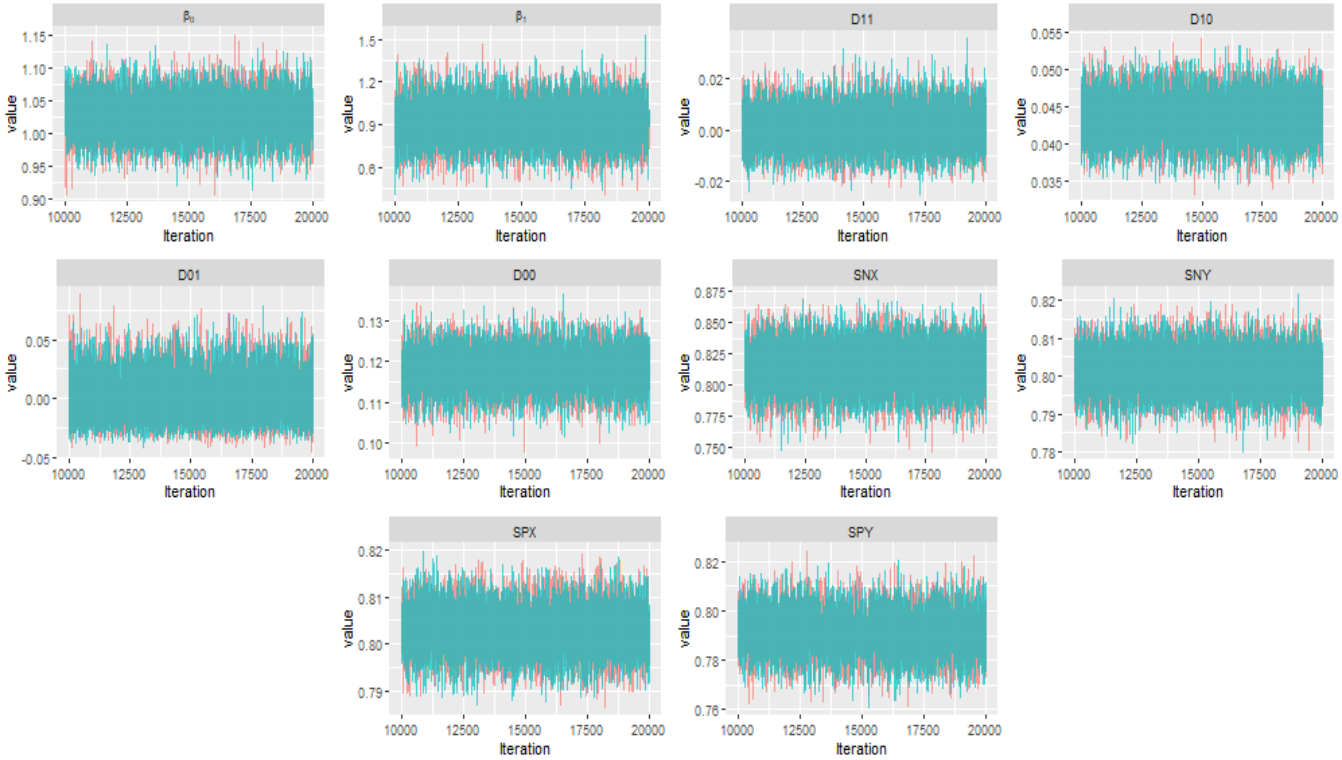


Figure 2.13: Trace plots for the posterior samples under the *Independent misclassification model* with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

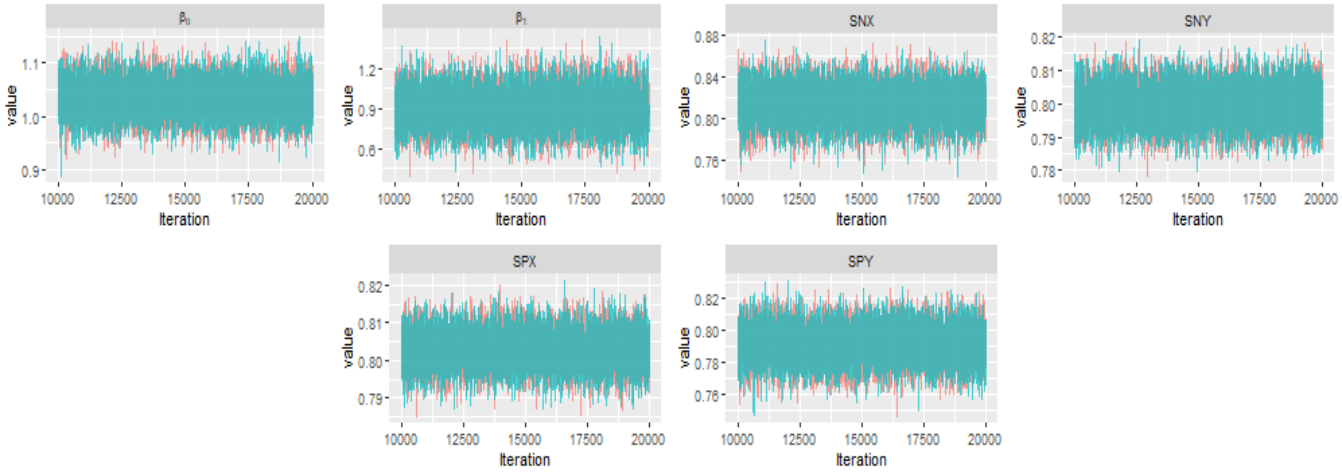


Figure 2.14: Trace plots for the posterior samples under the *Naive* Model with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

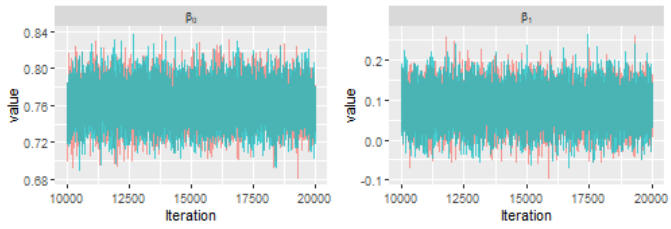


Figure 2.15: Density plots for the posterior samples under the *Dependent misclassification error model* with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

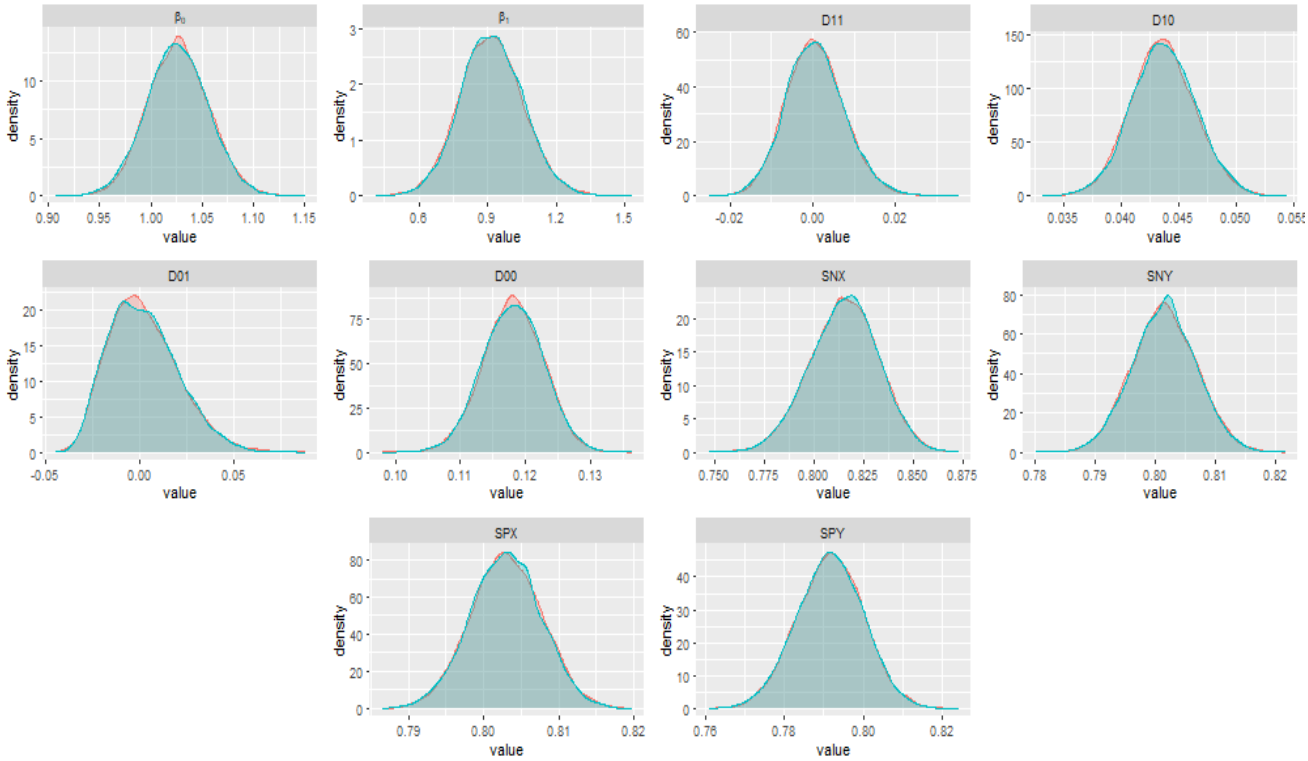


Figure 2.16: Density plots for the posterior samples under the *Independent misclassification model* with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

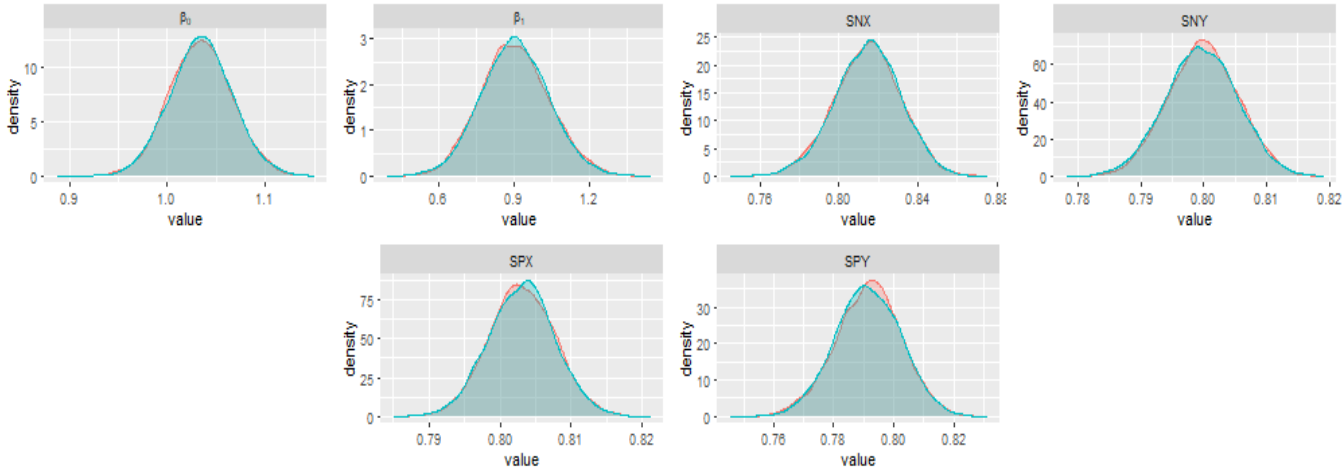


Figure 2.17: Density plots for the posterior samples under the *Naive* Model with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

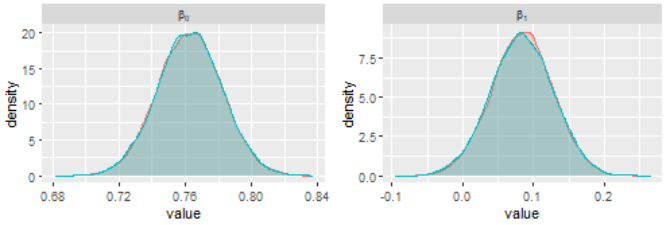


Figure 2.18: Autocorrelation plots for the posterior samples under the *Dependent misclassification error model* with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

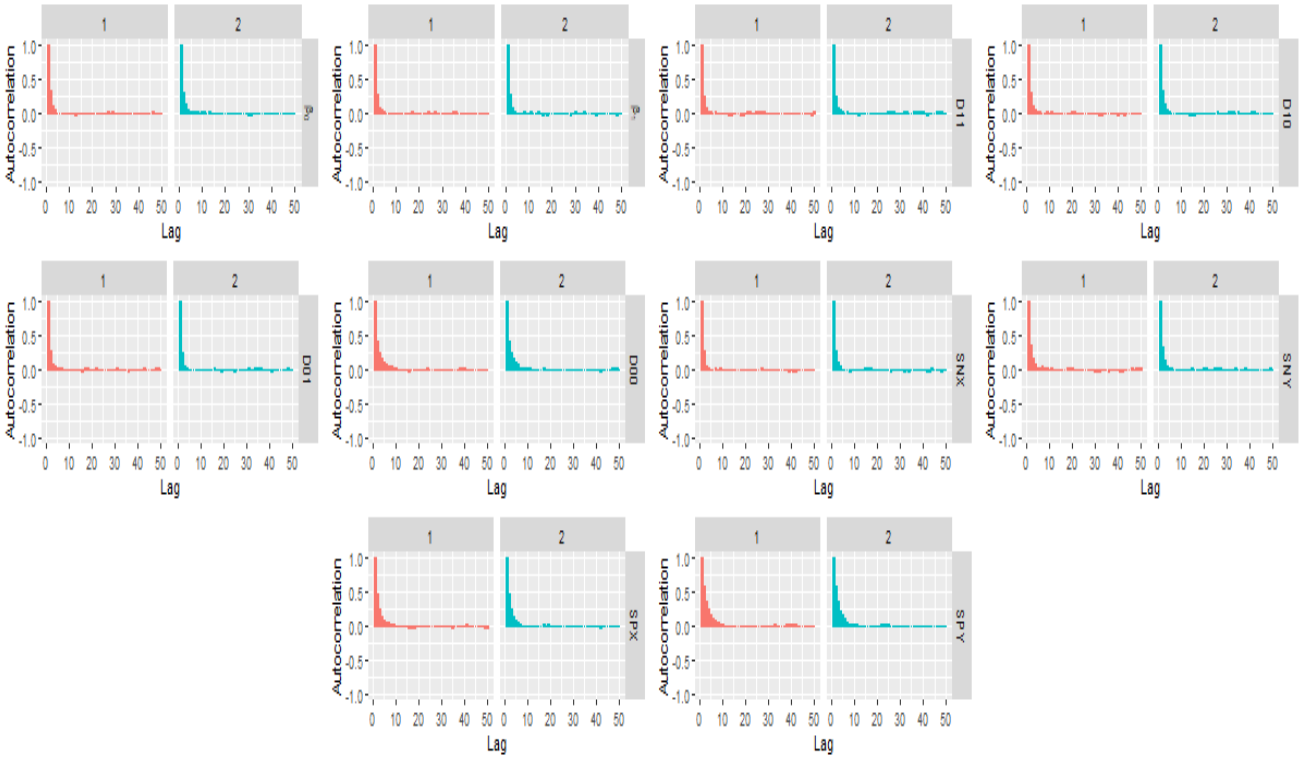


Figure 2.19: Autocorrelation plots for the posterior samples under the *Independent misclassification error model* with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)

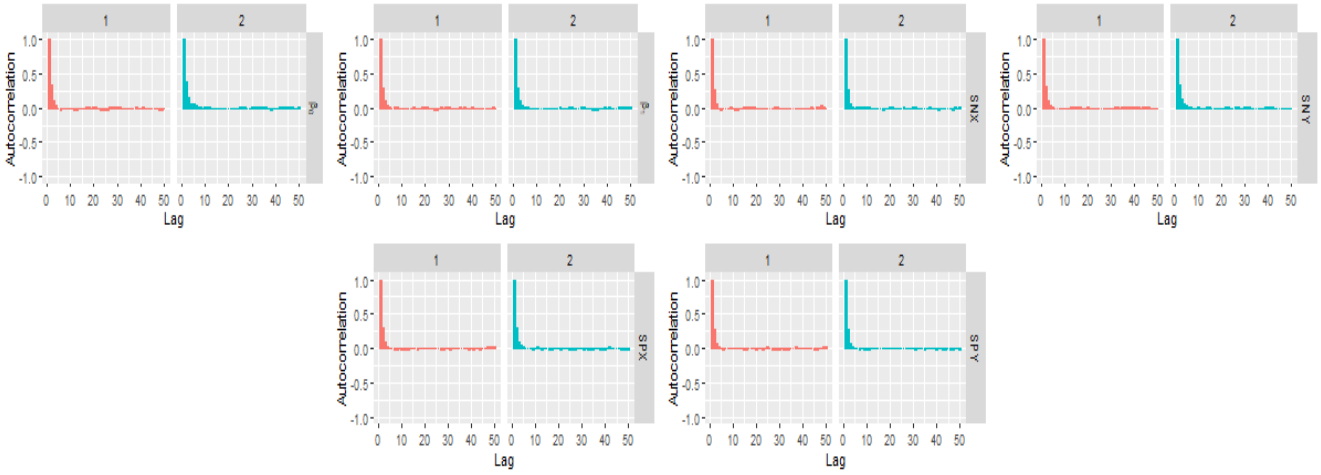
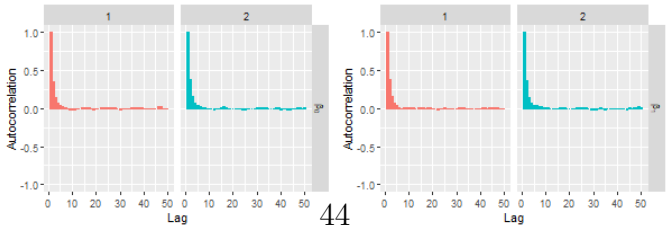


Figure 2.20: Autocorrelation plots for the posterior samples under the *Naive* Model with 50% validation data for Scenario 2.($SNX = 0.8, SNY = 0.8, SPX = 0.8, SPY = 0.8$, Low δ)



2.4 Discussion

In this chapter, a model that accounts for dependent misclassification error in a binary response variable and binary covariate was introduced. I considered a data structure with validation data available. Simulation study is conducted with the objective of checking the consequences of fitting an *independent misclassification error model* and a *naive model* to a data generated from a *dependent misclassification error model*.

Findings based on the simulation studies summarize that, when 10% of validation data is employed, the *dependent misclassification error model* demonstrates that it was best at producing estimates that are closer to the true value than the *naive model* and *independent misclassification error model* for all scenarios considered. Making comparison based on the β_1 estimates, the model that produced the largest bias for high δ scenarios is the *independent misclassification error model*. However, for low δ, both the *dependent* and *independent misclassification error model* accurately estimated the true values of the parameters under consideration, except Scenario 8, where the naive model better estimated the β_1 than the *independent misclassification error model*. Although the *independent misclassification error model* was the model that produced tjhe largest bias for cases where 10% validation data was used, it performed better in most scenarios than the naive model in the case where 50% validation data was employed. Overall, it is observed that for low misclassification scenarios, the relative bias for the β_1 parameter is low.

The simulation studies revealed that ignoring dependence error can be worse than ignoring misclassification altogether when validation data is small (10% of validation data). However, for substantial validation (50% of validation data), when the δ is low, both the *dependent* and *independent misclassification error models* accurately estimate the true values when δ is low. In the next chapter (Chapter 3), I extend the proposed model in this chapter to a multi-category response variable and a multi-category covariate.

Chapter 3

Analysis of Categorical Data Subject to Misclassification

3.1 Introduction

This chapter describes a general approach to joint misclassification in category data. It is an extension and a generalization of the binary response variable and binary covariate case proposed in chapter 2. Category data are a common occurence in many fields of study, especially epidemiology. An example includes the association between socio-economic status (SES) and malnourished children [52]. The covariate SES is a three level category variable (low, middle and high). Wealth index (asset index) is widely used as a measure of SES [46]. Wealth index is created by measuring an individual's assets. However, not possessing certain assets may not necessarily mean one cannot afford them. Employing wealth index can introduce misclassification error in categorizing SES. A child's nutrition status is categorized as severely undernourished, moderately undernourished, and nourished. Nutrition status is categorized by comparing a child's weight and height with reference standards [55]. However, these athropometric indicators may be influenced by chronic diseases or genetics but not lack of nutrient availability. This can therefore lead to misclassification in the categorization of nutrient status. This kind of data is subject to joint misclassification errors, but these errors are often ignored in the analysis since techniques for adjustment have been least explored. The chapter is organized as follows: notations and prelimary concepts for category misclassification error in both the response variable and covariate are introduced in Section 3.2. A Bayesian method for the estimation of the model parameters is discussed in Section 3.3. In Section 3.4, a comprehensive simulation study is conducted to assess the consequences

of ignoring dependence of misclassification errors or completely ignoring misclassification in a trinary response variable and a trinary covariate.

3.2 Model and Notation

Let Y and X denote the actual response variable and covariate of interest and let Y^* and X^* denote their surrogate (i.e., error-prone) variables respectively. The number of categories of the response variable and covariate are represented by g_Y and g_X, respectively.

Now, the dependence parameters D_{ijkl} is defined by,

$$D_{ijkl} = P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)P(X^* = j|X = l). \tag{3.1}$$

Similar reasoning as in the binary case, each dependence parameter is bounded to lie within an interval.

$$\begin{aligned} Max\big\{ - M_Y[i,k]M_X[j,l]; -(1 - M_Y[i,k])(1 - M_X[j,l])\big\} &\leq \boldsymbol{D_{ijkl}} \\ &\leq Min\big\{(1 - M_Y[i,k])M_X[j,l]; M_Y[i,k](1 - M_X[j,])\big\} \end{aligned} \tag{3.2}$$

Please refer to Appendix (A) for the proof of the boundaries of the dependence parameters. Let the vector of the joint probabilities of the error-prone response variable and the error-prone covariate p^*_{ij} be represented by p^*, then $p^*_i = (p^*_{i1}, ..., p^*_{ig_Y})', (i = 1, ..., g_X)$ and $p^* = ({p^*_1}', ..., {p^*_{g_X}}')'$ and the vector of the joint probabilities of the error-free response variable and the error-free covariate p_{ij} be represented by p, then $p_i = (p_{i1}, ..., p_{ig_X})', (i = 1, ..., g_Y)$ and $p = ({p_1}', ..., {p_{g_Y}}')'$.

