Bias Study of the Naive Estimator in a Longitudinal Binary Mixed-effects Model with Measurement Error and Misclassification in Covariates

by

©Ernest Dankwa

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirement for the Degree of

Master of Science

Department of Mathematics and Statistics

Memorial University

St. John's Newfoundland and Labrador, Canada August 2014

Abstract

When covariates in Longitudinal data are subject to errors, the naive estimates of the model parameters are often biased. In this research, we exploit a Dynamic Binary Mixed-effects Model using a Generalized Quasi-likelihood approach. Through simulations, we shall study the patterns in the bias of the naive estimator of the parameters that ignores the errors in the covariates.

Acknowledgements

First and foremost, I thank the Almighty God for watching over me over all these years of study. Next, my admiration and profound gratitude go to my supervisor, Dr. Taraneh Abarin. I am deeply indebted to her for all the support, inspiration, guidance and, most importantly, her commitment to excellence. They were very essential not only for the success of this work but also for my entire progress and well being. I am also grateful to all the faculty members and staff of Mathematics and statistics department for creating an enabling environment for us to study. I appreciate all the advice, teachings and invaluable services you offered to ensure that we have comfortable study environment. Finally, to my family, friends and colleagues, I will say that your comments and suggestions were fantastic! I thank everyone who has contributed in diverse ways to make my period of study successful.

Dedication

This work is specially dedicated to my mother, Mary Ashyia;

and to the memory of my father S.M. Dankwa,

in appreciation of their parental love and sacrifices.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Generalized Linear Models for Longitudinal Studies

Over the past 30 years, various statisticians and researchers have made significant contributions towards development of methodologies for longitudinal data analysis. According to Fitzmaurice, Davadian, Verberke and Molenberghs (2009), the earliest models for longitudinal data analysis were based on univariate repeated measure ANOVA. However, those models were very restrictive due to compound symmetry assumption of the correlation structure. Some researchers developed the repeated measure MANOVA models, to flex this restrictive assumption. However, those models were computationally demanding and could not support features of the longitudinal data such as unbalanced and incomplete data, time varying covariates and discrete responses. In the early 1980's, the development of linear mixed-effect models (Laird and Ware (1982)) became very essential in longitudinal data analysis. However, they were not appropriate for discrete responses. Generalized Linear Models (GLMs), (Nelder and Wedderburn (1972), McCullagh and Nelder (1983)) were developed to cater for discrete responses in longitudinal data analysis. GLM extensions include Marginal Models (models in which the response depends only on the covariates), Mixed-effect models (models in which the response depends on both covariates and random effects), and Transitions models (models in which response depends on past responses), (Diggle, Heagety ,Liang, and Zeger (2002)).

1.2 Mixed Logit and Probit link functions in GLMMs

In most statistical literature, Generalized linear mixed-effect models, GLMMs are written as

$h^{-1}[E\,(Y_{it}|X_{it}, Z_{it}, b_i)] = X'_{it}\beta + Z'_{it}b_i, \;\; i = 1 \dots n, \; t = 1 \dots T;$

where for n number of individuals,

- $h^{-1}(\cdot)$ is a non-linear link function;

- Y_{it} is response of the ith individual at time point t;
- X_{it} is covariate for fixed effects with parameter β;
- Z_{it} is covariate for random effects and
- b_i are individual specific random effects.

The responses, which are collected from a number of individuals along a set of covariates over a given period of time, are assumed to be conditionally independent and follow the exponential family distribution. The link function used in GLMMs depends on the type of the response under study; it may be identity link for continuous response, log link for counts, logit or probit link for binary response.