The relationship between p^* and the error-free joint probability p is given by:

$$\boldsymbol{p^*} = (\boldsymbol{M_Y} \otimes \boldsymbol{M_X} + \boldsymbol{D})\boldsymbol{p} \tag{3.3}$$

where $\boldsymbol{p^*}$, $\boldsymbol{p}$ are defined accordingly above,

$$\boldsymbol{M_Y}[i,k] = P(Y^* = i|Y = k), \tag{3.4}$$

$$\boldsymbol{M_X}[j,l] = P(X^* = j|X = l), \tag{3.5}$$

and $\boldsymbol{D}$ are the dependence parameters defined in Eq.(3.1). Note that, $\boldsymbol{M_Y}[i,k]$ and $\boldsymbol{M_X}[j,l]$ represent the misclassification in the response variable and covariate respectively. In the

binary case the misclassification parameters are the sensitivities and specificities, defined as,

$$SN_Y = P(Y^* = 1|Y = 1), \qquad SN_X = P(X^* = 1|X = 1),$$
$$SP_Y = P(Y^* = 0|Y = 0), \qquad SP_X = P(X^* = 0|X = 0).$$

3.2.1 Misclassification in a trinary response variable and a trinary covariate

Let Y and X represent a three-category response variable and a three-category covariate respectively. Suppose both Y and X are misclassified, Y^* and X^* will instead be recorded. Then i, j, k, l in $\boldsymbol{M_Y}[i, k]$, $\boldsymbol{M_X}[j, l]$, and $\boldsymbol{D_{ijkl}}$ from Eq.(3.1), Eq.(3.4), and Eq.(3.5) takes on $1, 2, 3$.

Relating to the fact that ,

$$\sum_i D_{ijkl} = 0,$$
$$\sum_j D_{ijkl} = 0.$$

The number of independent dependence parameter is obtained from $(I-1)\times(J-1)\times I\times J$. Specifically for the three category response variable and three category covariate the number of dependence paranmeters obtained are $(3-1)\times(3-1)\times 3\times 3 = 36$. As the number of category increases the number of dependence parameters in the model increase. Large number of parameters in a model may cause computational problems.

3.3 Bayesian method for adjustment for misclassification error in a category data

Bayesian analysis using a JAGS (version) program similar to the previous chapter is performed. The likelihood functions for both the validation and main data are in like manner as the binary case stated in equations (2.20) and (2.23). The joint probabilities p_{ij} are reparameterized through a multinomial logistic regression. In this study, ordinal variables are considered in the multinomial logistic regression, which is common in epidemiologic studies.

p_{ij} is generated from an ordinal logistic regression model with a three-category response variable and a three-category covariate. The following indicator variables for X are defined. Let $X = 1$ be the reference category. For $j = 2, 3$, let $Z_j = 1$ if $X = j$ and zero otherwise. The underlying ordinal logistic regression model is,

$$logit\big[P(Y \leq i|X = j)\big] = \alpha_i + \beta_1 Z_1 + \beta_2 Z_2, \tag{3.6}$$

where $i = 1, 2$.

The regression model above results to :

$$\big[P(Y \leq 2|X = j)\big] = \frac{exp\big[\alpha_1 + \beta_1 Z_1 + \beta_2 Z_2\big]}{1 + exp\big[\alpha_1 + \beta_1 Z_1 + \beta_2 Z_2\big]}, \tag{3.7}$$

$$\big[P(Y \leq 3|X = j)\big] = \frac{exp\big[\alpha_2 + \beta_1 Z_1 + \beta_2 Z_2\big]}{1 + exp\big[\alpha_2 + \beta_1 Z_1 + \beta_2 Z_2\big]}. \tag{3.8}$$

Note that $\forall\ p_{ij} \in [0, 1]$, $\alpha_1 < \alpha_2$.

The conditional probabilities for the first Y category is given by,

$$P(Y = 1|X = j) = P(Y \leq 1|X = j).$$

Subsequent conditional probabilities for Y are obtained from,

$$P(Y = i|X = j) = P(Y \leq i|X = j) - P(Y \leq i - 1|X = j).$$

$$P(Y = i, X = j) = P(Y = i|X = j)P(X = j),$$

where $P(X = j)$ are the marginal probabilities. Please refer to Appendix (B) for specific joint probabilities.

3.4 Simulation Studies

In section simulation studies are conducted to investigate the impact of fitting an independent misclassification error model and a naive model to data generated from a dependent misclassification error model when the response variable and covariate are trinary. The following models are fitted:

1. The model for **dependent misclassification errors**: In this model, the misclassification errors in the response variable depends on the misclassification errors in the covariate and vice versa.

2. The model for **independent misclassification errors** in the response variable and the covariate.

3. The **naive** model, which assumes no misclassification error.

The dependence parameter and the misclassification parameters are varied to observe if the magnitude of dependence or the extent of misclassification has an impact on the models. For the dependence parameters, the function $\phi_c(Y^*, X^*|Y, X)$ is used to control the dependence in the model. Please refer to section 3.4.1 for details of the $\phi_c(Y^*, X^*|Y, X)$ function

Also the proportion of validation data $\frac{n_v}{N}$ are varied by considering 10% proportion of validation data and 50% proportion of validation data. Here the following assumptions are made for the simulation studies, these are:

- all correct classifications are equal:

$$\begin{aligned} M_Y[1,1] &= M_Y[2,2] = M_Y[3,3] = C_Y \\ M_X[1,1] &= M_X[2,2] = M_X[3,3] = C_X \end{aligned}$$

- misclassification occurs only in classification parameters adjacent to the correct classification. Follow the discussion in Swartz et al [48], who purported that misclassification is less likely as the incorrect category moves away from the true category.

Since $\sum_{i=1}^{3} M_Y[i,k] = 1$ and $\sum_{j=1}^{3} M_X[j,l] = 1$ where, k and l takes on $1, 2, 3$, the following are obtained

$$M_Y[2,1] = M_Y[2,3] = 1 - C_Y \qquad M_X[2,1] = M_X[2,3] = 1 - C_X$$
$$M_Y[1,2] = M_Y[3,2] = \frac{1 - C_Y}{2} \qquad M_X[1,2] = M_X[3,2] = \frac{1 - C_X}{2}$$

From the above assumptions, the following results:

$$\boldsymbol{M_Y} = \begin{bmatrix} C_Y & \frac{1-C_Y}{2} & 0 \\ 1-C_Y & C_Y & 1-C_Y \\ 0 & \frac{1-C_Y}{2} & C_Y \end{bmatrix}, \quad \boldsymbol{M_X} = \begin{bmatrix} C_X & \frac{1-C_X}{2} & 0 \\ 1-C_X & C_X & 1-C_X \\ 0 & \frac{1-C_X}{2} & C_X \end{bmatrix}.$$

The kronecker product of the M_Y and M_X gives the matrix $M_Y \otimes M_X$.

Recall,

$$D_{ijkl} = P(Y^* = i, X^* = j | Y = k, X = l) - P(Y^* = i | Y = k) P(X^* = j | Y = l)$$

$$P(Y^* = i, X^* = j | Y = k, X = l) \leq P(Y^* = i | Y = k, X = l) = P(Y^* = i | Y = k)$$

this implies,

$$\begin{aligned} D_{ijkl} &\leq P(Y^* = i | Y = k) - P(Y^* = i | Y = k) P(X^* = j | X = l) \\ D_{ijkl} &= P(Y^* | Y = k) - P(Y^* = i | Y = k) P(X^* = j | X = l) \end{aligned}$$

when $P(Y^* = i | Y = k)$=0

$$D_{ijkl} = 0$$

From the above, all entries in the matrix $M_Y \otimes M_X$ that have 0 entries have corresponding entries in D to be 0. This therefore, further reduces to 16 dependence parameters.

$$\boldsymbol{M_Y \otimes M_X} = \begin{bmatrix}
C_Y C_X & C_Y \frac{(1-C_X)}{2} & 0 & \frac{(1-C_Y)}{2} C_X & \frac{(1-C_Y)}{2}\frac{(1-C_X)}{2} & 0 & 0 & 0 & 0 \\
C_Y(1-C_X) & C_Y C_X & C_Y(1-C_X) & \frac{(1-C_Y)}{2}(1-C_X) & \frac{(1-C_Y)}{2} C_X & \frac{(1-C_Y)}{2}(1-C_X) & 0 & 0 & 0 \\
0 & C_Y \frac{(1-C_X)}{2} & C_Y C_X & 0 & \frac{(1-C_Y)}{2}\frac{(1-C_X)}{2} & \frac{(1-C_Y)}{2} C_X & 0 & 0 & 0 \\
(1-C_Y)C_X & (1-C_Y)\frac{(1-C_X)}{2} & 0 & C_Y C_X & C_Y \frac{(1-C_X)}{2} & 0 & (1-C_Y)C_X & (1-C_Y)\frac{(1-C_X)}{2} & 0 \\
(1-C_Y)(1-C_X) & (1-C_Y)C_X & (1-C_Y)(1-C_X) & C_Y(1-C_X) & C_Y C_X & C_Y(1-C_X) & (1-C_Y)(1-C_X) & (1-C_Y)C_X & (1-C_Y)(1-C_X) \\
0 & (1-C_Y)\frac{(1-C_X)}{2} & (1-C_Y)C_X & 0 & C_Y \frac{(1-C_X)}{2} & C_Y C_X & 0 & (1-C_Y)\frac{(1-C_X)}{2} & (1-C_Y)C_X \\
0 & 0 & 0 & \frac{(1-C_Y)}{2} C_X & \frac{(1-C_Y)}{2}\frac{(1-C_X)}{2} & 0 & C_Y C_X & C_Y \frac{(1-C_X)}{2} & 0 \\
0 & 0 & 0 & \frac{(1-C_Y)}{2}(1-C_X) & \frac{(1-C_Y)}{2} C_X & \frac{(1-C_Y)}{2}(1-C_X) & C_Y(1-C_X) & C_Y C_X & C_Y(1-C_X) \\
0 & 0 & 0 & 0 & \frac{(1-C_Y)}{2}\frac{(1-C_X)}{2} & \frac{(1-C_Y)}{2} C_X & 0 & C_Y \frac{1-C_X}{2} & C_Y C_X
\end{bmatrix}.$$

Prior Construction

1. Priors for the misclassification parameters C_Y and C_X: Truncated beta distributions are assigned to the priors of the misclassification parameters. An equal-tail 95% CI (0.6,0.95) are used to obtain the priors for the misclassification parameters. The distribution is truncated to lie within [0.5,1], that is :

$$\begin{aligned} C_Y &\sim Beta(14.19, 3.38)I(C_Y > 0.5); \\ C_X &\sim Beta(14.19, 3.38)I(C_X > 0.5). \end{aligned}$$

   where, $I(C_Y > 0.5)$ and $I(C_X > 0.5)$ are indicator functions with value equal to 1 if the input is greater than 0.5 and 0 otherwise.

2. Priors for the regression parameters α_i, β_i, where $i = 1, 2$: The priors for the multi-category regression parameters are weakly informative priors, hence normal distribution with a large variance are assumed. However, for an ordinal regression model, α_2 should necessarily be greater than α_1, that is $\alpha_2 > \alpha_1$, hence α_1 is an upper truncated normal distribution with an upper bound α_2.

$$\begin{aligned} \beta_i &\sim N(0, 1000); \\ \alpha_2 &\sim N(0, 1000); \\ \alpha_1 &\sim N(0, 1000)I(\alpha_1 < \alpha_2). \end{aligned} \tag{3.9}$$

3. Priors for the dependence parameters D_{ijkl}. For the dependence parameters, unifom distributions constrained within the boundaries of Eq. (3.2) are chosen.

Simulation Setup

Based on the ordinal logistic regression model considered in Section (3.3), I derive the joint distribution p_{ij} for (Y, X). The values of $\alpha_1 = -1$, $\alpha_2 = 0.5$, $\beta_1 = 0.5$ and $\beta_2 = 1$ are set. The marginal probabilities are also set at $P(X = 1) = 0.3$ and $P(X = 2) = 0.3$. To introduce non-differential and dependent misclassification errors in Y and X, true value of M_Y,M_X,

and D are set up. In trinary misclassification with validation, the 81 distinct patterns of the validation data are derived from $P(Y^* = i, X^* = j, Y = k, X = l)$ where $i, j, k, l = 1, 2, 3$.

In the simulation studies, two different proportions of the validation data are considered, and each has a sample size 100,000, including both main data and validation data. (a) 10% validation data: 90,000 are main data observations, that is, there are observations for only $Y^*$and X^*, and 10,000 are validation data observations, that is observations for all $Y^*, X^*, Y$and X. (b) 50% of the sampling unit as validation data; the main data and the validation data have 50,000 observations. There are eight scenarios for each proportion of the validation data, and I repeatedly generated 1000 data sets for each scenario for the simulation study. The average of each of the parameter's estimates are calculated. Table (3.1) presents the dependence value for each of the dependence parameters for various misclassification scenarios.

3.4.1 Choice of Dependence

The extent of dependence between the response variable and covariate in the categorical setting is characterized by $E(\phi_c(Y^*, X^*|Y, X))$ which is derived from the concept of Cramer's V (Liu et al. 2020) [29] . In the trinary categorization, ϕ_c is defined by:

$$\phi_c(Y^*, X^*|Y = k, X = l) = \sqrt{\frac{\chi^2}{2n}}, \tag{3.10}$$

where,

$$\begin{aligned} \chi^2 &= n\sum_{ij} \frac{P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)P(X^* = j|X = l)}{P(Y^* = i|Y = k)P(X^* = j|X = l)}, \\ &= n\sum_{ij} \frac{(D_{ijkl})^2}{M_Y[i,k]\ M_X[j,l]}. \end{aligned}$$

Hence,

$$\phi_c(Y^*, X^*|Y = k, X = l) = \sqrt{\frac{1}{2}\sum_{ij} \frac{(D_{ijkl})^2}{M_Y[i,k]\ M_X[j,l]}}, \tag{3.11}$$

Please note that, $\phi_c(Y^*, X^*|Y = k, X = l)$ reduces to $\delta_r = E(Y^*, X^*|Y = k, X = l)$ in the binary case. The function $\phi_c(Y^*, X^*|Y = k, X = l)$ is optimized to obtain the low and high

value for each dependence parameter. The boundaries for the dependence parameters are used as constraints in the optimization process, this included all non-linear constraints. An indicator function is included in the MCMC smpling in jags by using the step() function. This function is used to impose all the non-linear constraints in the models. Only MCMC samples which satisfy the nonlinear constraint were included in the posterior summaries.

Table 3.1: Dependence value for the dependence parameter for various scenarios.

	$(C_Y = 0.8, C_X = 0.8)$		$(C_Y = 0.8, C_X = 0.95)$		$(C_Y = 0.95, C_X = 0.8)$		$(C_Y = 0.95, C_X = 0.95)$	
D parameters	Low $\phi_c = 0.0667$ (Scenario 1)	High $\phi_c = 0.4733$ (Scenario 2)	Low $\phi_c = 0.0161$ (Scenario 3)	High $\phi_c = 0.2523$ (Scenario 4)	Low $\phi_c = 0.1115$ (Scenario 5)	High $\phi_c = 0.2569$ (Scenario 6)	Low $\phi_c = 0.0176$ (Scenario 7)	High $\phi_c = 0.6041$ (Scenario 8)
D1	-4.788E-05	0.0600	1.562E-04	0.0150	-4.043E-05	0.0150	-3.654E-05	0.0225
D2	-1.292E-04	0.0600	-5.205E-04	0.0250	-4.195E-05	0.0150	2.505E-04	0.0225
D3	-1.199E-04	-0.0300	-6.575E-05	-0.0100	-3.699E-05	-0.0075	-8.072E-05	-0.0112
D4	4.421E-05	-0.0600	-5.065E-05	-0.0150	1.103E-04	-0.0150	-7.635E-06	-0.0225
D5	-4.502E-04	0.0600	-3.859E-04	0.0150	-5.184E-05	0.0233	3.102E-04	0.0225
D6	-4.713E-05	-0.0300	-6.687E-05	-0.0075	-1.157E-04	-0.0117	-7.746E-05	-0.0113
D7	-3.297E-05	0.0178	-5.094E-05	0.0058	-2.500E-03	0.0058	-2.827E-05	0.0063
D8	8.944E-03	-0.0356	4.748E-04	-0.0117	5.000E-03	-0.0117	3.449E-04	-0.0126
D9	-3.287E-05	-0.0356	-1.147E-04	-0.0117	-2.000E-02	-0.0117	-1.603E-04	-0.0126
D10	7.106E-02	0.0711	1.953E-02	0.0233	4.000E-02	0.0233	2.341E-02	0.0253
D11	3.520E-05	-0.0600	7.216E-05	-0.0150	-4.542E-05	-0.0233	-1.073E-04	-0.0225
D12	-1.575E-05	0.0300	-4.759E-05	0.0075	-3.115E-05	0.0117	-3.102E-05	0.0113
D13	-5.311E-05	0.0600	-1.962E-05	0.0150	-6.244E-05	0.0150	-4.365E-05	0.0225
D14	-4.935E-05	-0.0600	-5.080E-05	-0.0233	-2.179E-04	-0.0150	6.334E-06	-0.0225
D15	-2.047E-05	0.0300	-4.386E-05	0.0117	-3.747E-05	0.0075	-3.033E-05	0.0113
D16	-5.547E-05	0.0600	-1.956E-05	0.0150	-2.216E-05	0.0150	-3.871E-05	0.0225

3.4.2 Simulation Results

In the simulation study results, the outcome of fitting an independent misclassification error model and a naive model to data generated from a dependent misclassification model are shown. The results are presented for (1) 10% proportion of validation data, and (2) 50% proportion of validation data.

Results for simulation studies employing 10% validation data

Tables (3.2) -(3.9) shows the average posterior means and the 95% credible intervals for the regression and misclassification parameters in the three models for each of the 8 simulation scenarios when proportion of validation data is 10%. The average relative bias for each parameter estimate of the 1000 datasets under the various models is also presented in the tables. Since a three-category response variable and a three-category covariate are considered, the primary parameters of interest, which give the relationships between the two variables, are the β_1 and β_2 parameters. Also note that for the misclassification parameters, the higher the value, the less the misclassification error.

Table 3.2: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 1 (High ϕ_c) and scenario 2 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9995	(-1.0381 , -0.9609)	0.0005	-1.0295	(-1.0616 , -0.9973)	0.0295	-1.0745	(-1.0969 , -1.0520)	0.0745
α_2	0.5000	0.5008	(0.4648 , 0.5368)	0.0016	0.4838	(0.4528 , 0.5149)	0.0323	0.6783	(0.6565 , 0.7001)	0.3565
β_1	0.5000	0.4989	(0.4352 , 0.5628)	0.0022	0.4833	(0.4334 , 0.5333)	0.0333	0.4041	(0.3767 , 0.4316)	0.1917
β_2	1.0000	0.9990	(0.9501 , 1.0481)	0.0010	1.0743	(1.0355 , 1.1132)	0.0743	0.8518	(0.8235 , 0.8800)	0.1482
C_X	0.8000	0.8001	(0.7931 , 0.8071)	0.0001	0.8015	(0.7946 , 0.8085)	0.0019	-	-	-
C_Y	0.8000	0.8001	(0.7929 , 0.8073)	0.0002	0.8017	(0.7944 , 0.8089)	0.0021	-	- -	-
***Low* - ϕ_c**										
α_1	-1.0000	-0.9976	(-1.0367 , -0.9587)	0.0024	-0.8960	(-0.9279 , -0.8641)	0.1040	-0.9671	(-0.9893 , -0.9448)	0.0329
α_2	0.5000	0.5022	(0.4658 , 0.5385)	0.0045	0.5933	(0.5621 , 0.6245)	0.1866	0.7720	(0.7500 , 0.7941)	0.5441
β_1	0.5000	0.4963	(0.4308 , 0.5620)	0.0073	0.3069	(0.2568 , 0.3570)	0.3862	0.2698	(0.2423 , 0.2972)	0.4604
β_2	1.0000	0.9967	(0.9483 , 1.0454)	0.0033	0.8973	(0.8588 , 0.9358)	0.1027	0.6992	(0.6711 , 0.7273)	0.3008
C_X	0.8000	0.8000	(0.7933 , 0.8065)	0.0000	0.8016	(0.7947 , 0.8085)	0.0020	-	- -	-
C_Y	0.8000	0.7999	(0.7930 , 0.8068)	0.0001	0.7994	(0.7921 , 0.8067)	0.0008	-	- -	-

Table (3.2) shows the results of larger amount of misclassification error scenarios (that is, $C_Y = 0.8$ and $C_X = 0.8$). It is observed that the point estimates obtained for the misclassification parameters are closer to the true values in scenario 2 (low ϕ_c) than scenario 1 (high ϕ_c error) for both the *dependent* and *independent misclassification error models*. For the β_1 and β_2 parameters, the *dependent misclassification error model* gave point estimates that are closer to the true values than the *naive* and *independent misclassification error model*. However, the *independent misclassification error model* is superior to the *naive model*. Also, notice from Table 3.2 that all the β_1 and β_2 parameters for the *naive model* in both scenario 1 and scenario 2 have the estimated mean value to be smaller than the true value. The model which produced the largest bias in estimating the β_1 and β_2 parameters is the *naive model* with Low ϕ_c ($\hat{\beta}_1 = 0.2698$ and $\hat{\beta}_1 = 0.6992$).

From Table 3.3 it is seen that, where less misclassification error scenario is considered in the covariate ($C_X = 0.95$) and larger amount of misclassification error scenario is considered in the response variable ($C_Y = 0.8$) that the estimates of the misclassification error parameters are quite close to the true values in both scenario 3 and scenario 4 for the *dependent* and *independent misclassification error models*. The *dependent misclassification error model* produced the lowest bias in estimating the β_1 and β_2 parameters, followed by the *independent misclassification model* for both the high ϕ_c scenario and low ϕ_c scenario. The naive model with low ϕ_c produced the worst mean estimated values ($\hat{\beta}_1 = 0.3914$ and $\hat{\beta}_2 = 0.8164$).

A look at Table 3.4, where larger amount of misclassification error scenarios are considered for the covariate ($C_X = 0.8$) and less misclassification scenarios are considered for the response variable ($C_Y = 0.95$) shows that the misclassification parameters have been accurately estimated by both the dependent and independent misclassification error model for both scenario 5 and scenario 6. For the β_1 and β_2 parameters, a similar pattern is observed as in Tables 3.2 and 3.3, the dependent misclassification model gave estimates that are closest to true value. The naive models produce the largest bias in Table 3.4, that is ($\hat{\beta}_1 = 0.4534$ and $\hat{\beta}_2 = 0.8911$) for high ϕ_c scenario and ($\hat{\beta}_1 = 0.4549$ and $\hat{\beta}_2 = 0.8775$) for low ϕ_c scenario.

When less misclassification scenarios are considered for both the response variable and covariate (that is, $C_X = 0.95$ and $C_Y = 0.95$) as in Table 3.5, the general observation is that, estimates for all the parameters under consideration are closer to the true values than scenarios considered in Tables 3.2, 3.3 and 3.4. The misclassification parameters were accurately estimated by the *dependent* and *independent misclassification error models* for scenario 7 and scenario 8. Considering the β_1 and β_2 parameters, the dependent misclassification error model remains the best in estimating the true parameter values. For the low ϕ_c scenario, the naive model produced largest bias. However, when high ϕ_c scenario is considered, the estimates of β_1 and β_2 parameters are quite close to the true values than the *independent misclassification error model* ($\hat{\beta}_1 = 0.4744$ and $\hat{\beta}_1 = 0.9685$). A plot of the average relative bias for the estimated parameters in Figure 3.1, clearly shows that the naive model produces the largest relative bias.

3.4.3 MCMC Diagnostics

The posterior samples are based on two MCMC chains, each having total length 5,000 after a 10,000 burn-in period. Figure (3.3) - (3.20) shows the trace plots, density plots, and autocorrelation plots of the regression and misclassification parameters for the three models for 10% and 50% proportion of validation data. The trace plots show no visible patterns for the parameters; there is evidence of unimodal in the density plot. The autocorrelation drops with increasing lag; these are all indications of convergence and stationarity for the considered parameters.

Table 3.3: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 3 (High ϕ_c) and scenario 4 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-1.0004	(-1.0319 , -0.9689)	0.0004	-1.0128	(-1.0415 , -0.9841)	0.0128	-1.0932	(-1.1149 , -1.0715)	0.0932
α_2	0.5000	0.5004	(0.4699 , 0.5308)	0.0007	0.4917	(0.4635 , 0.5198)	0.0167	0.6640	(0.6430 , 0.6851)	0.3280
β_1	0.5000	0.4997	(0.4532 , 0.5463)	0.0006	0.5060	(0.4683 , 0.5439)	0.0121	0.4306	(0.4027 , 0.4585)	0.1388
β_2	1.0000	1.0003	(0.9615 , 1.0392)	0.0003	1.0231	(0.9882 , 1.0579)	0.0231	0.8545	(0.8274 , 0.8816)	0.1455
C_X	0.9500	0.9500	(0.9457 , 0.9541)	0.0000	0.9498	(0.9456 , 0.9539)	0.0002	-	- -	-
C_Y	0.8000	0.8002	(0.7929 , 0.8074)	0.0002	0.8007	(0.7934 , 0.8079)	0.0008	-	- -	-
***Low* - ϕ_c**										
α_1	-1.0000	-0.9975	(-1.0288 , -0.9663)	0.0025	-0.9785	(-1.0071 , -0.9498)	0.0215	-1.0650	(-1.0867 , -1.0433)	0.0650
α_2	0.5000	0.5013	(0.4712 , 0.5313)	0.0026	0.5193	(0.4911 , 0.5474)	0.0386	0.6879	(0.6668 , 0.7090)	0.3757
β_1	0.5000	0.4964	(0.4516 , 0.5413)	0.0071	0.4590	(0.4213 , 0.4968)	0.0821	0.3914	(0.3634 , 0.4193)	0.2172
β_2	1.0000	0.9985	(0.9606 , 1.0365)	0.0015	0.9795	(0.9447 , 1.0142)	0.0205	0.8164	(0.7894 , 0.8434)	0.1836
C_X	0.9500	0.9498	(0.9456 , 0.9539)	0.0002	0.9499	(0.9456 , 0.9540)	0.0001	-	-	-
C_Y	0.8000	0.7994	(0.7921 , 0.8066)	0.0007	0.7999	(0.7926 , 0.8072)	0.0001	-	- -	-

Table 3.4: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 5 (High ϕ_c) and scenario 6 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9999	(-1.0306 , -0.9693)	0.0001	-1.0063	(-1.0342 , -0.9784)	0.0063	-0.9737	(-0.9957 , -0.9516)	0.0263
α_2	0.5000	0.5003	(0.4714 , 0.5290)	0.0005	0.4958	(0.4692 , 0.5224)	0.0083	0.5769	(0.5554 , 0.5985)	0.1539
β_1	0.5000	0.4997	(0.4472 , 0.5524)	0.0006	0.4952	(0.4494 , 0.5411)	0.0096	0.4534	(0.4260 , 0.4807)	0.0932
β_2	1.0000	0.9998	(0.9615 , 1.0382)	0.0002	1.0183	(0.9840 , 1.0528)	0.0183	0.8911	(0.8630 , 0.9191)	0.1089
C_X	0.8000	0.7999	(0.7929 , 0.8069)	0.0001	0.8002	(0.7932 , 0.8071)	0.0002	-	-	-
C_Y	0.9500	0.9499	(0.9456 , 0.9540)	0.0001	0.9497	(0.9454 , 0.9539)	0.0003	-	-	-
***Low* - ϕ_c**										
α_1	-1.0000	-1.0000	(-1.0300 , -0.9699)	0.0000	-1.0021	(-1.0300 , -0.9743)	0.0021	-0.9689	(-0.9910 , -0.9468)	0.0311
α_2	0.5000	0.5000	(0.4716 , 0.5285)	0.0001	0.4981	(0.4715 , 0.5246)	0.0038	0.5800	(0.5585 , 0.6016)	0.1601
β_1	0.5000	0.4992	(0.4471 , 0.5513)	0.0016	0.5013	(0.4555 , 0.5472)	0.0026	0.4549	(0.4275 , 0.4822)	0.0903
β_2	1.0000	1.0005	(0.9635 , 1.0377)	0.0005	1.0045	(0.9702 , 1.0389)	0.0045	0.8775	(0.8494 , 0.9055)	0.1225
C_X	0.8000	0.7992	(0.7922 , 0.8061)	0.0011	0.7999	(0.7929 , 0.8069)	0.0001	-	- -	-
C_Y	0.9500	0.9499	(0.9455 , 0.9540)	0.0002	0.9497	(0.9454 , 0.9539)	0.0003	-	- -	-

Table 3.5: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 7 (High ϕ_c) and scenario 8 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9996	(-1.0258 , -0.9735)	0.0004	-1.0096	(-1.0340 , -0.9852)	0.0096	-1.0198	(-1.0413 , -0.9983)	0.0198
α_2	0.5000	0.5002	(0.4752 , 0.5251)	0.0004	0.4936	(0.4700 , 0.5172)	0.0128	0.5430	(0.5221 , 0.5638)	0.0859
β_1	0.5000	0.4994	(0.4608 , 0.5381)	0.0013	0.4952	(0.4617 , 0.5288)	0.0095	0.4744	(0.4466 , 0.5021)	0.0513
β_2	1.0000	1.0002	(0.9672 , 1.0333)	0.0002	1.0262	(0.9956 , 1.0568)	0.0262	0.9685	(0.9414 , 0.9956)	0.0315
C_X	0.9500	0.9498	(0.9455 , 0.9539)	0.0002	0.9500	(0.9458 , 0.9541)	0.0000	-	-	-
C_Y	0.9500	0.9497	(0.9454 , 0.9539)	0.0003	0.9499	(0.9457 , 0.9541)	0.0001	-	-	-
***Low* - ϕ_c**										
α_1	-1.0000	-0.9984	(-1.0238 , -0.9731)	0.0016	-0.9809	(-1.0052 , -0.9565)	0.0191	-0.9925	(-1.0140 , -0.9710)	0.0075
α_2	0.5000	0.5004	(0.4759 , 0.5249)	0.0009	0.5166	(0.4930 , 0.5402)	0.0332	0.5654	(0.5445 , 0.5863)	0.1307
β_1	0.5000	0.4971	(0.4609 , 0.5334)	0.0058	0.4624	(0.4288 , 0.4959)	0.0753	0.4432	(0.4154 , 0.4710)	0.1135
β_2	1.0000	0.9999	(0.9679 , 1.0319)	0.0001	0.9819	(0.9513 , 1.0124)	0.0181	0.9251	(0.8980 , 0.9522)	0.0749
C_X	0.9500	0.9491	(0.9451 , 0.9530)	0.0009	0.9498	(0.9456 , 0.9539)	0.0002	-	-	-
C_Y	0.9500	0.9491	(0.9451 , 0.9530)	0.0009	0.9496	(0.9453 , 0.9538)	0.0004	-	-	-

Figure 3.1: Graph of the average relative bias of the misclassification parameters and the regression coefficient for the multi-category model when 10% proportion of validation data is employed.

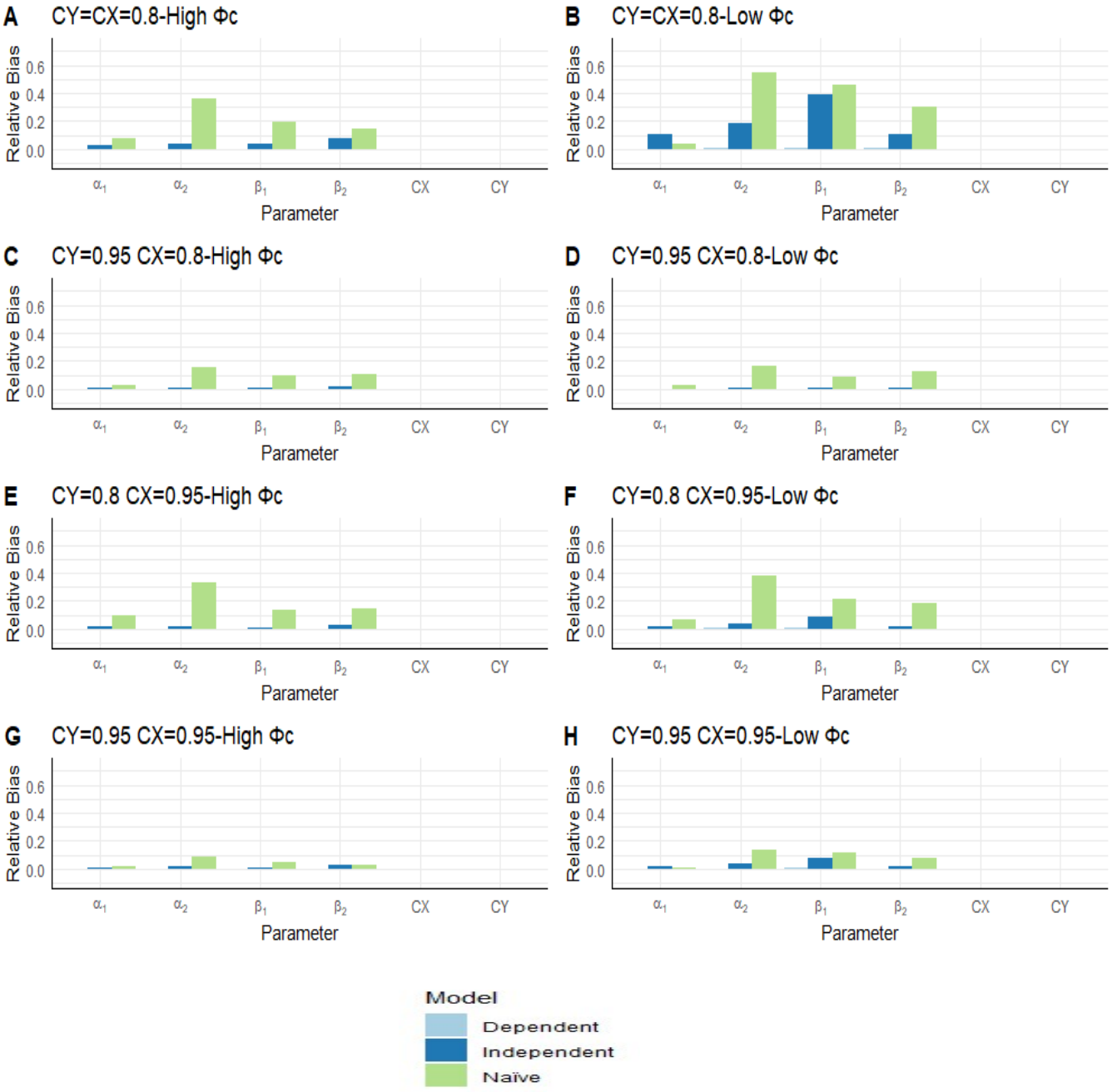


Table 3.6 considers larger amount of misclassification error scenarios for both the covariate and response variable (that is, $C_X = 0.8$ and $C_Y = 0.8$) when proportion of validation data is 50%. It is observed that the estimates of the misclassification parameters are close to the true values for both high ϕ_c (scenario 1) and low ϕ_c (scenario 2). For the β_1 and β_2 parameters, the *dependent misclassification error model* produced estimates that are closest to the true values than the *independent misclassification error model* and the *naive model*. In scenario 1 and scenario 2, the *independent misclassification error model* performs better in estimating the β_1 and β_2 parameters than the *naive model*.

From Table 3.7 where less misclassification error scenario for the covariate ($C_X = 0.95$) and larger amount of misclassification error scenarios for the response variable ($C_Y = 0.8$) are considered it is seen, that for the β_1 and β_2 parameters the estimates are closer to the true values in the *dependent misclassification error model* than the *naive* and *independent misclassification error models*. However, the *naive models* remains the worse performing models for β_1 and β_2 parameters, that is ($\hat{\beta}_1 = 0.4491$ and $\hat{\beta}_1 = 0.8943$) for scenario 3 and ($\hat{\beta}_1 = 0.4202$ and $\hat{\beta}_1 = 0.8670$) for scenario 4.

In Table 3.8 larger amount of misclassification error scenario in covariate ($C_X = 0.8$) and less misclassification scenario in the response variable ($C_Y = 0.95$) are considered. Patterns observed are similar to that in Table 3.7. The models that produced the largest bias in estimating the true β_1 and β_2 parameters value are the *naives models*, that is ($\hat{\beta}_1 = 0.4665$ and $\hat{\beta}_1 = 0.9225$) for scenario 5 and ($\hat{\beta}_1 = 0.4686$ and $\hat{\beta}_1 = 0.9128$) for scenario 6.

Table 3.9 shows less misclassification scenarios for both the covariate and response variable (that is, $C_X = 0.95$ and $C_Y] = 0.95$), although the β_1 and β_2 estimates for the *naive models* are closer to the true value than scenarios with larger amount of misclassification error scenarios, they produce the largest bias, for the high ϕ_c scenarios ($\hat{\beta}_1 = 0.4804$ and $\hat{\beta}_1 = 0.9767$). However, the *naive model* performed better than the *independent misclassification error model* in the high ϕ_c scenario.

Table 3.6: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 1 (High ϕ_c) and scenario 2 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9996	(-1.0256 , -0.9737)	0.0004	-1.0134	(-1.0396 , -0.9873)	0.0134	-1.0563	(-1.0753 , -1.0374)	0.0563
α_2	0.5000	0.5003	(0.4753 , 0.5252)	0.0006	0.4902	(0.4652 , 0.5153)	0.0195	0.6285	(0.6102 , 0.6469)	0.2571
β_1	0.5000	0.4988	(0.4625 , 0.5351)	0.0024	0.4997	(0.4629 , 0.5365)	0.0007	0.4289	(0.4053 0.4525)	0.1422
β_2	1.0000	0.9994	(0.9668 , 1.0320)	0.0006	1.0303	(0.9976 , 1.0629)	0.0303	0.8974	(0.8735 , 0.9212)	0.1026
C_X	0.8000	0.8000	(0.7967 , 0.8034)	0.0000	0.8003	(0.7969 , 0.8036)	0.0004	-	- -	-
C_Y	0.8000	0.8001	(0.7966 , 0.8035)	0.0001	0.8003	(0.7969 , 0.8037)	0.0004	-	- -	-
***Low*- ϕ_c**										
α_1	-1.0000	-1.0000	(-1.0268 , -0.9732)	0.0000	-0.9642	(-0.9902 , -0.9381)	0.0358	-0.9807	(-0.9995 , -0.9618)	0.0193
α_2	0.5000	0.5004	(0.4750 , 0.5258)	0.0008	0.5322	(0.5071 , 0.5573)	0.0644	0.6938	(0.6754 , 0.7123)	0.3876
β_1	0.5000	0.4997	(0.4622 , 0.5373)	0.0006	0.4399	(0.4032 , 0.4768)	0.1201	0.3330	(0.3094 , 0.3566)	0.3339
β_2	1.0000	0.9996	(0.9662 , 1.0331)	0.0004	0.9589	(0.9263 , 0.9915)	0.0411	0.7897	(0.7660 0.8134)	0.2103
C_X	0.8000	0.7999	(0.7968 , 0.8030)	0.0001	0.8000	(0.7966 , 0.8033)	0.0000	-	- -	-
C_Y	0.8000	0.8000	(0.7968 , 0.8032)	0.0000	0.7998	(0.7964 , 0.8032)	0.0002	-	- -	-

Table 3.7: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 3 (High ϕ_c) and scenario 4 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9996	(-1.0247 , -0.9745)	0.0004	-1.0053	(-1.0304 , -0.9803)	0.0053	-1.0682	(-1.0867 , -1.0498)	0.0682
α_2	0.5000	0.5005	(0.4763 , 0.5247)	0.0010	0.4960	(0.4719 , 0.5202)	0.0080	0.6194	(0.6015 , 0.6372)	0.2387
β_1	0.5000	0.4992	(0.4658 , 0.5326)	0.0017	0.5027	(0.4693 , 0.5360)	0.0053	0.4491	(0.4252 , 0.4730)	0.1018
β_2	1.0000	0.9994	(0.9681 , 1.0307)	0.0006	1.0102	(0.9790 , 1.0413)	0.0102	0.8943	(0.8712 , 0.9174)	0.1057
C_X	0.9500	0.9499	(0.9480 , 0.9518)	0.0001	0.9499	(0.9480 , 0.9518)	0.0001	-	- -	-
C_Y	0.8000	0.8000	(0.7966 , 0.8034)	0.0000	0.8001	(0.7967 , 0.8035)	0.0002	-	- -	-
***Low* - ϕ_c**										
α_1	-1.0000	-0.9995	(-1.0248 , -0.9742)	0.0005	-0.9903	(-1.0153 , -0.9653)	0.0097	-1.0481	(-1.0665 , -1.0296)	0.0481
α_2	0.5000	0.5008	(0.4765 , 0.5251)	0.0016	0.5092	(0.4851 , 0.5333)	0.0184	0.6367	(0.6188 , 0.6546)	0.2734
β_1	0.5000	0.4988	(0.4650 , 0.5325)	0.0024	0.4814	(0.4481 , 0.5148)	0.0372	0.4202	(0.3962 , 0.4441)	0.1596
β_2	1.0000	0.9990	(0.9676 , 1.0306)	0.0010	0.9898	(0.9586 , 1.0209)	0.0102	0.8670	(0.8439 , 0.8900)	0.1330
C_X	0.9500	0.9499	(0.9480 , 0.9518)	0.0001	0.9499	(0.9480 , 0.9518)	0.0001	-	- -	-
C_Y	0.8000	0.8000	(0.7966 , 0.8034)	0.0000	0.8001	(0.7967 , 0.8035)	0.0001	-	- -	-

Table 3.8: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 5 (High ϕ_c) and scenario 6 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9998	(-1.0246 , -0.9751)	0.0002	-1.0035	(-1.0282 , -0.9788)	0.0035	-0.9817	(-1.0004 , -0.9630)	0.0183
α_2	0.5000	0.5004	(0.4767 , 0.5240)	0.0008	0.4978	(0.4742 , 0.5214)	0.0044	0.5555	(0.5373 , 0.5737)	0.1110
β_1	0.5000	0.4993	(0.4637 , 0.5350)	0.0014	0.4994	(0.4637 , 0.5351)	0.0012	0.4665	(0.4429 , 0.4900)	0.0671
β_2	1.0000	0.9995	(0.9684 , 1.0307)	0.0005	1.0080	(0.9769 , 1.0390)	0.0080	0.9225	(0.8988 , 0.9462)	0.0775
C_X	0.8000	0.8000	(0.7966 , 0.8033)	0.0000	0.8000	(0.7967 , 0.8034)	0.0001	-	- -	-
C_Y	0.9500	0.9500	(0.9480 , 0.9518)	0.0000	0.9500	(0.9480 , 0.9518)	0.0001	-	- -	-
***Low* - ϕ_c**										
α_1	-1.0000	-0.9997	(-1.0246 , -0.9750)	0.0003	-1.0010	(-1.0257 , -0.9763)	0.0010	-0.9784	(-0.9971 , -0.9597)	0.0216
α_2	0.5000	0.5003	(0.4766 , 0.5241)	0.0007	0.4994	(0.4758 , 0.5230)	0.0013	0.5576	(0.5394 , 0.5758)	0.1151
β_1	0.5000	0.4995	(0.4636 , 0.5355)	0.0009	0.5007	(0.4650 , 0.5364)	0.0014	0.4686	(0.4450 , 0.4921)	0.0628
β_2	1.0000	0.9994	(0.9682 , 1.0308)	0.0006	1.0015	(0.9704 , 1.0325)	0.0015	0.9128	(0.8891 , 0.9366)	0.0872
C_X	0.8000	0.7997	(0.7964 , 0.8031)	0.0003	0.7999	(0.7966 , 0.8033)	0.0001	-	- -	-
C_Y	0.9500	0.9500	(0.9481 , 0.9519)	0.0000	0.9500	(0.9481 , 0.9519)	0.0000	-	- -	-

Figure 3.2: Graph of the average relative bias of the misclassification parameters and the regression coefficient for the multi-category model when 50% proportion of validation data is employed.