For n number of individuals and over T time-points, the logit link is defined as follows:

$logit\left(\mu_{it}\right) = log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \eta_{it}, \;\; i = 1 \ldots n, \; t = 1 \ldots T.$

On the other hand, the probit link is defined as follows:

$\phi^{-1}\left(\mu_{it}\right) = \eta_{it}, \quad i = 1 \ldots n, \; t = 1 \ldots T,$

In each case, $\mu_{it} = E\left(Y_{it} | X_{it}, Z_{it}, b_i\right)$ is the mean response,

$\eta_{ij} = X'_{it}\beta + Z'_{it}b_i$ is a linear predictor, and

ϕ^{-1} is the inverse standard normal cumulative distribution function.

1.3 Some review of literature on Measurement error and Misclassification

There exist a great number of literature on measurement error (ME) and misclassification. Examples are Gustafson (2003), Buzas, Jeffrey , Leonard , Stefanski, and Tosteson (2014). It is well-known that ME and/or misclassification have negative impacts on the estimating parameters. (Fuller (1986), Carroll and Stefanski (1985), Abarin, Li, Wang, and Briollais (2012), Abarin and Wang (2012)). As a result of that, naive estimators that ignore the errors in covariates are not typically consistent. However, the directions and the magnitude of the bias can be quite complex, as the naive estimate is a function of the unknown model parameters. The magnitude of the bias is treated in various cases: it can be ignored (Buzas, Jeffrey, Leonard, Stefanski, and Tosteson (2014)), reduced (Eisenhower, Mathiowetz and Morganstein (1991)), or corrected (Batistatou and McNamee (2012), Spiegelman, McDermott, and Rosner (1997)).

Greene and Cai (2004) discussed measurement error in marginal model for multivariate failure time data. They demonstrated the properties of Simulation Extrapolation (SIMEX), (Cook and Stefanki (1994)) to account for ME in covariate in the marginal hazard model. Buonaccorsi (2010) and Carroll and Stefanski (1985), are among several works that exploit the SIMEX to account for Measurement errors. Tao and Fan (2011) investigated the generalized quasi-likelihood estimating approach for misclassified binary data. In the study, only the response variable was subject to misclassification. Ji and Fan (2010) used a similar approach for misclassified binary response model comparing results to those obtained from maximum likelihood estimation. Thomas, Stefanski, Davidian (2011) also explored the moment adjusted imputation to account for measurement error in logistic regression model. Fan, Sutradhar and Rao (2012) have proposed bias-corrected generalized method of moment and generalized quasi-likelihood for panel data with measurement error.

1.4 Modelling Strategies

In our research work, we focus on longitudinal binary mixed-effect model with measurement error and misclassification in covariates. We assume the responses are not subject to misclassification. We also assume responses are influence by individual random effects and conditioning on the individual random effect, there is a dynamic linear relationship between successive responses. We make use of generalized quasi-likelihood (GQL) estimation approach in our study. We study the bias patterns as parameters in the model change with the others held constant.

1.5 Organization of work

This thesis is not meant to be a comprehensive treatment of measurement error analyses. It is laid out to give readers useful insights of bias patterns when covariates are subject to measurement error and misclassification and hopefully to motivate bias corrected inferences. The structure of our work is as follows:

In Chapter 2, we present the error-free model which is a longitudinal binary mixed-effect model without errors in covariates. We explain the derivation of moments and the estimation process. In the first part of Chapter 3, we describe the theoretical work for a model with measurement error. This is followed by results of simulation studies of various scenarios designed to help us study bias pattern of the naive estimator, which ignores the errors in the covariates. Chapter 4 covers the model with misclassification in covariates, both the theoretical work and simulation results. A final model is presented in Chapter 5, which has both measurement error and misclassification in one model. Discussion of the entire research work is presented in Chapter 6.

Chapter 2

Longitudinal Binary Mixed-effects Model

In Longitudinal studies, responses collected over time are usually correlated since they are repeated measurements taken on the same individuals. Quite often, it is of scientific interest to determine the influence of individual specific random effect on the response under consideration.

We consider a Generalized Longitudinal Binary Mixed-effect Model in which, conditional on individual specific effect, γ_i, successive responses have a linear dynamic relationship. The model is given by:

$$P(Y_{it} = 1|\gamma_i, Y_{it-1}) = \pi_{it} + \rho(Y_{it-1} - \pi_{it-1}), \quad t = 2, \ldots, T \tag{2.1}$$

$$P(Y_{i1} = 1|X_{i1}, G_i, \gamma_i) = \pi_{i1}, \tag{2.2}$$

where,

$$\pi_{it} = \frac{exp(X'_{it}\beta + \gamma_i + G_i\alpha)}{1 + exp(X'_{it}\beta + \gamma_i + G_i\alpha)}, \quad t = 1, \ldots, T.$$

We define the variables in the model as follows:

- Y_{it} is the binary response of the ith individual, at time point t, $\quad t = 1, \ldots, T$.
- $X_i = [X_{i1} \ldots X_{it} \ldots X_{ip}]'$ is a matrix of continuous covariates;
- $G_i \in I\!R$ is a categorical time-invariant predictor
- $X_{it} \in I\!R^p$ is the p dimensional continuous covariate, independent of G_i
- $\beta \in I\!R^p$ is a vector of fixed effects parameter, and $\alpha \in I\!R$ is the coefficient of the variable G_i.
- $\gamma_i \in I\!R$ is the individual specific random effect; we assume it is normally distributed with mean zero and constant variance σ^2_γ.

With these model specifications, the conditional moments are obtained from the properties of the binary distribution as follows:

$$
\begin{aligned}
E(Y_{it}|X_{it}, G_i, \gamma_i) &= \pi_{it} && (2.3)\\
Var(Y_{it}|X_{it}, G_i, \gamma_i) &= \sigma_{itt} = \pi_{it}(1-\pi_{it})\\
Cov(Y_{iu}, Y_{it}|X_{it}, G_i, \gamma_i) &= \rho^{|t-u|}\sigma_{iuu} = \rho^{|t-u|}\pi_{iu}(1-\pi_{iu}), \quad \text{(Sutradhar (2011))}\\
Corr(Y_{iu}, Y_{it}|X_{it}, G_i, \gamma_i) &= \begin{cases} \rho^{t-u}\left[\frac{\sigma_{iuu}}{\sigma_{itt}}\right]^{1/2}, & u \le t \\ \\ \rho^{u-t}\left[\frac{\sigma_{itt}}{\sigma_{iuu}}\right]^{1/2}, & u > t. \end{cases} && (2.4)
\end{aligned}
$$

Assuming $f(\gamma_i)$ to be the probability density function of γ_i, we have the marginal moments as follows:

$$
\begin{aligned}
\mu_{it} = E(Y_{it}|X_{it}, G_i) &= E_{\gamma_i}(E(Y_{it}|X_{it}, G_i, \gamma_i)) = \int \pi_{it} f(\gamma_i) d\gamma_i\\
Var(Y_{it}|X_{it}, G_i) &= E_{\gamma_i}(Var(Y_{it}|X_{it}, G_i, \gamma_i)) + Var_{\gamma_i}(E(Y_{it}|X_{it}, G_i, \gamma_i)) = \mu_{it}(1-\mu_{it})
\end{aligned}
$$

In Appendix A, it has been shown that both the marginal and the conditional response have binary distribution.

Now, when $u < t$,

$$\begin{aligned}
Cov(Y_{iu}, Y_{it}|X_{it}, G_i) &= E_{\gamma_i}\left(Cov(Y_{iu}, Y_{it}|X_{it}, G_i, \gamma_i)\right) \\
&+ Cov_{\gamma_i}(E(Y_{iu}|X_{it}, G_i, \gamma_i), E(Y_{it}|X_{it}, G_i, \gamma_i)) \\
&= E_{\gamma_i}(\rho^{t-u}\sigma_{iuu}) + Cov_{\gamma_i}(\pi_{iu}, \pi_{it}) \\
&= \rho^{t-u}\left[\int \pi_{iu}(1-\pi_{iu})f(\gamma_i)d\gamma_i\right] \\
&+ \int \pi_{iu}\pi_{it}f(\gamma_i)d\gamma_i - \mu_{iu}\mu_{it} \\
&= \rho^{t-u}\left[\int \left[\pi_{iu} - (\pi_{iu})^2\right] f(\gamma_i)d\gamma_i\right] \\
&+ \int \pi_{iu}\pi_{it}f(\gamma_i)d\gamma_i - \mu_{iu}\mu_{it} \\
&= \rho^{t-u}\left[\mu_{iu} - \mu_{iuu}\right] + \left[\mu_{iut} - \mu_{iu}\mu_{it}\right],
\end{aligned}$$