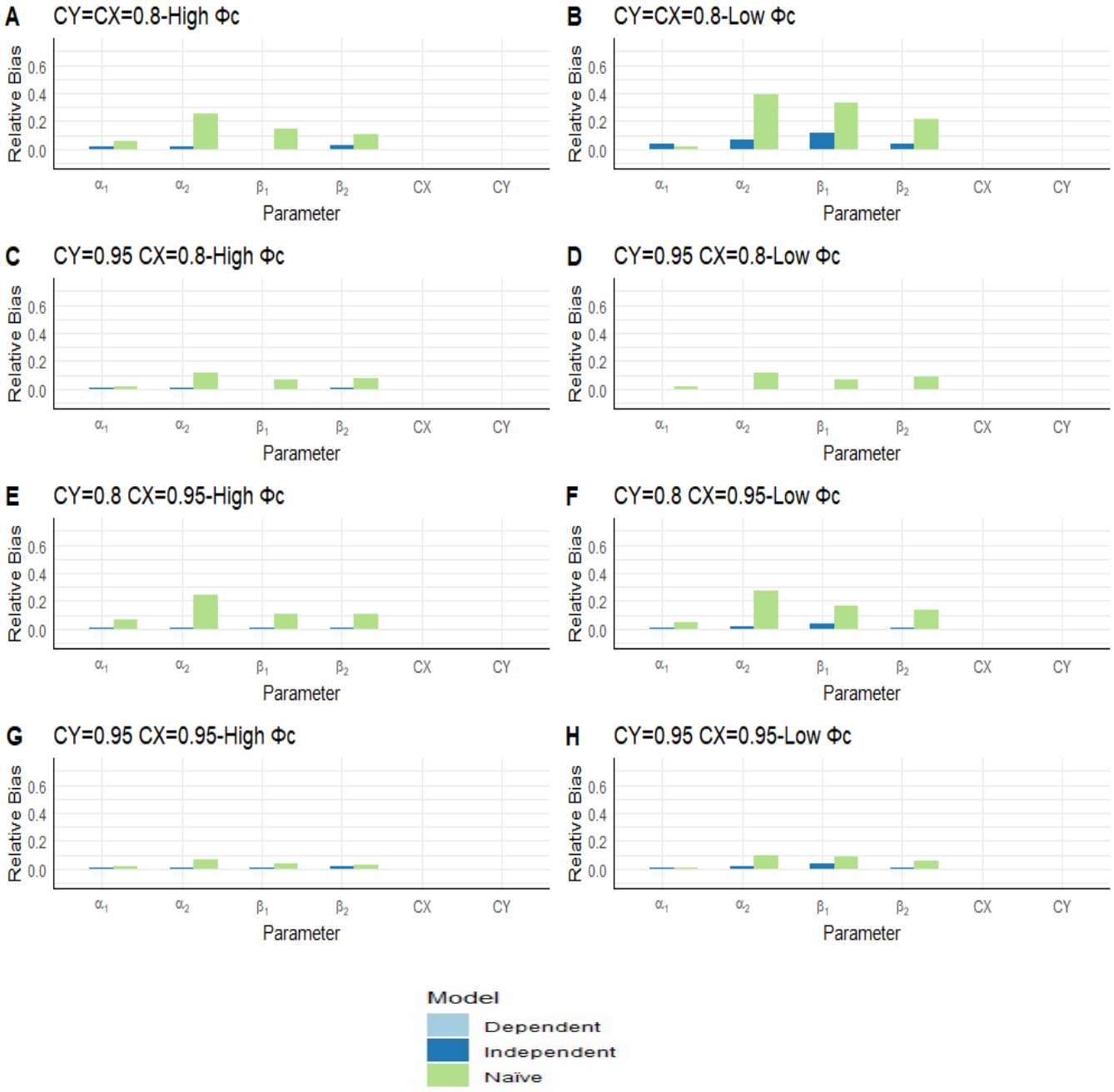


Figure 3.3: Trace plots for the posterior samples under the *dependent Model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

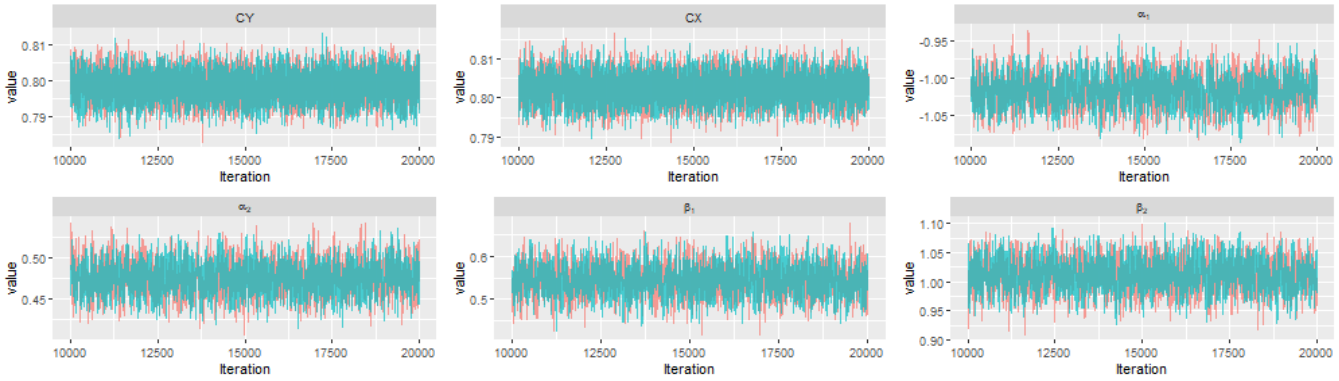


Figure 3.4: Trace plots for the posterior samples under the *independent Model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

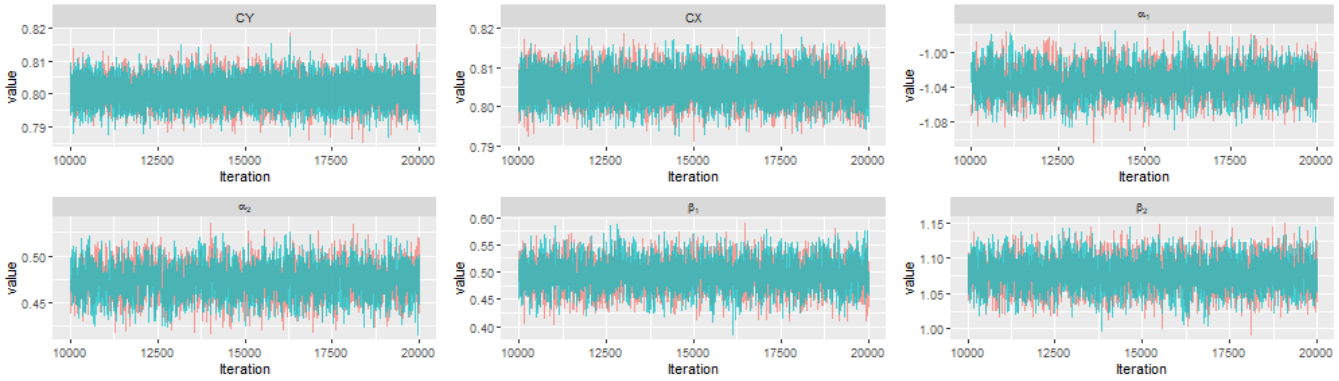


Figure 3.5: Trace plots for the posterior samples under the *naive model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

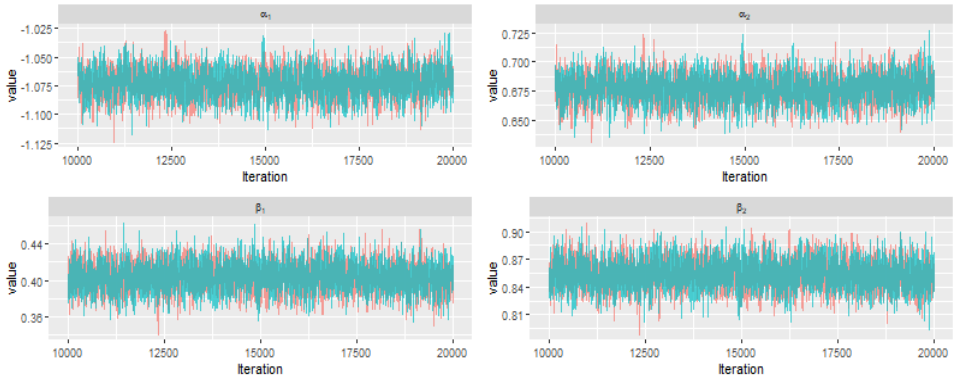


Figure 3.6: Density plots for the posterior samples under the *dependent Model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

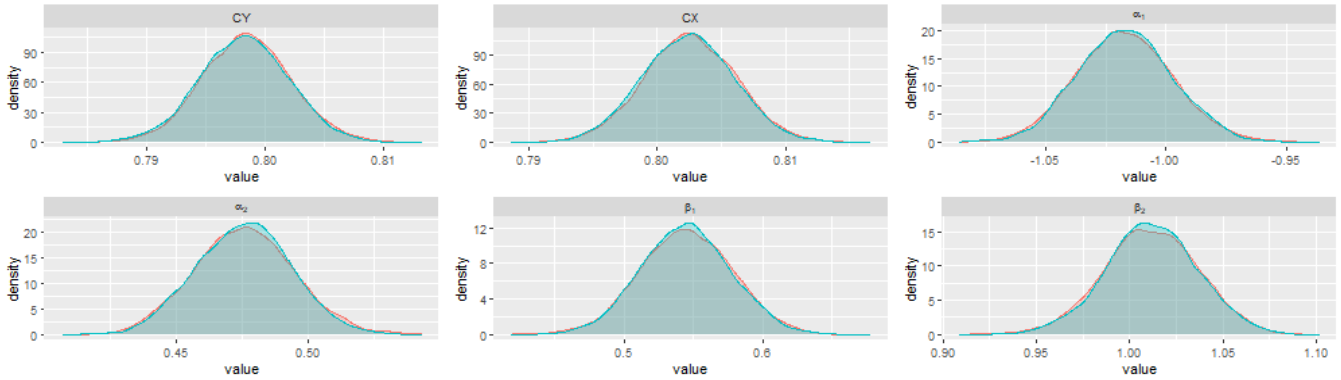


Figure 3.7: Density plots for the posterior samples under the *independent Model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

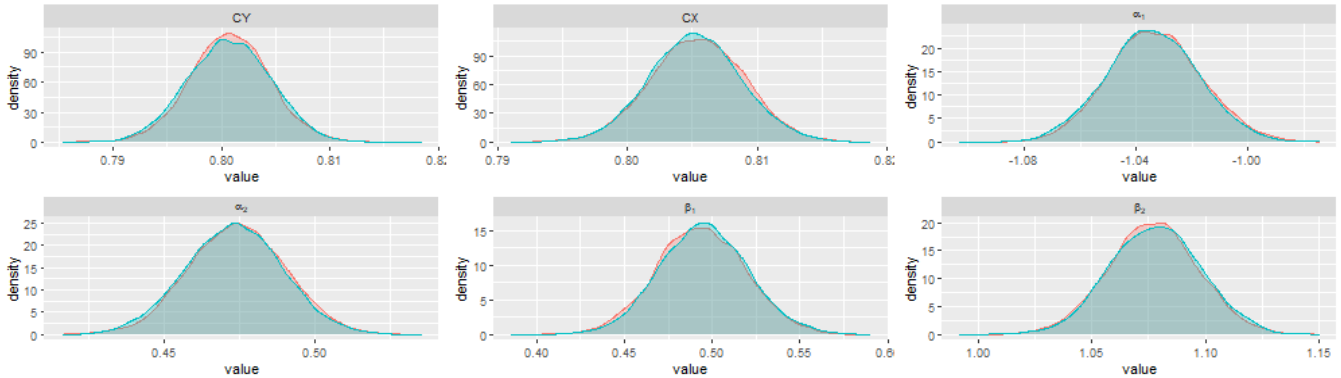


Figure 3.8: Density plots for the posterior samples under the *naive Model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

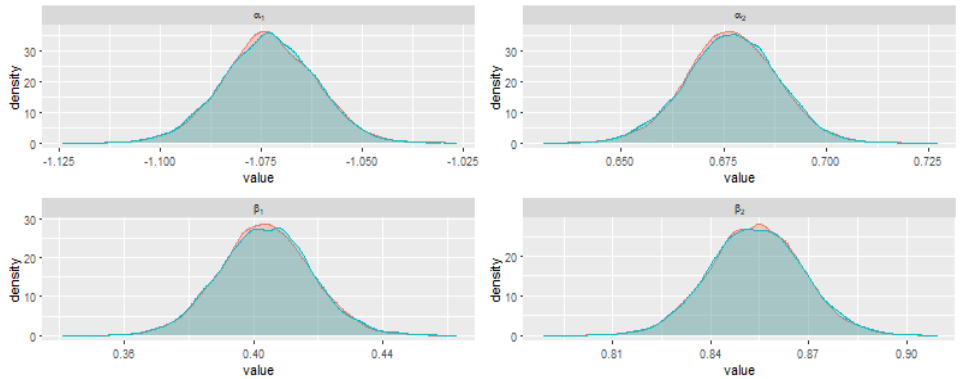


Figure 3.9: Autocorrelation plots for the posterior samples under the *dependent Model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

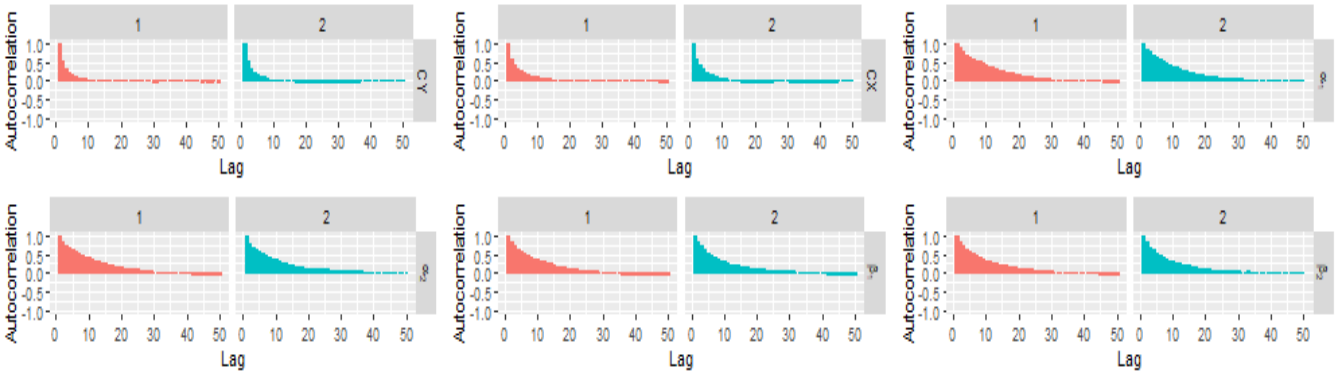


Figure 3.10: Autocorrelation plots for the posterior samples under the *independent model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

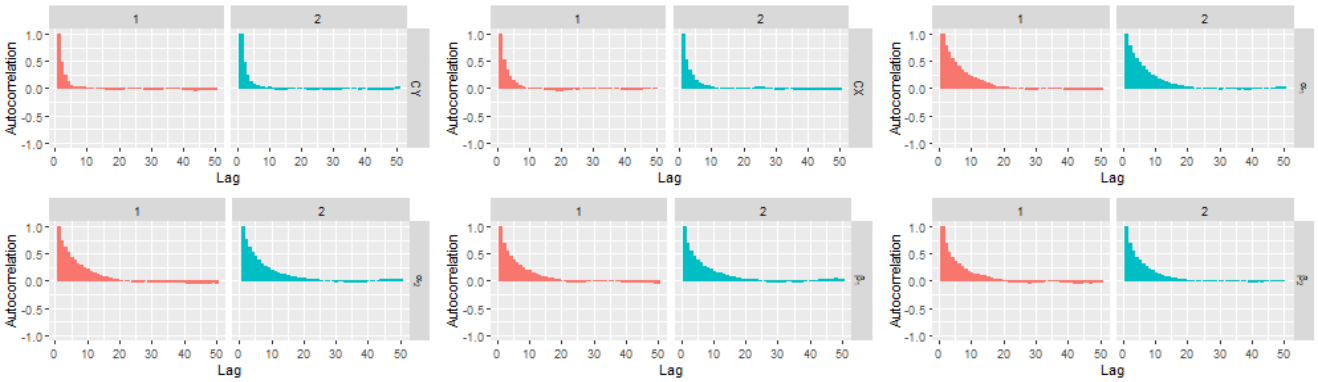


Figure 3.11: Autocorrelation plots for the posterior samples under the *naive model* of a trinary misclassification in both the response variable and covariate, with 10% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

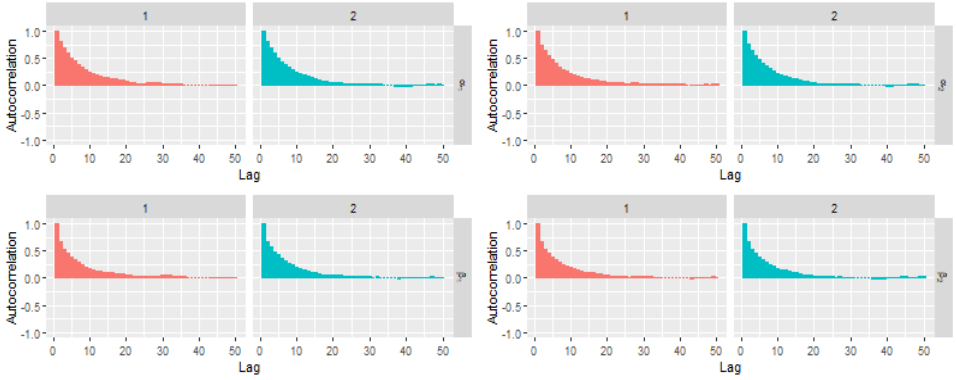


Figure 3.12: Trace plots for the posterior samples under the *dependent Model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

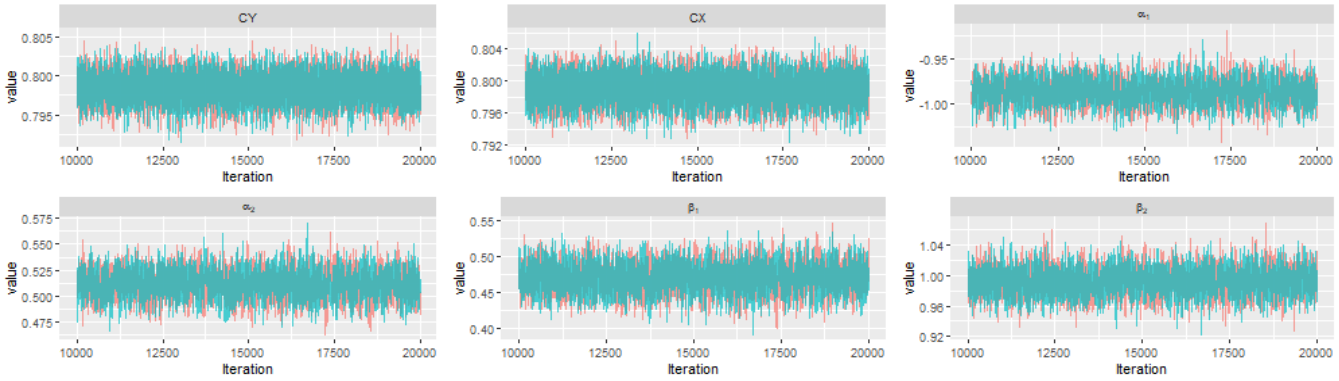


Figure 3.13: Trace plots for the posterior samples under the *independent Model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

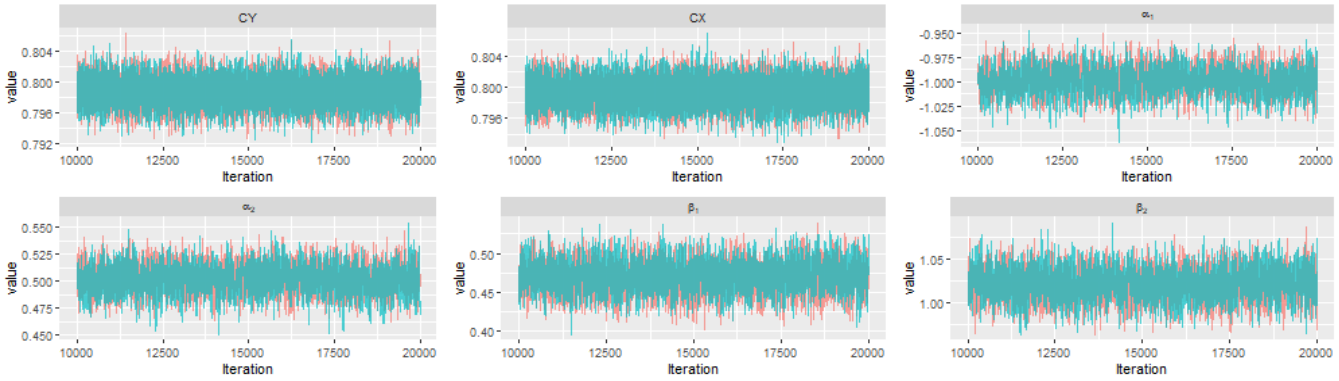


Figure 3.14: Trace plots for the posterior samples under the *naive model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$,, High ϕ_c)