where,

$$\begin{aligned}
\mu_{iuu} &= E_{\gamma_i}\left[(\pi_{iu})^2\right] = \int (\pi_{iu})^2 f\left(\gamma_i\right) d\gamma_i. \\
\mu_{iut} &= E_{\gamma_i}\left[\pi_{iu}\pi_{it}\right] = \int \pi_{iu}\pi_{it}f(\gamma_i)d\gamma_i.
\end{aligned}$$

More details on obtaining μ_{it}, μ_{iut} and μ_{iuu} are given in Appendix B

2.1 Estimation Method: Generalized Quasi-likelihood

To estimate the parameters in model (2.1), various techniques have been proposed by different authors, such as Generalized Estimating Equations (GEE) (Liang and Zeger (1986)), Marginal Quasi-likelihood (MQL), and Penalized Quasi-likelihood (PQL) methods (Goldstein (1991); Breslow and Clayton, (1993)). In this dissertation, we apply the Generalized Quasi-likelihood (GQL) method. This methodology, which was first suggested by Wedderburn (1972), requires only the mean and variance of the response to be specified, and shown to have asymptotic normal distribution. It has also been shown to be more efficient (Sutradhar (2011)), relative to some of the methods mentioned above.

We let $Y_i = (Y_{i1} \dots Y_{iT})'$ to be the $T \times 1$ vector of the observed longitudinal binary responses, for $(i = 1 \dots n)$;
$E(Y_{it}|X_{it}, G_i) = \mu_{it} = (\mu_{i1} \dots \mu_{iT})'$ be the marginal mean of response variable;
$Cov(Y_i|X_i, G_i) = \Omega_i$ be the $T \times T$ covariance matrix of response variable, and $\theta = (\beta', \alpha)'$ to be the vector of parameters of interest. Then, the marginal GQL estimating equation for the parameters, is given by the following equations.

$$\sum_{i=1}^{n} \frac{\partial \mu_i'}{\partial \theta} \Omega_i^{-1} (Y_i - \mu_i) = 0,$$

where $\frac{\partial \mu_i'}{\partial \theta}$ is a $(p+1) \times T$ matrix of first derivative of the marginal mean response.

We note that under the normality assumption for γ_i, the marginal mean response is of the form

$$\begin{aligned}\mu_{it} &= \int_{-\infty}^{\infty} \pi_{it} f(\gamma_i) d(\gamma_i) \\ &= \int_{-\infty}^{\infty} \frac{\exp(X_{it}'\beta + \gamma_i + G_i\alpha)}{1 + \exp(X_{it}'\beta + \gamma_i + G_i\alpha)} \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} e^{-\frac{\gamma_i^2}{2\sigma_\gamma^2}} d(\gamma_i)\end{aligned}$$

The above integral cannot be evaluated in closed-form. However, an approximation is suggested by some authors such as (Fitzmaurice, Davadian, Verberke, Molenberghs (2009)), (Diggle, Heagety ,Liang, and Zeger (2002)), and Monahan and Stefanski (1992).

That is, for a logistic regression model with a single random intercept γ_i, where we have $logit[E(Y_{it}|X_{it}, \gamma_i)] = X_{it}'\beta + \gamma_i$ and $\gamma_i \sim N(0, \sigma_\gamma^2)$, the following approximation holds:

$$logit[E(Y_{it}|X_{it})] \approx \frac{X_{it}'\beta}{\sqrt{1 + \sigma_\gamma^2/k^2}},$$

where $k^2 = 1.7$. Hence,

$$E(Y_{it}|X_{it}) \approx \frac{\exp\left(\frac{X_{it}'\beta}{\sqrt{1+\sigma_\gamma^2/k^2}}\right)}{1 + \exp\left(\frac{X_{it}'\beta}{\sqrt{1+\sigma_\gamma^2/k^2}}\right)}.$$

Based on the above approximation, we have the marginal mean of the response for our model, as follows.

$$E(Y_{it}|X_{it},G_i) = \mu_{it} \approx \frac{\exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}}\right)}{1+exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}}\right)}. \quad (2.5)$$

Therefore, $\frac{\partial\mu_{it}}{\partial\theta'} = (\frac{\partial\mu_{it}}{\partial\beta}, \frac{\partial\mu_{it}}{\partial\alpha})'$, where

$$\frac{\partial\mu_{it}}{\partial\beta} \approx \frac{\exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}}\right)\frac{X_{it}\mathbf{1}_p}{\sqrt{1+\sigma^2_\gamma/k^2}}}{\left[1+\exp(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})\right]^2},$$

and

$$\frac{\partial\mu_{it}}{\partial\alpha} \approx \frac{\exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}}\right)\frac{G_i}{\sqrt{1+\sigma^2_\gamma/k^2}}}{\left[1+\exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}}\right)\right]^2}.$$

$\mathbf{1}_p$ is a unit vector of length p. The solution to this equation may be obtained by solving Gauss-Newton iterative equation given by

$$\hat{\theta}_{GQL}(r+1) = \hat{\theta}_{GQL}(r) + \left[\sum_{i=1}^{k}\frac{\partial\mu'_i}{\partial\theta}\Omega_i^{-1}\frac{\partial\mu_i}{\partial\theta}\right]^{-1}\left[\sum_{i=1}^{k}\frac{\partial\mu'_i}{\partial\theta}\Omega_i^{-1}(Y_i-\mu_i)\right]_r.$$

The GQL estimator is asymptotically normal with mean θ, and the following covariance matrix (Sutradhar (2011)),

$$\left[\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\theta'}\Omega_i^{-1}\frac{\partial\mu_i}{\partial\theta}\right]^{-1},$$

where, $\frac{\partial\mu_i}{\partial\theta'} = (\frac{\partial\mu_{i1}}{\partial\theta}, \frac{\partial\mu_{i2}}{\partial\theta}, \ldots, \frac{\partial\mu_{iT}}{\partial\theta})'$.

Chapter 3

Simulation Studies on Longitudinal Binary Mixed-effects Model with Measurement Error

In practice, the true predictor X_{it} is often unobservable; instead, another variable W_{it} independent of G_i is observed, which is prone to some measurement error, U_{it}. We consider a Classical (additive) measurement error model as

$$w_{itv} = x_{itv} + u_{itv}, \quad v = 1 \ldots p.$$

Here, u_{itv} is an unobservable measurement error with mean 0 and variance σ_v^2. One can write the measurement error model as:

$$W_i = X_i + U_i,$$

where,

$W_i = [w_{i(1)} \dots w_{i(v)} \dots w_{i(p)}]$ with $w_{i(v)} = (w_{i1v} \dots w_{itv} \dots w_{iTv})'$ and

$U_i = [U_{i(1)} \dots U_{i(v)} \dots U_{i(p)}]$ with $u_{i(v)} = (u_{i1v} \dots u_{itv} \dots u_{iTv})'$.