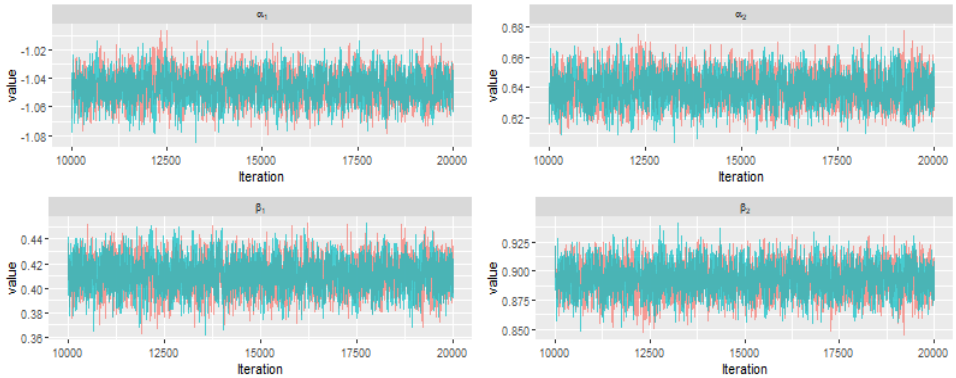


Figure 3.15: Density plots for the posterior samples under the *dependent Model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

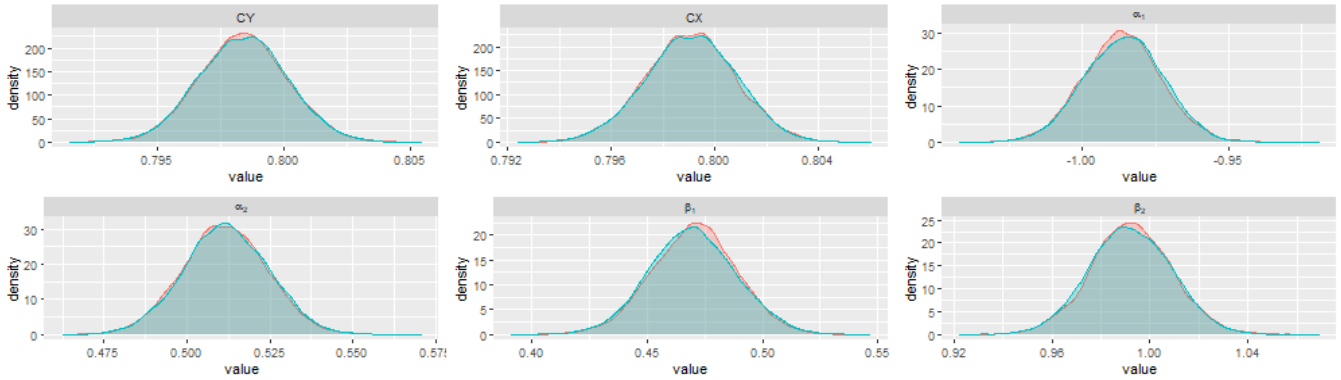


Figure 3.16: Density plots for the posterior samples under the *independent Model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

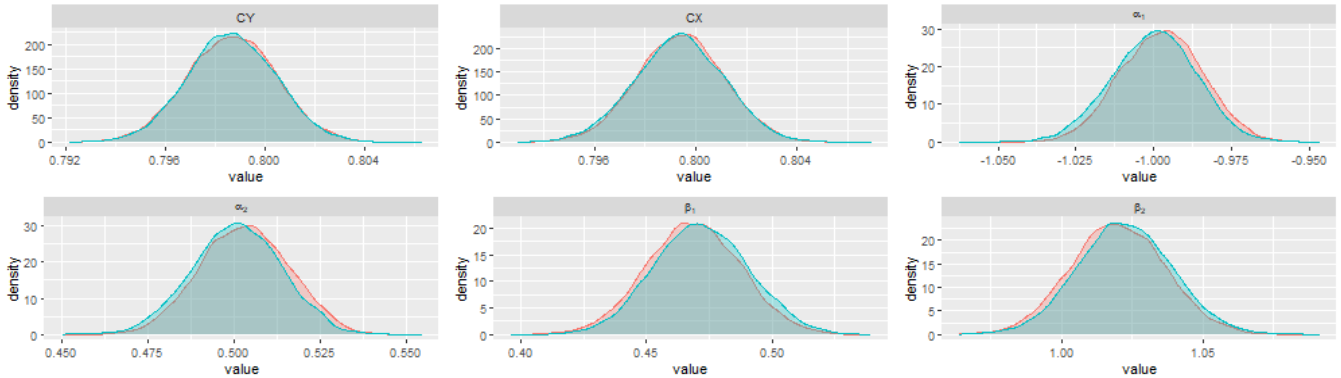


Figure 3.17: Density plots for the posterior samples under the *naive Model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

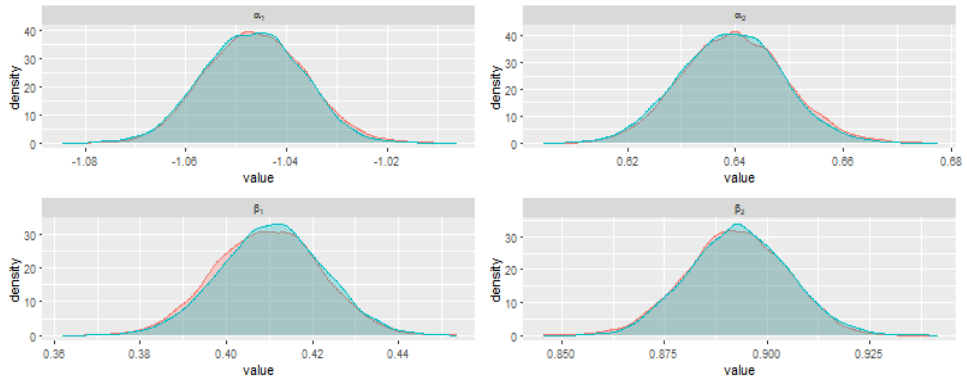


Figure 3.18: Autocorrelation plots for the posterior samples under the *dependent Model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

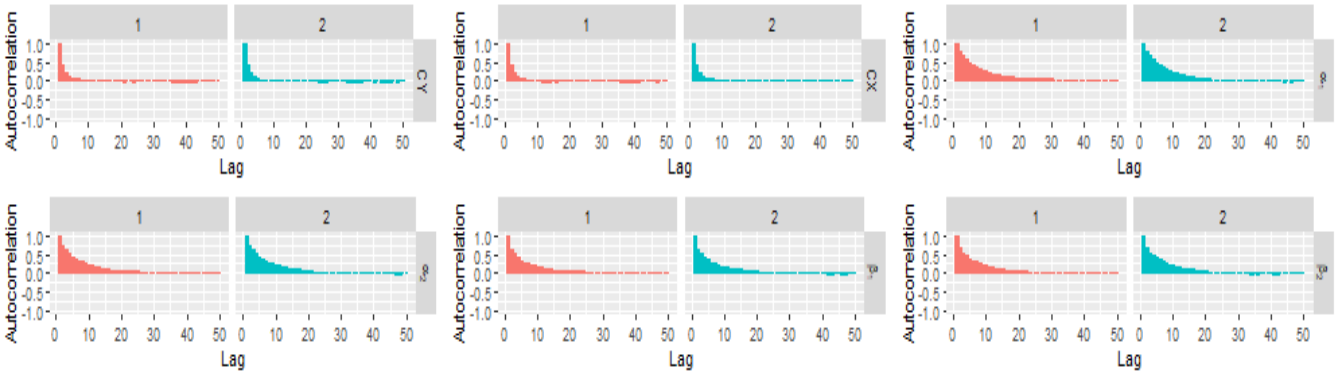


Figure 3.19: Autocorrelation plots for the posterior samples under the *independent model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

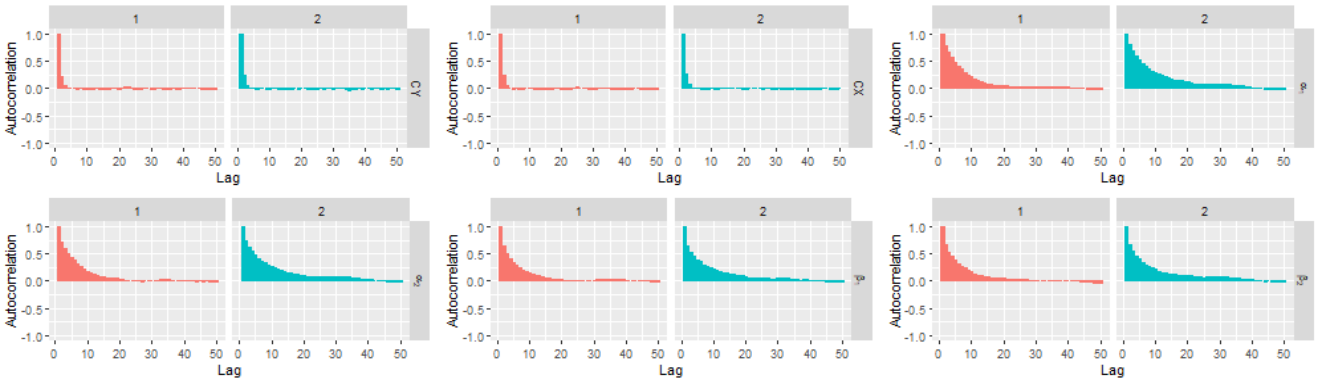


Figure 3.20: Autocorrelation plots for the posterior samples under the *naive model* of a trinary misclassification in both the response variable and covariate, with 50% proportion of validation data for scenario 1. ($C_Y = 0.8$, $C_X = 0.8$, High ϕ_c)

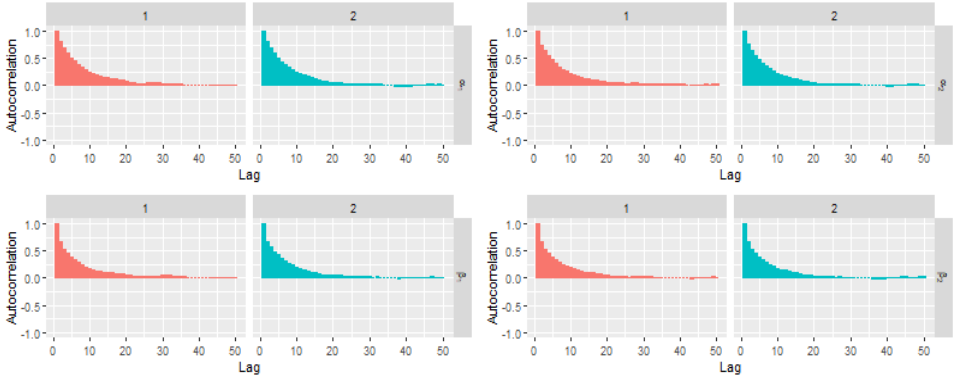


Table 3.9: Posterior summaries and relative bias for the model parmeters under the Dependent Model, Independent Model and Naive Model in multi-category misclassification using simulated data (N=100,000) with 10% proportion of validation data. Table employs settings from scenario 7 (High ϕ_c) and scenario 8 (Low ϕ_c).

		Dependent Model			**Independent Model**			**Naive Model**		
Parameter	True Value	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias	Mean	95% CI	Relative Bias
***High* - ϕ_c**										
α_1	-1.0000	-0.9996	(-1.0229 , -0.9762)	0.0004	-1.0049	(-1.0283 , -0.9815)	0.0049	-1.0143	(-1.0327 , -0.9959)	0.0143
α_2	0.5000	0.5006	(0.4781 , 0.5231)	0.0012	0.4970	(0.4744 , 0.5195)	0.0061	0.5319	(0.5142 , 0.5498)	0.0639
β_1	0.5000	0.4991	(0.4677 , 0.5305)	0.0019	0.4975	(0.4660 , 0.5291)	0.0050	0.4804	(0.4565 , 0.5042)	0.0393
β_2	1.0000	0.9996	(0.9703 , 1.0289)	0.0004	1.0131	(0.9837 , 1.0425)	0.0131	0.9767	(0.9535 , 0.9998)	0.0233
C_X	0.9500	0.9499	(0.9480 , 0.9518)	0.0001	0.9500	(0.9480 , 0.9518)	0.0000	-	- -	-
C_Y	0.9500	0.9499	(0.9480 , 0.9518)	0.0001	0.9500	(0.9481 , 0.9518)	0.0000	-	- -	-
***Low* - ϕ_c**										
α_1	-1.0000	-0.9998	(-1.0233 , -0.9762)	0.0002	-0.9904	(-1.0138 , -0.9671)	0.0096	-0.9948	(-1.0132 , -0.9764)	0.0052
α_2	0.5000	0.5006	(0.4780 , 0.5233)	0.0012	0.5090	(0.4864 , 0.5315)	0.0179	0.5481	(0.5303 , 0.5659)	0.0962
β_1	0.5000	0.4993	(0.4676 , 0.5311)	0.0014	0.4812	(0.4497 , 0.5128)	0.0377	0.4583	(0.4344 , 0.4822)	0.0834
β_2	1.0000	0.9996	(0.9700 , 1.0291)	0.0004	0.9901	(0.9607 , 1.0194)	0.0099	0.9453	(0.9221 0.9684)	0.0547
C_X	0.9500	0.9499	(0.9482 , 0.9516)	0.0001	0.9500	(0.9481 , 0.9519)	0.0000	-	- -	-
C_Y	0.9500	0.9499	(0.9482 , 0.9515)	0.0001	0.9500	(0.9481 , 0.9518)	0.0000	-	- -	-

Discussion

In chapter 3, the model that accounts for dependent misclassification errors discussed in Chapter 2 was extended to a multi-category setting. Simulation studies were conducted on a trinary response variable and a trinary covariate. Data were generated from a *dependent error misclassification error model*; however, an independent misclassification error model and a naive model were fitted to the data to learn the consequencies of fitting the wrong model. In the case where 10% proportion of validation data is employed, it was observed that, although the *dependent misclassification error model* and the *independent misclassification error model* estimates β_1 and β_2 were close to the true values, the *dependent misclassification error model* was better than the *independent misclassification error model*. The model that produced the largest bias was the *naive model* for scenarios which considered both high ϕ_c and low ϕ_c. The patterns observed for cases that considered 50% proportion of validation data are similar to those observed for 10% proportion of validation data; however, in general, the relative biases are less for 50% proportion of validation data. In chapter 4, the proposed model for joint misclassification in both the response variable and covariate is illustrated through a real data example.

Chapter 4

Real Data Example

4.1 Data Description

This chapter aims to illustrate the proposed model discussed in previous chapters using a real data example. This data example is a cross-sectional analysis of the Bacterial Vaginosis (BV) and Trichomoniasis (TRICH) status of women enrolled in the HIV Epidemiology Research Study (HERS). This dataset was obtained from Tang et al. (2013). The HERS study is a multi-center prospective study that enrolled 1310 women from four cities in the United States from 1993 to 1995.

A unique feature that makes this data example suitable for the illustration of our proposed model is that two different methods measure both the response variable and covariate; an error-prone method and an error-free method arguably a gold standard method. Here the response variable is BV; the error-prone method for BV is a clinical-based method (CLIN), which employs a modified Amsel's criteria. The error-free method is a Laboratory-based test (LAB). The covariate Trichomoniasis diagnosis for the error-prone method is a microscopic evaluation of wet preparation of genital secretion, referred mostly to as a wet mount procedure (WET) and a culture (CULT) was used to assess the error-free method. The error-prone procedures for the two conditions are relatively low cost and convenient.

A total number of 916 patients with complete observations on both the error-free and error-prone diagnosis of TRICH and BV at the fourth HERS visit are considered. The justification for using the fourth visit data of the HERS study is stated in Tang et al. (2003). A random subsample, which selected a quarter of the total sample size, was used as the validation dataset. Table (4.1) and (4.2) summarizes the resulting main and validation samples respectively. The main and validation sample sizes are $Nm = 687$ and $Nv = 229$

Table 4.1: Main Data of the Fourth HERS Visit (Tang et al (2013))

			TRICH (WET) X^*		
			1	**0**	Total
BV	Y^*	**1**	29	138	167
(CLIN)		**0**	23	497	520
		Total	52	635	687

respectively. From the samples employed, the prevalence of BV through the clinical method is 7.5%, while the more expensive LAB method serving as the gold standard method has a prevalence of 18.2%. The prevalence of TRICH is 24.5% when assessed by the wet mount; however, the culture method's prevalence of TRICH is 40%.

Background

Trichomoniasis and Bacterial Vaginosis are two of the three diseases most frequently associated with abnormal vaginal discharge, elevated vaginal pH, and a shift in vaginal flora. TRICH is the most prevalent curable Sexually Transmitted Disease (STD), which is more common in women than men. The global annual incidence of TRICH is estimated to be over 170 million cases [47]. More than 8 million new cases are reported annually in North America. BV, on the other hand, although not considered an STD, has multiple sexual partners included in its risk factors. Other risk factors include douching, smoking, and low socioeconomic status.

Both BV and TRICH are associated with increase risk of HIV acquistion [54], preterm birth and other adverse pregnancy outcome including premature rupture of membranes [15], infertility [41] and pelvic inflamatory diseases[12, 38]. These two diseases often occur concurrently and studies have shown an association between them [6, 43].

Table 4.2: Validation Data of the Fourth HERS Visit (Tang et al (2013))

Study	CLIN Y^*	WET X^*	LAB Y	CULT X	Count
	1	**1**	**1**	**1**	7
	1	**1**	**1**	**0**	0
	1	**1**	**0**	**1**	3
	1	**1**	**0**	**0**	0
	1	**0**	**1**	**1**	11
Validation	**1**	**0**	**1**	**0**	28
Data	**1**	**0**	**0**	**1**	0
	1	**0**	**0**	**0**	8
	0	**1**	**1**	**1**	2
	0	**1**	**1**	**0**	0
	0	**1**	**0**	**1**	4
	0	**1**	**0**	**0**	1
	0	**0**	**1**	**1**	11
	0	**0**	**1**	**0**	34
	0	**0**	**0**	**1**	11
	0	**0**	**0**	**0**	109

4.2 Analysis and Results

This analysis's particular interest is to establish how the association between TRICH and BV is affected by dependence error misclassification. We fit the Fourth HERS visit data to three different models: The dependent misclassification error model (adjusts for misclassification and dependence error). The independent misclassification error model (adjusts for misclassification but ignores dependence error). The naive model (ignore both misclassification). Bayesian MCMC sampling in R via the Rjags Package is implemented [42] for parameter estimation of the three models. Please refer to section (2.2.1) for likelihood functions used

in the analysis. The specific prior distributions for the parameters are:

- the misclassification parameters SN, SP, a fairly diffuse prior is selected such that the 95% equal-tail interval is (0.55, 0.95). The crude sensitivity and specificity from the real data example is quite low, hence a more flexible constraint $SP + SN - 1 > 0$ is employed. Truncated beta distributions is used for the priors of the sensitivity and specificity parameters; that is,

$$\begin{aligned} SN &\sim Beta(11, 3)I(1 - SP,), \\ SP &\sim Beta(11, 3), \end{aligned}$$

  where, $I(1 - SP)$ is an indicator function that equal 1 if input is greater than $1 - SP$ and 0 otherwise.

- weakly informative priors are chosen for the regression parameters for the models because no essential prior information is available, that is,

$$\beta_k \sim N(0, 1000), k = 0, 1.$$

- the dependence parameters $D_{11}, D_{10}, D_{01}, D_{00}$ are constrained to lie within an interval determined by the misclassification parameters, therefore the priors for dependence parameters are chosen to follow a uniform distribution also constrained within the said interval.

For the posterior samples of the three models, two MCMC chains are used, each with 10000 iterations and the first half is discarded as burn-in. Table (4.3) shows the posterior mean, standard deviation (SD) and credible interval (CI) of parameters for each model. The sample size for the real data is lower than that used in the simulation because it is a challenge to obtain a large sample size for validation data. However, there is control over the size to use in the validation data. The larger the size the better.

Notice from 4.3 that all three models employed differ by the estimated β_1 parameter (1.885 for the dependent error model, 2.433 for the independent error model, and 1.533 for the naive model). This indicates the potential benefits of adjusting for misclassification and dependence of error. The trace plots for the posterior samples of the regression parameters

for the three models under consideration are shown in Figure (4.1)-(4.3); the absence of a trend in the trace plots for the two chains is an indication of good mixing an indication of stationarity. From Figure (4.7)-(4.9), it is observed that the autocorrelation plots of the regression parameters show a quick drop from 1 to 0, an indication of stationarity. Density plots of the regression parameter shown in Figure (4.4)-(4.6) for the three models indicate stationarity as there is no evidence of unexpected peaks. In addition to the visual plots discussed above, the Gelman-Rubin R statistics are employed; the value 1 obtained for all posterior estimates confirms stationarity.

4.3 Model selection

In this analysis, the Likelihood Ratio Test (LRT) proposed by Neyman and Pearson (1928) is employed. LRT is strictly reserved for comparing "nested" models. Two models are nested if one is considered as a special case of the other model. The more complex model (model with the most parameters) is compared to a simpler model (model with least parameters) to see if it fits a dataset significantly better. LRT is mostly used to compare models fit by Maximum Likelihood Estimation (MLE). However, in large samples, MLE and confidence interval coincides with the posterior mean and the credible interval of Bayesian [44]. It has also been established that when flat or weakly informative priors are used in Bayesian analysis, estimates obtained are very close to MLE. The LRT statistic approximately follows a chi-squared distribution, where the degree of freedom is the number of additional parameters in the more complex model. Considering the dependent and independent models in the study, notice that the independent model differs from the dependent model by the addition of the four dependence parameters $D_{11}, D_{10}, D_{01}, D_{00}$. Therefore, it can be said that these two models are hierarchically nested, a crucial requirement of the LRT.

Let η and θ represent the collection of the parameters for the independent and dependent model respectively, that is,

$$\begin{aligned} \eta &= (SN_X, SP_X, SN_Y, SP_Y, p_X, \beta_0, \beta_1), \\ \theta &= (SN_X, SP_X, SN_Y, SP_Y, p_X, \beta_0, \beta_1, D_{11}, D_{10}, D_{01}, D_{00}), \end{aligned} \tag{4.1}$$

where, degrees of freedom for the test equals the difference in the number of parameters for

the two models. Degree of freedom $= 11 - 7 = 4$. The hypothesis $H_0 : D_{11} = D_{10} = D_{01} = D_{00} = 0$ is tested against H_1 : at least one of the D_{ij} parameters is not equal to 0.

Let the likelihood functions of the dependent and independent model be represented by $L(\eta)$ and $L(\theta)$ respectively. 1This leads to,

$$\begin{aligned} LRT = & \quad -2\log\left(\frac{L(\eta)}{L(\theta)}\right). \\ = & \qquad 52.6078 \end{aligned} \tag{4.2}$$

The p-value for this test is 1.029e-10. This suggests a strong evidence against the null hypothesis. The dependent model is a significant improvement over the independent model.

Table 4.3: Posterior Mean, Standard deviation (SD) and the 95% Credible Interval (CI) of Parameters for the fourth HERS visit data under the Dependent Model.

	Dependent Model		
Parameter	Mean	SD	95% CI
β_0	-1.326	0.224	(-1.794 , -0.915)
β_1	1.158	0.325	(0.547 , 1.802)
SNX	0.274	0.034	(0.211 , 0.345)
SNY	0.583	0.051	(0.478 , 0.680)
SPX	0.961	0.011	(0.937 , 0.979)
SPY	0.963	0.023	(0.908 , 0.995)
D_{11}	0.015	0.024	(-0.032 , 0.061)
D_{10}	-0.011	0.007	(-0.023 , 0.003)
D_{01}	-0.006	0.012	(-0.035 , 0.013)
D_{00}	0.006	0.006	(-0.001 , 0.022)
	Independent Model		
Parameter	Mean	SD	95% CI
β_0	-1.452	0.216	(-1.901 , -1.057)
β_1	1.351	0.313	(0.753 , 1.975)
SNX	0.281	0.033	(0.220 , 0.351)
SNY	0.615	0.047	(0.520 , 0.706)
SPX	0.967	0.009	(0.947 , 0.983)
SPY	0.969	0.019	(0.924 , 0.996)
	Naive Model		
Parameter	Mean	SD	95% CI
β_0	-1.567	0.087	(-1.744 , -1.399)
β_1	0.930	0.172	(0.594 , 1.262)

Figure 4.1: Trace plots for posterior samples of $\beta_k (k = 0, 1)$ under the Dependent Model.

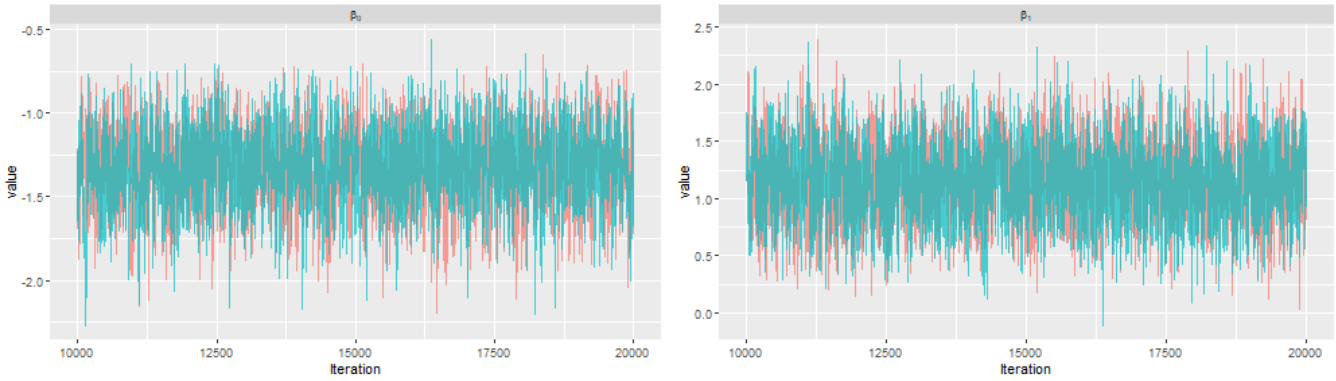


Figure 4.2: Trace plots for posterior samples of $\beta_k (k = 0, 1)$ under the Independent Model.

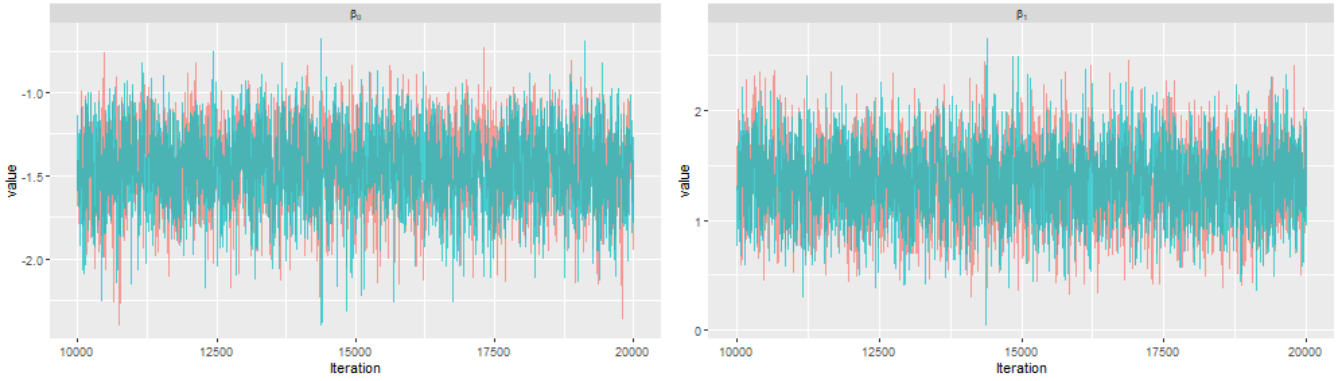


Figure 4.3: Trace plots for posterior samples of $\beta_k (k = 0, 1)$ under the Naive Model.

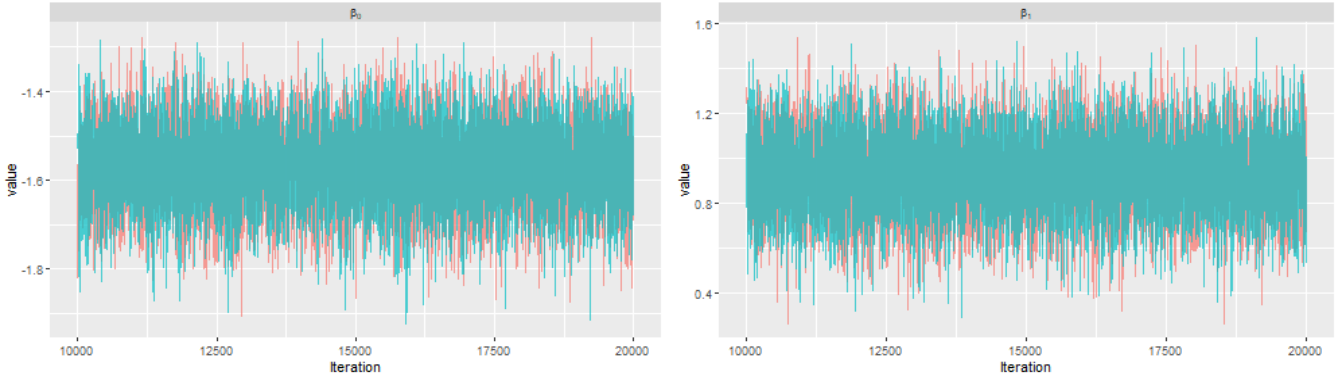


Figure 4.4: Density plots for the posterior samples under the Dependent Model.

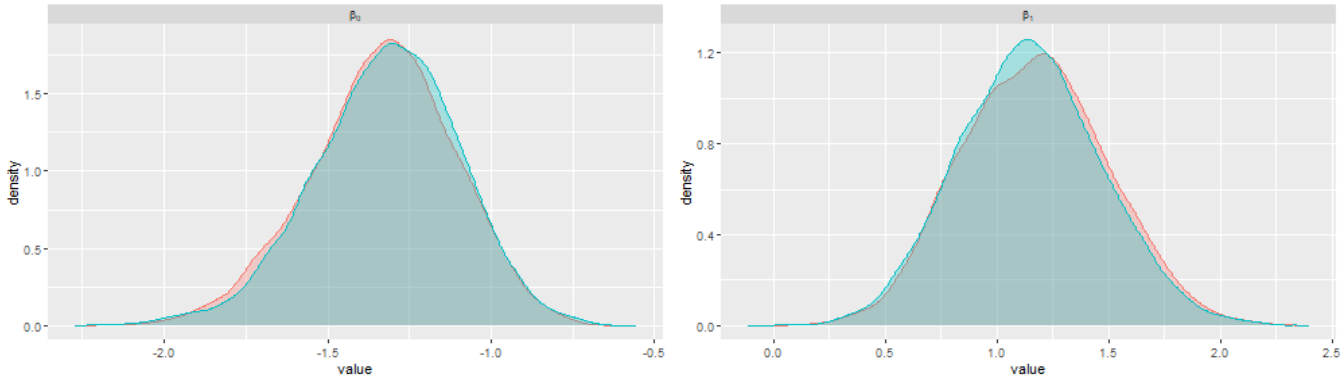


Figure 4.5: Density plots for the posterior samples under the Independent Model.

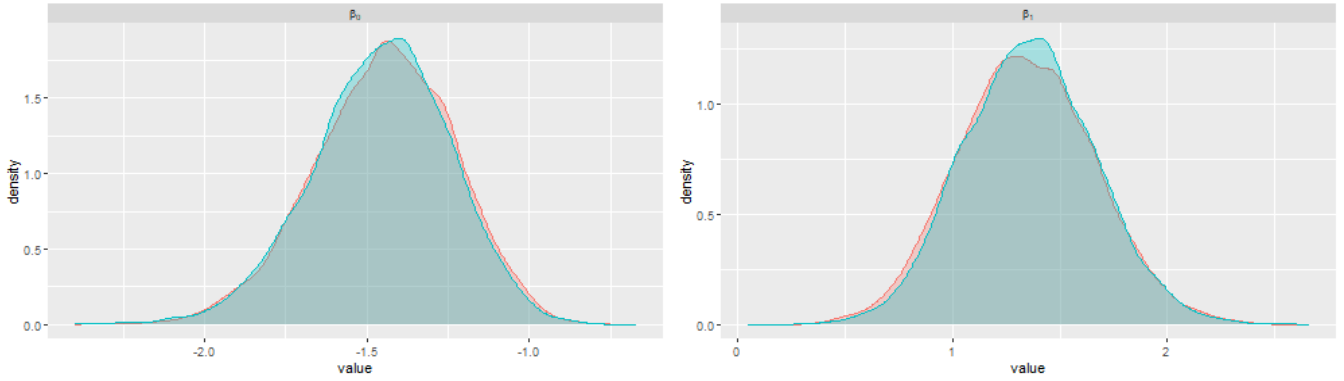


Figure 4.6: Density plots for the posterior samples under the Naive Model.

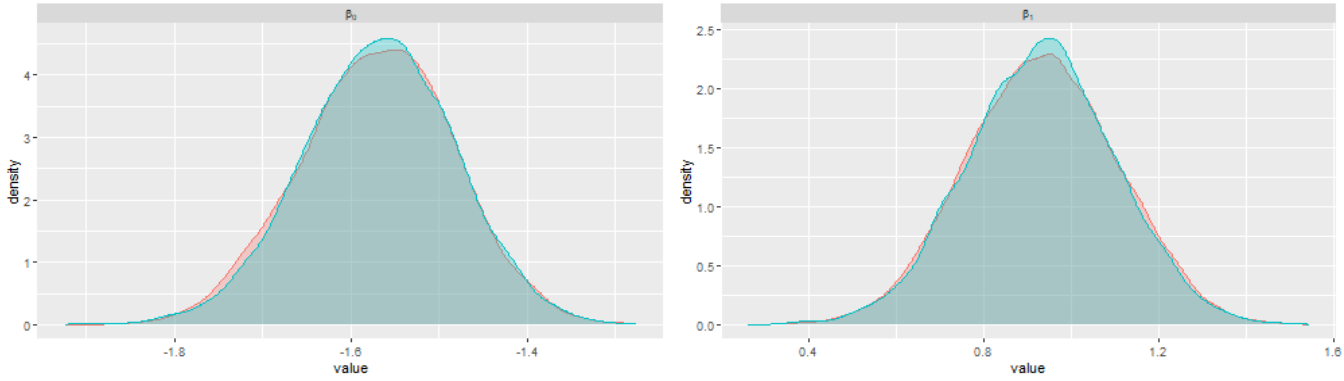


Figure 4.7: Autocorrelation plots for posterior samples of $\beta_k (k = 0, 1)$ under the Dependent Model.

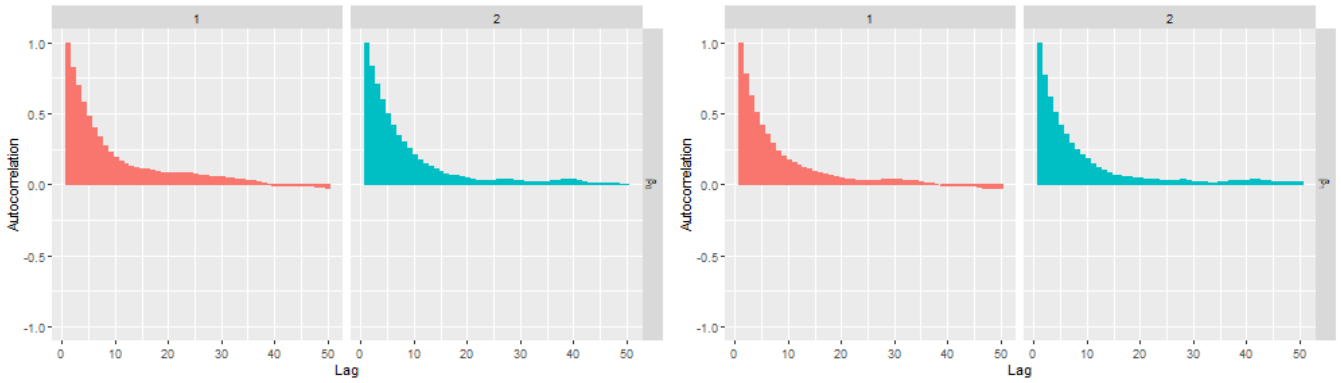


Figure 4.8: Autocorrelation plots for posterior samples of $\beta_k (k = 0, 1)$ under the Independent Model.

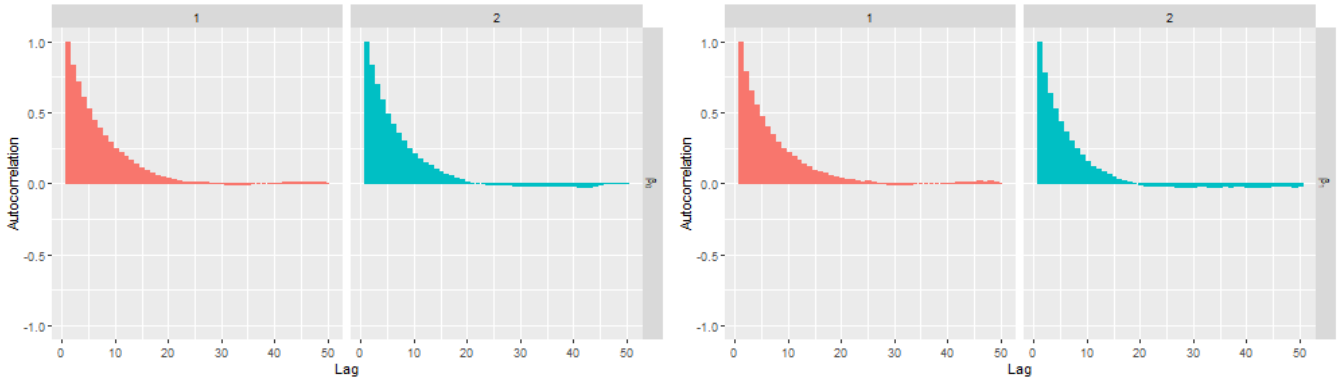
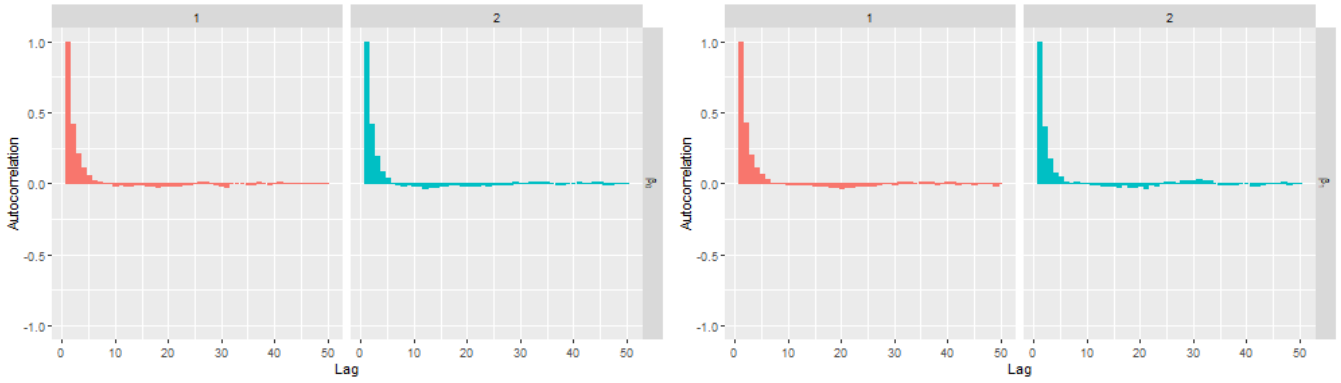


Figure 4.9: Autocorrelation plots for posterior samples of $\beta_k (k = 0, 1)$ under the Naive Model.



Chapter 5

Discussion

5.1 Findings and Conclusion

This thesis aims to establish the importance of dependence on joint misclassification errors in both the response variable and covariate. Misclassification error studies have relied on the conditional independence assumption[30, 51]. However, when information on both the response variable and covariate status are from the same source, their errors are likely dependent. First, a model that accounted for dependent misclassification errors in a binary response variable and a binary covariate were introduced. Different from the works of Tang et al.(2013) [50], Tang et al.(2015)[49], and Salway et al.(2019) [45] where dependence was captured through conditional probabilities, covariance-like parameters characterized the dependent misclassification errors in this study [53]. Although Brenner et al. (1993)[4] and Vogel et al. (2005)[53] first introduced the covariance like parameters to capture dependence, their work mostly focussed on how the bias (e.g., relative risk), due to complete ignorance of the misclassification errors, depends on the correlation in the misclassification errors when both the response variable and covariate are misclassified.

The objective of my thesis was to conduct a comprehensive simulation study to check the consequences of fitting an independent misclassification error model and a naive model to data generated from a dependent misclassification error model. The scenarios of the simulation studies were selected based on varying the misclassification parameters (SN_Y, SP_Y, SN_X and SP_X), the proportion of validation data, and the dependence strength. The dependence strength was assessed by a δ function, where δ is the expected value of the conditional covariance between the error-prone response variable and error-prone covariate (Y^* and X^*) given the error-free response variable and error-free covariate (Y and X). The

simulation studies show that misfitting the joint misclassification error model can be worse than simply ignoring misclassification errors when low proportions of the validation data are used. However, when a higher proportion of validation data is employed, the independent misclassification model's performance is similar to the dependent misclassification model, and they produce point estimates that are closer to the true value.

Categorical variables are often encountered in practice. For instance, in the medical context, the severity of a case may influence the choice of treatment better than the mere absence or presence of a disease. For example, it is more appropriate to address the diagnosis of cancer in stages (absence of cancer, I, II, III, IV). To address categorical misclassification, the models considered in chapter 2 were extended to a multi-category setting. This study is the first to address dependent misclassification in both a categorical response variable and a categorical covariate to the best of knowledge. Greenland and Kleinbaum (1983) briefly mentioned a general form for categorical misclassification in both the response variable and covariate while assuming conditional independence. To learn the impact of ignoring dependence in categorical data, simulation studies were conducted for a trinary response variable and a trinary covariate. The simulation study showed that when both low and high proportions of validation data are used in the misclassification process, the dependent misclassification error model is better than the independent misclassification error model and the naive model. However, the estimates from the dependent and independent misclassification error models are close to each other.

The proposed model presented in Chapter 2 was illustrated through a real data example by establishing the true association between Trichomoniasis and Bacterial Vaginosis, using data from the HIV Epidemiology Research Study (HERS). The data was fitted to a dependent misclassification error model, an independent misclassification error model, and a naive model. A comparison of the dependent misclassification error model and the independent misclassification error model by a Likelihood Ratio Test concluded that the dependent misclassification error model fits the dataset significantly better. This is consistent with the conclusion of Tang et al. (2013) [50], where AIC was used for the comparison of the models that accounted for dependence and models that ignored dependence.