We assume that U_i, is independent of W_i. Moreover, the measurement errors, for any two covariates are independent. Thus,

$$E[U_i U_i'] = T diag[\sigma_1^2, \dots, \sigma_2^2, \dots, \sigma_p^2]. \tag{3.1}$$

Furthermore, we assume that measurement errors $u_{i1v} \dots u_{itv} \dots u_{iTv}$, on the same vth covariate, at different time points are likely to be correlated due to common instrumental random effect, hence

$$E(U_{i(v)} U_{i(v)}') = \sigma_v^2[\phi_u \mathbf{1}_T \mathbf{1}_T' + (1 - \phi_u)\mathbf{I}_T],$$

where ϕ_u denotes the equicorrelation between u_{iv} *and* u_{it} and $\mathbf{I}_T$ is the $T \times T$ identity matrix. Since in here, we do not observe X_{it}, the marginal moments of the response can be written in terms of the observed covariate W_{it}.

Therefore, by the model assumptions and the law of iterative expectations, we have:

$$
\begin{aligned}
& E(Y_{it}|W_{it}, G_i) \\
= \quad & E_{X|W}\left(E(Y_{it}|X_{it}, W_{it}, G_i)|W_{it}, G_i\right) \\
= \quad & E_{X|W}(E(Y_{it}|X_{it}, G_i)|W_{it}, G_i) \qquad (3.2) \\
= \quad & E_{X|W}(\mu_{it}|W_{it}, G_i) \\
\approx \quad & E_{X|W}\left(\frac{\exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}}\right)}{1+exp\left(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_{\gamma}/k^2}}\right)}, |W_{it}, G_i\right), \quad k^2 = 1.7 \qquad (3.3) \\
\approx \quad & E_{X|W}(\mu_{it}|W_{it}, G_i) \qquad (3.4) \\
\approx \quad & \mu^*_{1it} \qquad (3.5)
\end{aligned}
$$

Equation (3.2) is obtained by assuming that W is surrogate, meaning that it can not provide any additional information about the distribution of the response when X is given. Moreover, equation (3.3) is obtained by substituting the marginal mean approximating in (2.5). Moreover, further information on equation (3.4) is given in Appendix (C).

The marginal variance and covariance of the response can be obtained in a similar fashion as follows:

$$
\begin{aligned}
& var(Y_{it}|W_{it}, G_i) \\
= \;& var_{X|W}(E(Y_{it}|X_{it}, W_{it}, G_i)|W_{it}, G_i) + E_{X|W}(var(Y_{it}|X_{it}, W_{it}, G_i)|W_{it}, G_i) \\
= \;& var_{X|W}(E(Y_{it}|X_{it}, G_i)|W_{it}, G_i) + E_{X|W}(var(Y_{it}|X_{it}, G_i)|W_{it}, G_i), && (3.6) \\
\approx \;& E_{X|W}(\mu_{it}|W_{it}, G_i)(1 - E_{X|W}(\mu_{it}|W_{it}, G_i)), && (3.7) \\
\approx \;& \mu^*_{1it}(1 - \mu^*_{1it}) && (3.8)
\end{aligned}
$$

where equation (3.6) results from the surrogacy assumption on W. Also, equation (3.7) is obtained by the marginal properties of binary distribution, as illustrated in Appendix A.

Moreover, for $t > u$

$$
\begin{aligned}
& cov(Y_{it}, Y_{iu}|W_{it}, G_i) \\
= \;& cov_{X|W}(E((Y_{it}|X_{it}, W_{it}, G_i)|W_{it}, G_i), E((Y_{iu}|X_{iu}, W_{iu}, G_i)|W_{iu}, G_i)) \\
+ \;& E_{X|W}(cov((Y_{it}|X_{it}, W_{it}, G_i)|W_{it}, G_i, (Y_{iu}|X_{iu}, W_{iu}, G_i)|W_{iu}, G_i)) \\
= \;& cov_{X|W}(E((Y_{it}|X_{it}, G_i)|W_{it}, G_i), E((Y_{iu}|X_{iu}, G_i)|W_{iu}, G_i)) && (3.9) \\
+ \;& E_{X|W}(cov((Y_{it}|X_{it}, G_i)|W_{it}, G_i), ((Y_{iu}|X_{iu}, G_i)|W_{iu}, G_i)) \\
= \;& cov_{X|W}(\mu_{iu}, \mu_{it}|W_{it}, W_{iu}, G_i) + E_{X|W}(\rho^{t-u}[\mu_{iu} - \mu_{iuu}] + [\mu_{iut} - \mu_{iu}\mu_{it}]).
\end{aligned}
$$