An important implication that this study's results have on epidemiologic studies is to add

to the current body of knowledge that refutes the assertion that non-differential misclassification will always bias the effect measure towards the null. This is evident in our findings in the simulation studies that; the β_1 estimates were either greater or lower than the true values, an indication that bias is either away from the null or towards the null. In conclusion, dependence error may have an impact on joint misclassification; therefore, care has to be taken in constructing misclassification models.

5.2 Limitations

A key limitation of this study is that the validation/main study design may not always apply to real applications. This thesis considered the (internal) validation/main study design but not other study designs, e.g., multiple measurement for response variable and/or covariates when internal validation data is too expensive to collect. Internal validation data is not readily available for categorical studies; although external validation can be employed, the transportability assumption is a crucial assumption that usually cannot be verified by the available data. When validation data is not available, model identification becomes a major challenge, especially for the models with multi-categorical data considered in this study.

5.3 Future Studies

Other remaining issues can be considered for future studies; these include:

- Theoretical investigation for asymptotic bias. In this study, relative bias was employed in quantifying the impact of ignoring joint misclassification errors in both the response variable and covariate in the simulation studies. Future studies can consider the impact of joint misclassification errors on parameter estimation as the sample size increases by studying the asymptotic bias.

- Dependence misclassification errors were accounted for in our studies as covariance-like parameters. Another area to explore is to consider an alternative way to account for dependence in misclassification error. For example, Tang et al. [49] in considering

differential and dependent misclassification error, modeled dependence error through conditional probabilities.

- Alternative parameterization of the misclassification parameters. In the characterization of the misclassification process, the error-prone variable were conditioned on the error-free variable, that is in binary settings known as sensitivity and specificity. Future studies can consider the alternative characterization procedure of conditioning the error-free variable on the error-prone variable. In binary setting known as predictive values. Predictive values are also of clinical relevance in practice. Using the law of total probability and definition of conditional probability, the error-free joint probability can be connected with the error prone joint probability of the response variable and covariate.

$$P(Y=i,X=j)=\sum_{k=0}^{1}\sum_{l=0}^{1}P(Y=i,X=j|Y^*=k,X^*=l)P(Y^*=k,X^*=l). \quad (5.1)$$

Here, PPV and NPV are employed as the misclassification parameters and, the dependence parameters are defined as,

$$D_{ij}=P(Y=i,X=j|Y^*=i,X^*=j)-P(Y=i|Y^*=i)P(X=j|X^*=j). \quad (5.2)$$

The matrix formulation of this alternative parameterization is given as

$$\boldsymbol{p}=(\boldsymbol{M_Y}\bigotimes\boldsymbol{M_X}+\boldsymbol{D})\boldsymbol{p}^*, \quad (5.3)$$

where $\boldsymbol{p}^*$ and $\boldsymbol{p}$ are the vectors of the misclassified and true probabilities, respectively, that is, $\boldsymbol{p}^*=(p^*_{11},p^*_{10},p^*_{01},p^*_{00})'$ and $\boldsymbol{p}=(p_{11},p_{10},p_{01},p_{00})'$. The matrix $\boldsymbol{M_Y}$ and $\boldsymbol{M_X}$ are as follows:

$$\boldsymbol{M_Y}=\begin{bmatrix} PPV_Y & 1-NPV_Y \\ 1-PPV_Y & NPV_Y \end{bmatrix},\ \boldsymbol{M_X}=\begin{bmatrix} PPV_X & 1-NPV_X \\ 1-PPV_X & NPV_X \end{bmatrix}.$$

The dependence matrix $\boldsymbol{D}$, is composed of the dependence parameters D_{ij} and are bounded. The specific boundaries of the dependence parameter are obtained in like manner when sensitivity and specificity are employed.

References

[1] Wolfgang Ahrens and Iris Pigeot. *Handbook of epidemiology*, volume 451. Springer, 2014.

[2] Bruce A Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, pages 414–418, 1977.

[3] Gloria LA Beckles, David F Williamson, Arleen F Brown, Edward W Gregg, Andrew J Karter, Catherine Kim, R Adams Dudley, Monika M Safford, Mark R Stevens, and Theodore J Thompson. Agreement between self-reports and medical records was only fair in a cross-sectional study of performance of annual eye examinations among adults with diabetes in managed care. *Medical care*, 45(9):876–883, 2007.

[4] Hermann Brenner, David A Savitz, and Olaf Gefeller. The effects of joint misclassification of exposure and disease on epidemiologic measures of association. *Journal of clinical epidemiology*, 46(10):1195–1202, 1993.

[5] Irwin Bross. Misclassification in 2 x 2 tables. *Biometrics*, 10(4):478–486, 1954.

[6] Rebecca M Brotman, Emily J Erbelding, Roxanne M Jamshidi, Mark A Klebanoff, Jonathan M Zenilman, and Khalil G Ghanem. Findings associated with recurrence of bacterial vaginosis among adolescents attending sexually transmitted diseases clinics. *Journal of pediatric and adolescent gynecology*, 20(4):225–231, 2007.

[7] John P Buonaccorsi. *Measurement error: models, methods, and applications*. CRC press, 2010.

[8] Raymond J Carroll, Mitchell H Gail, and Jay H Lubin. Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88(421):185–199, 1993.

[9] Raymond J Carroll, David Ruppert, Ciprian M Crainiceanu, and Leonard A Stefanski. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[10] Michael Chavance, Georges Dellatolas, and Joseph Lellouch. Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. *International journal of epidemiology*, 21(3):537–546, 1992.

[11] Qixuan Chen, Hanga Galfalvy, and Naihua Duan. Effects of disease misclassification on exposure–disease association. *American journal of public health*, 103(5):e67–e73, 2013.

[12] Thomas L Cherpes, Harold C Wiesenfeld, Melissa A Melan, Jeffrey A Kant, Lisa A Cosentino, Leslie A Meyn, and Sharon L Hillier. The associations between pelvic inflammatory disease, trichomonas vaginalis infection, and positive herpes simplex virus type 2 serology. *Sexually transmitted diseases*, 33(12):747–752, 2006.

[13] John R Cook and Leonard A Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89(428):1314–1328, 1994.

[14] Karen T Copeland, Harvey Checkoway, Anthony J McMichael, and Robert H Holbrook. Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology*, 105(5):488–495, 1977.

[15] Mary Frances Cotch, II Joseph G Pastorek, Robert P Nugent, Sharon L Hillier, Ronald S Gibbs, David H Martin, David A Eschenbach, Robert Edelman, Christopher J Carey, Joan A Regan, et al. Trichomonas vaginalisassociated with low birth weight and preterm delivery. *Sexually transmitted diseases*, 24(6):353–360, 1997.

[16] Robert E Dales, Harry Zwanenburg, Richard Burnett, and Claire A Franklin. Respiratory health effects of home dampness and molds among canadian children. *American journal of epidemiology*, 134(2):196–203, 1991.

[17] Mustafa Dosemeci, Sholom Wacholder, and Jay H Lubin. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American journal of epidemiology*, 132(4):746–748, 1990.

[18] Sander Greenland. The effect of misclassification in the presence of covariates. *American journal of epidemiology*, 112(4):564–569, 1980.

[19] Sander Greenland. Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. *Journal of Statistical Planning and Inference*, 138(2):528–538, 2008.

[20] Paul Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

[21] Jerry A Hausman and Fiona Scott Morton. *Misclassification of a dependent variable in a discrete response setting*. Cambridge, Mass.: Dept. of Economics, Massachusetts Institute of Technology, 1994.

[22] Charles Hennekens, J Buring, and SL Mayrent. *Epidemiology in Medicine*. Philadelphia, PA: Lippincott Williams and Wilkins, 1987.

[23] Anne M Jurek, George Maldonado, and Sander Greenland. Adjusting for outcome misclassification: the importance of accounting for case-control sampling and other forms of outcome-related selection. *Annals of epidemiology*, 23(3):129–135, 2013.

[24] Ruth H Keogh, Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Helmut Küchenhoff, Janet A Tooze, Michael P Wallace, Victor Kipnis, et al. Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—basic theory and simple methods of adjustment. *Statistics in Medicine*, 2020.

[25] David G Kleinbaum, Lawrence L Kupper, and Hal Morgenstern. *Epidemiologic research: principles and quantitative methods*. John Wiley & Sons, 1982.

[26] Peter Kristensen. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*, pages 210–215, 1992.

[27] Helmut Küchenhoff, Samuel M Mwalili, and Emmanuel Lesaffre. A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics*, 62(1):85–96, 2006.

[28] Charles E Lance and Robert J Vandenberg. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*. Taylor & Francis, 2009.

[29] Juxin Liu, Annshirley Afful, and Yanyuan Ma. *Consequences of incorrect misclassification assumption when both response variable and covariate are misclassified*. "submitted", 2020.

[30] Juxin Liu, Paul Gustafson, and Dezheng Huo. Bayesian adjustment for the misclassification in both dependent and independent variables with application to a breast cancer study. *Statistics in medicine*, 35(23):4252–4263, 2016.

[31] Robert H Lyles. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. *Biometrics*, 58(4):1034–1036, 2002.

[32] Robert H Lyles, Li Tang, Hillary M Superak, Caroline C King, David D Celentano, Yungtai Lo, and Jack D Sobel. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology (Cambridge, Mass.)*, 22(4):589, 2011.

[33] Laurence S Magder and James P Hughes. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2):195–203, 1997.

[34] Roger J Marshall. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43(9):941–947, 1990.

[35] Thierry Mertens. Estimating the effects of misclassification. *The Lancet*, 342(8868):418–421, 1993.

[36] Mary J Morrissey and Donna Spiegelman. Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, 55(2):338–344, 1999.

[37] Samuel Musili Mwalili. Bayesian and frequentist approaches to correct for misclassification error with applications to caries research. 2006.

[38] Roberta B Ness, Kevin E Kip, Sharon L Hillier, David E Soper, Carol A Stamm, Richard L Sweet, Peter Rice, and Holly E Richter. A cluster analysis of bacterial vaginosis–associated microflora and pelvic inflammatory disease. *American journal of epidemiology*, 162(6):585–590, 2005.

[39] Sallie A Newell, Afaf Girgis, Rob W Sanson-Fisher, and Nina J Savolainen. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. *American journal of preventive medicine*, 17(3):211–229, 1999.

[40] Juha Pekkanen, Jordi Sunyer, and Susan Chinn. Nondifferential disease misclassification may bias incidence risk ratios away from the null. *Journal of clinical epidemiology*, 59(3):281–289, 2006.

[41] Donatella Pellati, Ioannis Mylonakis, Giulio Bertoloni, Cristina Fiore, Alessandra Andrisani, Guido Ambrosini, and Decio Armanini. Genital tract infections and infertility. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 140(1):3–11, 2008.

[42] M Plummer. Jags version 4.3. 0 user manual. 28 june 2017, 2017.

[43] Sujit D Rathod, Karl Krupp, Jeffrey D Klausner, Anjali Arun, Arthur L Reingold, and Purnima Madhivanan. Bacterial vaginosis and risk for trichomonas vaginalis infection: a longitudinal analysis. *Sexually transmitted diseases*, 38(9):882, 2011.

[44] FJ Rubio, Adam M Johansen, et al. A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654, 2013.

[45] Travis Salway, Martin Plöderl, Juxin Liu, and Paul Gustafson. Effects of multiple forms of information bias on estimated prevalence of suicide attempts according to sexual orientation: An application of a bayesian misclassification correction method to data from a systematic review. *American journal of epidemiology*, 188(1):239–249, 2019.

[46] MA Yushuf Sharker, Mohammed Nasser, Jaynal Abedin, Benjamin F Arnold, and Stephen P Luby. The risk of misclassifying subjects within principal component based asset index. *Emerging themes in epidemiology*, 11(1):6, 2014.

[47] Jack D Sobel. What's new in bacterial vaginosis and trichomoniasis? *Infectious disease clinics of North America*, 19(2):387–406, 2005.

[48] Tim B Swartz, Yoel Haitovsky, Albert Vexler, and Tae Y Yang. Bayesian identifiability and misclassification in multinomial data. *Canadian Journal of Statistics*, 32(3):285–302, 2004.

[49] Li Tang, Robert H Lyles, Caroline C King, David D Celentano, and Yungtai Lo. Binary regression with differentially misclassified response and exposure variables. *Statistics in medicine*, 34(9):1605–1620, 2015.

[50] Li Tang, Robert H Lyles, Ye Ye, Yungtai Lo, and Caroline C King. Extended matrix and inverse matrix methods utilizing internal validation data when both disease and exposure status are misclassified. *Epidemiologic methods*, 2(1):49–66, 2013.

[51] Mushfiqur R Tarafder, Hélène Carabin, Stephen T McGarvey, Lawrence Joseph, Ernesto Balolong Jr, and Remigio Olveda. Assessing the impact of misclassification error on an epidemiological association between two helminthic infections. *PLoS neglected tropical diseases*, 5(3):e995, 2011.

[52] Ellen Van de Poel, Ahmad Reza Hosseinpoor, Niko Speybroeck, Tom Van Ourti, and Jeanette Vega. Socioeconomic inequality in malnutrition in developing countries. *Bulletin of the World Health Organization*, 86:282–291, 2008.

[53] C Vogel, H Brenner, A Pfahlberg, and O Gefeller. The effects of joint misclassification of exposure and disease on the attributable risk. *Statistics in medicine*, 24(12):1881–1896, 2005.

[54] Chia C Wang, R Scott McClelland, Marie Reilly, Julie Overbaugh, Sandra R Emery, Kishorchandra Mandaliya, Bhavna Chohan, Jeckoniah Ndinya-Achola, Job Bwayo, and Joan K Kreiss. The effect of treatment of vaginal infections on shedding of human immunodeficiency virus type 1. *The Journal of infectious diseases*, 183(7):1017–1022, 2001.

[55] Jeffrey A Wright, Carole A Ashenburg, and Robert C Whitaker. Comparison of methods to categorize undernutrition in children. *The Journal of pediatrics*, 124(6):944–946, 1994.

[56] Grace Y Yi. *Statistical Analysis with Measurement Error Or Misclassification*. Springer, 2016.

Appendix A

Boundaries for the Dependence Parameters

The boundaries for the dependence parameters are obtained as follows;
Let $D_{ijkl} = P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)P(X^* = j|X = l)$.

$$\begin{aligned} P(Y^* = i|Y = k)P(X^* = j|X = l) + D_{ijkl} \leq & P(Y^* = i|Y = k) \\ D_{ijkl} \leq & P(Y^* = i|Y = k) \\ & - P(Y^* = i|Y = k)P(X^* = j|X = l) \\ D_{ijkl} \leq & P(Y^* = i|Y = k)(1 - P(X^* = j|X = l)) \quad . \end{aligned} \tag{A.1}$$

$$\begin{aligned} P(Y^* = i|Y = k)(1 - P(X^* = j|X = l) - D_{ijkl} \leq & P(Y^* = i|Y = k) \\ -D_{ijkl} \leq & P(Y^* = i|Y = k) \\ & - (P(Y^* = i|Y = k)(1 - P(X^* = j|X = l)) \\ D_{ijkl} \geq & - P(Y^* = i|Y = k)P(X^* = j|X = l). \end{aligned} \tag{A.2}$$

$$\begin{aligned} (1 - P(Y^* = i|Y = k))P(X^* = j|X = j) - D_{ijkl} \leq & 1 - P(Y^* = i|Y = k) \\ -D_{ijkl} \leq & 1 - P(Y^* = i|Y = k) \\ & - (1 - P(Y^* = i|Y = k))P(X^* = j|X = j) \\ D_{ijkl} \geq & -(1 - P(Y^* = i|Y = k))(1 - P(X^* = j|X = l)). \end{aligned} \tag{A.3}$$

$$\begin{aligned} (1 - P(Y^* = i|Y = k))(1 - P(X^* = j|X = l)) + D_{ijkl} \leq & 1 - P(Y^* = i|Y = k) \\ D_{ijkl} \leq & 1 - P(Y^* = i|Y = k) \\ & - (1 - P(Y^* = i|Y = k))(1 - P(X^* = j|X = l) \\ D_{ijkl} \leq & (1 - P(Y^* = i|Y = k))(P(X^* = j|X = l). \end{aligned} \tag{A.4}$$

Appendix B

Misclassified joint probabilities

Details of the misclassified joint probability p_{10}, p_{01} and p_{00} is given below:

$$\begin{aligned} p^*_{10} &= \sum_{k=0}^{1}\sum_{l=0}^{1} P(Y^*=1, X^*=0|Y=k, X=l)p_{kl} \qquad \text{(B.1)} \\ &= \Big[SN_Y(1-SN_X)+D_{11}\Big]p_{11} + \Big[SN_Y SP_X - D_{11}\Big]p_{10} \\ &\quad + \Big[(1-SP_Y)(1-SN_X)-D_{11}\Big]p_{01} + \Big[(1-SP_Y)SP_X - D_{11}\Big]p_{00} \end{aligned}$$

$$\begin{aligned} p^*_{01} &= \sum_{k=0}^{1}\sum_{l=0}^{1} P(Y^*=0, X^*=1|Y=k, X=l)p_{kl} \qquad \text{(B.2)} \\ &= \Big[(1-SN_Y)SN_X+D_{11}\Big]p_{11} + \Big[(1-SN_Y)(1-SP_X) - D_{11}\Big]p_{10} \\ &\quad + \Big[SP_Y SN_X - D_{11}\Big]p_{01} + \Big[SP_Y(1-SP_X)-D_{11}\Big]p_{00} \end{aligned}$$

$$\begin{aligned} p^*_{00} &= \sum_{k=0}^{1}\sum_{l=0}^{1} P(Y^*=1, X^*=1|Y=k, X=l)p_{kl} \qquad \text{(B.3)} \\ &= \Big[(1-SN_Y)(1-SN_X)+D_{11}\Big]p_{11} + \Big[(1-SN_Y)SP_X - D_{11}\Big]p_{10} \\ &\quad + \Big[SP_Y(1-SN_X)-D_{11}\Big]p_{01} + \Big[SP_Y SP_X - D_{11}\Big]p_{00} \end{aligned}$$

Appendix C

Proof of the delta parameter

***EstimatingCOV*$(Y^*, X^*|Y, X)$**

	$Y^* = 1$	$Y^* = 0$	
$X^* = 1$	$P(Y^* = 1, X^* = 1\|Y, X)$	$P(Y^* = 0, X^* = 1\|Y, X)$	$P(X^* = 1\|Y, X)$
$X^* = 0$	$P(Y^* = 1, X^* = 0\|Y, X)$	$P(Y^* = 0, X^* = 0\|Y, X)$	$P(X^* = 0\|Y, X)$
	$P(Y^* = 1\|Y, X)$	$P(Y^* = 0\|Y, X)$	

From the definition of covariance,

$$COV(Y^*, X^*|Y, X) = E(Y^*, X^*|Y, X) - E(Y^*|Y, X)E(X^*|Y, X)$$

For the binary case,

$$\begin{aligned} E(Y^*|Y, X) &= \sum_{y^*} y^* P(Y^* = y^*|Y, X) \\ E(X^*|Y, X) &= \sum_{x^*} x^* P(X^* = x^*|Y, X) \\ E(Y^*, X^*|Y, X) &= \sum_{x^* y^*} x^* y^* P(Y^* = y^*, X^* = y^*|Y, X) \end{aligned}$$

$$\begin{aligned} E(Y^*|Y, X) &= 1 \times P(Y^* = 1|Y, X) + 0 \times P(Y^* = 0|Y, X) = P(Y^* = 1|Y, X) \\ E(X^*|Y, X) &= 1 \times P(X^* = 1|Y, X) + 0 \times P(X^* = 0|Y, X) = P(X^* = 1|Y, X) \end{aligned}$$

$$\begin{aligned} E(Y^*, X^*|Y, X) &= [1 \times 1 \times P(Y^* = 1, X^* = 1|Y, X)] + [0 \times 1 \times P(Y^* = 0, X^* = 1|Y, X)] \\ &+ [1 \times 0 \times P(Y^* = 1, X^* = 0|Y, X)] + [0 \times 0 \times P(Y^* = 0, X^* = 0|Y, X)] \\ &= P(Y^* = 1, X^* = 1|Y, X) \end{aligned}$$

$$COV(Y^*, X^*|Y, X) = P(Y^* = 1, X^* = 1|Y, X) - P(Y^* = 1|Y, X)P(X^* = 1|Y, X)$$

Let $g(Y, X) = Cov(Y^*, X^*|Y, X)$

$$\begin{aligned} g(Y, X) &= Cov(Y^*, X^*|Y, X) \\ &= P(Y^* = 1, X^* = 1|Y, X) - P(Y^* = 1|Y, X)P(X^* = 1|Y, X) \end{aligned}$$

Considering non-differential misclassification error,

$$g(Y,X) = P(Y^* = 1, X^* = 1|Y,X) - P(Y^* = 1|Y)P(X^* = 1|X)$$

If $g(y,x)$ is a real- value function defined for all possible values of (x,y) of the discrete erandom vector (Y,X). Then $g(Y,X)$ itself is a random variable with expected value $\boldsymbol{E}g(Y,X)$ (Casella and Berger,2002) is given by

$$\boldsymbol{E}g(Y,X) = \sum_{allx,y} g(y,x)f(y,x)$$

where,

$$\sum_{allx,y} f(y,x) = P(Y,X) = 1$$

$$\begin{aligned}
\boldsymbol{E}g(Y,X) &= \Big[\sum_{allx,y} P(Y^* = 1, X^* = 1|Y,X) - P(Y^* = 1|Y)P(X^* = 1|X)\Big]P(Y,X) \\
&= \Big[P(Y^* = 1, X^* = 1|Y = 1, X = 1) - P(Y^* = 1|Y = 1)P(X^* = 1|X = 1)\Big]P(Y = 1, X = 1) \\
&+ \Big[P(Y^* = 1, X^* = 1|Y = 1, X = 0) - P(Y^* = 1|Y = 1)P(X^* = 1|X = 0)\Big]P(Y = 1, X = 0) \\
&+ \Big[P(Y^* = 1, X^* = 1|Y = 0, X = 1) - P(Y^* = 1|Y = 0)P(X^* = 1|X = 1)\Big]P(Y = 0, X = 1) \\
&+ \Big[P(Y^* = 1, X^* = 1|Y = 0, X = 0) - P(Y^* = 1|Y = 0)P(X^* = 1|X = 0)\Big]P(Y = 0, X = 0)
\end{aligned}$$

From the definition of the dependence parameter,

$$D_{kl} = P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)P(X^* = j|X = l)$$

$$\begin{aligned}
&\Big[P(Y^* = 1, X^* = 1|Y = 1, X = 1) - P(Y^* = 1|Y = 1)P(X^* = 1|X = 1)\Big]P(Y = 1, X = 1) \\
= \;& D_{11} \times p_{11} \\
= \;& \Big[P(Y^* = 1, X^* = 1|Y = 1, X = 0) - P(Y^* = 1|Y = 1)P(X^* = 1|X = 0)\Big]P(Y = 1, X = 0) \\
= \;& -D_{10} \times p_{10} \\
= \;& \Big[P(Y^* = 1, X^* = 1|Y = 0, X = 1) - P(Y^* = 1|Y = 0)P(X^* = 1|X = 1)\Big]P(Y = 0, X = 1) \\
= \;& -D_{01} \times p_{01} \\
= \;& \Big[P(Y^* = 1, X^* = 1|Y = 0, X = 0) - P(Y^* = 1|Y = 0)P(X^* = 1|X = 0)\Big]P(Y = 0, X = 1) \\
= \;& -D_{00} \times p_{00}
\end{aligned}$$

$$\boldsymbol{E}g(Y,X) = D_{11} \times p_{11} - D_{10} \times p_{10} - D_{01} \times p_{01} + D_{00} \times p_{00}$$

$$\boldsymbol{E}g(Y,X) = [D_{11}, -D_{10}, -D_{01}, D_{00}] \times \begin{bmatrix} p_{11} \\ p_{10} \\ p_{01} \\ p_{00} \end{bmatrix}$$

$$\delta = \boldsymbol{E}g(Y,X) = \sum_{ij} (-1)^{i+j} D_{ij} P_{ij} \tag{C.1}$$

Let p^* and p represent the vectors of the misclassified and true probabilities respectively,i.e.,

$$p^* = \begin{bmatrix} p^*_{11} & p^*_{10} & p^*_{01} & p^*_{00} \end{bmatrix}^T, \; p = \begin{bmatrix} p_{11} & p_{10} & p_{01} & p_{00} \end{bmatrix}^T$$

The relationship between p and p^* is given by,

$$\begin{aligned} p^* &= (\boldsymbol{Q_Y Q_X} + \boldsymbol{D})p \\ p^* &= (\boldsymbol{Q_Y Q_X})p + (\boldsymbol{D})p \end{aligned}$$

The matrix $\boldsymbol{Q_Y}$ and $\boldsymbol{Q_X}$ are composed of sensitivities and specificities in term of the response variable and the covariate and D is a matrix composed of the dependence parameters.

$$\boldsymbol{D} = \begin{bmatrix} D_{11} & -D_{10} & -D_{01} & D_{00} \\ -D_{11} & D_{10} & D_{01} & -D_{00} \\ -D_{11} & D_{10} & D_{01} & -D_{00} \\ D_{11} & -D_{10} & -D_{01} & D_{00} \end{bmatrix}$$

$$\begin{bmatrix} p^*_{11} \\ p^*_{10} \\ p^*_{01} \\ p^*_{00} \end{bmatrix} = (\boldsymbol{Q_Y Q_X})p + \begin{bmatrix} \boldsymbol{E}g(Y,X) \\ -\boldsymbol{E}g(Y,X) \\ -\boldsymbol{E}g(Y,X) \\ \boldsymbol{E}g(Y,X) \end{bmatrix}$$

Appendix D

Proof of the matrix form for Joint misclassification error model

From the law of total probability,

$$p^*_{ij} = P(Y^* = i, X^* = j) = \sum_{k=1}^{2}\sum_{l=1}^{2} P(Y^* = i, X^* = j|Y = k, X = l)p_{kl}. \tag{D.1}$$

$$\begin{aligned} p^*_{ij} &= \sum_{k=1}^{2}\sum_{l=1}^{2} P(Y^* = i, X^* = j|Y = k, X = l)p_{kl}. \\ &= \sum_{k=1}^{2}\sum_{l=1}^{2} \Big[P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)(X^* = j|X = l) + P(Y^* = i|Y = k)(X \\ &= \sum_{k=1}^{2}\sum_{l=1}^{2} \Big[P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)(X^* = j|X = l) + P(Y^* = i|Y = k)(X \end{aligned}$$

From the definition of the dependence parameter,

$$D_{ijkl} = P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)(X^* = j|X = l). \tag{D.3}$$

$$p^*_{ij} = \sum_{k=1}^{2}\sum_{l=1}^{2} \Big[P(Y^* = i|Y = k)(X^* = j|X = l) + D_{ijkl} \Big] p_{kl}. \tag{D.4}$$

Considering the RHS:

when $i = 1$, $j = 1$:

$$
\begin{aligned}
p^*_{11} = \quad & (P(Y^*=1|Y=1)(X^*=1|X=1)+D_{1111})\ p_{11} \\
& +(P(Y^*=1|Y=1)(X^*=1|X=2)+D_{1112})\ p_{12} \\
& +(P(Y^*=1|Y=1)(X^*=1|X=3)+D_{1113})\ p_{13} \\
& +(P(Y^*=1|Y=2)(X^*=1|X=1)+D_{1121})\ p_{21} \\
& +(P(Y^*=1|Y=2)(X^*=1|X=2)+D_{1122})\ p_{22} \\
& +(P(Y^*=1|Y=2)(X^*=1|X=3)+D_{1123})\ p_{23} \\
& +(P(Y^*=1|Y=3)(X^*=1|X=1)+D_{1131})\ p_{31} \\
& +(P(Y^*=1|Y=3)(X^*=1|X=2)+D_{1132})\ p_{32} \\
& +(P(Y^*=1|Y=3)(X^*=1|X=3)+D_{1133})\ p_{33}
\end{aligned}
$$

when $i=1$, $j=2$:

$$
\begin{aligned}
p^*_{12} = \quad & (P(Y^*=1|Y=1)(X^*=2|X=1)+D_{1211})\ p_{11} \\
& +(P(Y^*=1|Y=1)(X^*=2|X=2)+D_{1212})\ p_{12} \\
& +(P(Y^*=1|Y=1)(X^*=2|X=3)+D_{1213})\ p_{13} \\
& +(P(Y^*=1|Y=2)(X^*=2|X=1)+D_{1221})\ p_{21} \\
& +(P(Y^*=1|Y=2)(X^*=2|X=2)+D_{1222})\ p_{22} \\
& +(P(Y^*=1|Y=2)(X^*=2|X=3)+D_{1223})\ p_{23} \\
& +(P(Y^*=1|Y=3)(X^*=2|X=1)+D_{1231})\ p_{31} \\
& +(P(Y^*=1|Y=3)(X^*=2|X=2)+D_{1232})\ p_{32} \\
& +(P(Y^*=1|Y=3)(X^*=2|X=3)+D_{1233})\ p_{33}
\end{aligned}
$$

when $i=1$, $j=3$:

$$
\begin{aligned}
p^*_{13} = \quad & (P(Y^*=1|Y=1)(X^*=3|X=1)+D_{1311})\ p_{11} \\
& +(P(Y^*=1|Y=1)(X^*=3|X=2)+D_{1312})\ p_{12} \\
& +(P(Y^*=1|Y=1)(X^*=3|X=3)+D_{1313})\ p_{13} \\
& +(P(Y^*=1|Y=2)(X^*=3|X=1)+D_{1321})\ p_{21} \\
& +(P(Y^*=1|Y=2)(X^*=3|X=2)+D_{1322})\ p_{22} \\
& +(P(Y^*=1|Y=2)(X^*=3|X=3)+D_{1323})\ p_{23} \\
& +(P(Y^*=1|Y=3)(X^*=3|X=1)+D_{1331})\ p_{31} \\
& +(P(Y^*=1|Y=3)(X^*=3|X=2)+D_{1332})\ p_{32} \\
& +(P(Y^*=1|Y=3)(X^*=3|X=3)+D_{1333})\ p_{33}
\end{aligned}
$$

when $i=2$, $j=1$:

$$
\begin{aligned}
p^*_{21} = \quad & (P(Y^*=2|Y=1)(X^*=1|X=1)+D_{2111})\ p_{11} \\
& +(P(Y^*=2|Y=1)(X^*=1|X=2)+D_{2112})\ p_{12} \\
& +(P(Y^*=2|Y=1)(X^*=1|X=3)+D_{2113})\ p_{13} \\
& +(P(Y^*=2|Y=2)(X^*=1|X=1)+D_{2121})\ p_{21} \\
& +(P(Y^*=2|Y=2)(X^*=1|X=2)+D_{2122})\ p_{22} \\
& +(P(Y^*=2|Y=2)(X^*=1|X=3)+D_{2123})\ p_{23} \\
& +(P(Y^*=2|Y=3)(X^*=1|X=1)+D_{2131})\ p_{31} \\
& +(P(Y^*=2|Y=3)(X^*=1|X=2)+D_{2132})\ p_{32} \\
& +(P(Y^*=2|Y=3)(X^*=1|X=3)+D_{2133})\ p_{33}
\end{aligned}
$$

when $i=2$, $j=2$:

$$
\begin{aligned}
p^*_{22} = \quad & (P(Y^*=2|Y=1)(X^*=2|X=1)+D_{2211})\ p_{11} \\
& +(P(Y^*=2|Y=1)(X^*=2|X=2)+D_{2212})\ p_{12} \\
& +(P(Y^*=2|Y=1)(X^*=2|X=3)+D_{2213})\ p_{13} \\
& +(P(Y^*=2|Y=2)(X^*=2|X=1)+D_{2221})\ p_{21} \\
& +(P(Y^*=2|Y=2)(X^*=2|X=2)+D_{2222})\ p_{22} \\
& +(P(Y^*=2|Y=2)(X^*=2|X=3)+D_{2223})\ p_{23} \\
& +(P(Y^*=2|Y=3)(X^*=2|X=1)+D_{2231})\ p_{31} \\
& +(P(Y^*=2|Y=3)(X^*=2|X=2)+D_{2232})\ p_{32} \\
& +(P(Y^*=2|Y=3)(X^*=2|X=3)+D_{2233})\ p_{33}
\end{aligned}
$$

when $i=2$, $j=3$:

$$
\begin{aligned}
p^*_{23} = \quad & (P(Y^*=2|Y=1)(X^*=3|X=1)+D_{2311})\ p_{11} \\
& +(P(Y^*=2|Y=1)(X^*=3|X=2)+D_{2312})\ p_{12} \\
& +(P(Y^*=2|Y=1)(X^*=3|X=3)+D_{2313})\ p_{13} \\
& +(P(Y^*=2|Y=2)(X^*=3|X=1)+D_{2321})\ p_{21} \\
& +(P(Y^*=2|Y=2)(X^*=3|X=2)+D_{2322})\ p_{22} \\
& +(P(Y^*=2|Y=2)(X^*=3|X=3)+D_{2323})\ p_{23} \\
& +(P(Y^*=2|Y=3)(X^*=3|X=1)+D_{2331})\ p_{31} \\
& +(P(Y^*=2|Y=3)(X^*=3|X=2)+D_{2332})\ p_{32} \\
& +(P(Y^*=2|Y=3)(X^*=3|X=3)+D_{2333})\ p_{33}
\end{aligned}
$$

when $i=3$, $j=1$:

$$
\begin{aligned}
p^*_{31} = \quad & (P(Y^*=3|Y=1)(X^*=1|X=1)+D_{3111})\ p_{11} \\
& +(P(Y^*=3|Y=1)(X^*=1|X=2)+D_{3112})\ p_{12} \\
& +(P(Y^*=3|Y=1)(X^*=1|X=3)+D_{3113})\ p_{13} \\
& +(P(Y^*=3|Y=2)(X^*=1|X=1)+D_{3121})\ p_{21} \\
& +(P(Y^*=3|Y=2)(X^*=1|X=2)+D_{3122})\ p_{22} \\
& +(P(Y^*=3|Y=2)(X^*=1|X=3)+D_{3123})\ p_{23} \\
& +(P(Y^*=3|Y=3)(X^*=1|X=1)+D_{3131})\ p_{31} \\
& +(P(Y^*=3|Y=3)(X^*=1|X=2)+D_{3132})\ p_{32} \\
& +(P(Y^*=3|Y=3)(X^*=1|X=3)+D_{3133})\ p_{33}
\end{aligned}
$$

when $i=3$, $j=2$:

$$
\begin{aligned}
p^*_{32} = \quad & (P(Y^*=3|Y=1)(X^*=2|X=1)+D_{3211})\ p_{11} \\
& +(P(Y^*=3|Y=1)(X^*=2|X=2)+D_{3212})\ p_{12} \\
& +(P(Y^*=3|Y=1)(X^*=2|X=3)+D_{3213})\ p_{13} \\
& +(P(Y^*=3|Y=2)(X^*=2|X=1)+D_{3221})\ p_{21} \\
& +(P(Y^*=3|Y=2)(X^*=2|X=2)+D_{3222})\ p_{22} \\
& +(P(Y^*=3|Y=2)(X^*=2|X=3)+D_{3223})\ p_{23} \\
& +(P(Y^*=3|Y=3)(X^*=2|X=1)+D_{3231})\ p_{31} \\
& +(P(Y^*=3|Y=3)(X^*=2|X=2)+D_{3232})\ p_{32} \\
& +(P(Y^*=3|Y=3)(X^*=2|X=3)+D_{3233})\ p_{33}
\end{aligned}
$$

when $i=3$, $j=3$:

$$
\begin{aligned}
p^*_{33} = \quad & (P(Y^*=3|Y=1)(X^*=3|X=1)+D_{3311})\ p_{11} \\
& +(P(Y^*=3|Y=1)(X^*=3|X=2)+D_{3312})\ p_{12} \\
& +(P(Y^*=3|Y=1)(X^*=3|X=3)+D_{3313})\ p_{13} \\
& +(P(Y^*=3|Y=2)(X^*=3|X=1)+D_{3321})\ p_{21} \\
& +(P(Y^*=3|Y=2)(X^*=3|X=2)+D_{3322})\ p_{22} \\
& +(P(Y^*=3|Y=2)(X^*=3|X=3)+D_{3323})\ p_{23} \\
& +(P(Y^*=3|Y=3)(X^*=3|X=1)+D_{3331})\ p_{31} \\
& +(P(Y^*=3|Y=3)(X^*=3|X=2)+D_{3332})\ p_{32} \\
& +(P(Y^*=3|Y=3)(X^*=3|X=3)+D_{3333})\ p_{33}
\end{aligned}
$$

Resulting in the matrix below

$$
\boldsymbol{p}^* = (\boldsymbol{M_Y} \otimes \boldsymbol{M_X} + \boldsymbol{D})\boldsymbol{p} \tag{D.5}
$$

where

$$\begin{aligned} \boldsymbol{M_Y}[i,k] &= P(Y^* = i|Y^* = k), \\ \boldsymbol{M_X}[j,l] &= P(X^* = j|X^* = l), \end{aligned}$$

The dependence parameters $\boldsymbol{D}$ are,

$$\boldsymbol{D_{ijkl}} = P(Y^* = i, X^* = j|Y = k, X = l) - P(Y^* = i|Y = k)P(X^* = j|X = l), \qquad \text{(D.6)}$$

where i, j, k, l takes on $1, 2, 3$.

Appendix E

Parameterizing the joint probabilities p_{ij}'s through an ordinal logistic regression model

Consider a multinomial logit model with a multicategory covariate. Let Y^* denote the observed ordinal categorical response with J categories and X^* denote observed covariate with $(m \geq 2)$ categories. For X^* the following indicator variables are defined with category 1 as reference category. For $i = 2, ..., m$. Let $I_{i-1} = 1$ if $X^* = i$ and zero otherwise.

$$logit\big[P(Y^* \leq j|X^* = x^*)\big] = \alpha_j + \beta_1^* I_1 + \beta_2^* I_2 + ... + \beta_{m-1}^* I_{m-1}$$

where $j = 1, ..., J-1$

$$logit\big[P(Y^* \leq j|X^* = x^*)\big] = log\left[\frac{P(Y^* \leq j|X^* = x^*)}{1 - P(Y^* \leq j|X^* = x^*)}\right]$$

$$P(Y^* = j|X^* = x^*) = P(Y^* \leq j|X^* = x^*) - P(Y^* \leq j-1|X^* = x^*)$$

Let consider $J = 3$ and $m = 3$

	$Y^* = 2$	$Y^* = 1$	$Y^* = 0$
$X^* = 2$	$(2, 2)$	$(1, 2)$	$(0, 2)$
$X^* = 1$	$(2, 1)$	$(1, 1)$	$(0, 1)$
$X^* = 0$	$(2, 0)$	$(1, 0)$	$(0, 0)$

$$\big[P(Y^* \leq 1|X^* = x^*)\big] = \frac{exp\big[\alpha_1 + \beta_1^* I_1 + \beta_2^* I_2\big]}{1 + exp\big[\alpha_1 + \beta_1^* I_1 + \beta_2^* I_2\big]}$$

$$\big[P(Y^* \leq 2|X^* = x^*)\big] = \frac{exp\big[\alpha_2 + \beta_1^* I_1 + \beta_2^* I_2\big]}{1 + exp\big[\alpha_2 + \beta_1^* I_1 + \beta_2^* I_2\big]}$$

The cumulative probabilities reflect the ordering with $P(Y^* \leq 1|X^* = x^*) \leq P(Y^* \leq 2|X^* = x^*) \leq ... \leq P(Y^* \leq J)$.

The estimated probabilty for the first category is given by:

$$P(Y^* = 1|X^* = x^*) = P(Y^* \leq 1|X^* = x^*)$$

Other category probabilities are obtained from the difference between two consecutive cumulative probabilities. For example,

$$P(Y^* = 2|X^* = x^*) = P(Y^* \leq 2|X^* = x^*) - P(Y^* \leq 1|X^* = x^*)$$

The final cumulative probability is necessarily equals to 1.

$$P(Y^* \leq 3|X^* = x^*) = 1$$

$$P(Y^* = 3|X^* = x^*) = 1 - P(Y \leq 2|X^* = x^*)$$

The cell probabilities can be obtained from:

$$
\begin{aligned}
P(Y^*=1, X^*=1) &= \left[\frac{exp\big(\alpha_1\big)}{1+exp\big(\alpha_1\big)}\right] \times P(X^*=1) \\
P(Y^*=1, X^*=2) &= \left[\frac{exp\big(\alpha_1+\beta_1^*\big)}{1+exp\big(\alpha_1+\beta_1^*\big)}\right] \times P(X^*=2) \\
P(Y^*=1, X^*=3) &= \left[\frac{exp\big(\alpha_1+\beta_2^*\big)}{1+exp\big(\alpha_1+\beta_2^*\big)}\right] \times P(X^*=3) \\
P(Y^*=2, X^*=1) &= \left[\frac{exp\big(\alpha_2\big)}{1+exp\big(\alpha_2\big)} - \frac{exp\big(\alpha_1\big)}{1+exp\big(\alpha_1\big)}\right] \times P(X^*=1) \\
&= \left[\frac{exp\big(\alpha_2\big)}{1+exp\big(\alpha_2\big)} - P(Y^*=1|X^*=1)\right] \times P(X^*=1) \\
P(Y^*=2, X^*=2) &= \left[\frac{exp\big(\alpha_2+\beta_1^*\big)}{1+exp\big(\alpha_2+\beta_1^*\big)} - \frac{exp\big(\alpha_1+\beta_1^*\big)}{1+exp\big(\alpha_1+\beta_1^*\big)}\right] \times P(X^*=2) \\
&= \left[\frac{exp\big(\alpha_2+\beta_1^*\big)}{1+exp\big(\alpha_2+\beta_1^*\big)} - P(Y^*=1|X^*=2)\right] \times P(X^*=2) \\
P(Y^*=2, X^*=3) &= \left[\frac{exp\big(\alpha_2+\beta_2^*\big)}{1+exp\big(\alpha_2+\beta_2^*\big)} - \frac{exp\big(\alpha_1+\beta_2^*\big)}{1+exp\big(\alpha_1+\beta_2^*\big)}\right] \times P(X^*=3) \\
&= \left[\frac{exp\big(\alpha_2+\beta_2^*\big)}{1+exp\big(\alpha_2+\beta_2^*\big)} - P(Y^*=1|X^*=3)\right] \times P(X^*=3) \\
P(Y^*=3, X^*=1) &= \left[1 - \frac{exp\big(\alpha_2\big)}{1+exp\big(\alpha_2\big)}\right] \times P(X^*=1) \\
&= \left[1 - P(Y^* \leq 2|X^*=1)\right] \times P(X^*=1) \\
P(Y^*=3, X^*=2) &= \left[1 - \frac{exp\big(\alpha_2+\beta_1^*\big)}{1+exp\big(\alpha_2+\beta_1^*\big)}\right] \times P(X^*=2) \\
&= \left[1 - P(Y^* \leq 2|X^*=2)\right] \times P(X^*=2) \\
P(Y^*=3, X^*=3) &= \left[1 - \frac{exp\big(\alpha_2+\beta_2^*\big)}{1+exp\big(\alpha_2+\beta_2^*\big)}\right] \times P(X^*=3) \\
&= \left[1 - P(Y^* \leq 2|X^*=3)\right] \times P(X^*=3)
\end{aligned}
$$