$$(3.10)$$

where equation (3.9) results from the surrogacy assumption on W.

In order to estimate the parameters of the model with measurement error, we need to solve the GQL estimating equations. The *naive* estimator of θ that ignores the measurement error in X, is determined as follows.

$$\sum_{i=1}^{n} \frac{\partial \mu_{1i}^{*}}{\partial \theta'} \Omega_{1i}^{-1} (y_i - \mu_{1i}^{*}) = 0, \tag{3.11}$$

where $\mu_{1i}^{*} = (\mu_{1i1}^{*} \cdots \mu_{1iT}^{*})'$ is the $T \times 1$ vector of marginal mean of response variable based on W and G, $(i = 1 \ldots n)$;

$Cov(Y_i|W_i, G) = \Omega_{1i}^{*}$ is the $T \times T$ covariance matrix of response variable,

$\frac{\partial \mu_{1i}^{*}}{\partial \theta'}$ is a $(p+1) \times T$ vector of first derivative of marginal mean response based on W and G.

$\frac{\partial \mu_{1i}^{*}}{\partial \theta'} = (\frac{\partial \mu_{1i1}^{*}}{\partial \theta}, \frac{\partial \mu_{1i2}^{*}}{\partial \theta}, \ldots, \frac{\partial \mu_{1iT}^{*}}{\partial \theta})'$.

3.1 Simulation Studies

It is well-known that measurement error often affects both bias and variability of the parameter estimators (Fuller (1986), Carroll and Stefanski (2006)). Naive estimators are typically inconsistent. However, the directions and the magnitude of the bias can be quite complex. In the last chapter, we provided the GQL estimating equations. It should be noted that for the estimates, it was assumed that the covariance matrix of the response is known. In application, this assumption is unrealistic, as the covariance matrix is a function of unknown model parameters. As a result, the bias in the naive estimators change with the change in the model parameters. In this section, using simulation studies, we examine the direction and magnitude of the naive estimators, as a function of the model parameters.

For every simulation scenario, we provide the set-ups and the results, separately. Here, we first present the common set-ups for all scenarios of this chapter. For simplicity, we considered two covariates in the model;

For $T = 4$ time points, we generated one-dimensional true continuous time-dependent predictors as follows:

$X_{i1} \sim U(0, 0.2)$,

$X_{i2} \sim U(0.1, 0.3)$,

$X_{i3} \sim U(0.2, 0.4)$,

$X_{i4} \sim U(0.3, 0.5)$.

The true categorical time-invariant covariate, G_i generated from a binary distribution with probability of success, $P(G = 1) = \mathbf{p}$ where $\mathbf{p} = 0.3$. The individual specific random-effect, γ_i, is generated from standard normal distribution with mean 0 and variance $\sigma_\gamma^2 = 1$. The lag correlation coefficient, ρ, is highly restricted due to model assumptions, it is therefore fixed at $\rho = 0.1$. We consider $n = 300$ individuals. The regression model parameters were set to be $\alpha = 0.5$ and $\beta = 1$, except in the scenarios that they changed.

We consider an additive measurement error model and assume that measurement errors between two different covariates are independent. Then we fix the variance of measurement error variable at $\sigma_u^2 = 2$. Because of the complexity of the model that requires several iterative steps and approximations, the simulation studies were extremely time consuming. Therefore, for each sample size, 100 Monte Carlo replicates were simulated and the Monte-Carlo mean estimates and standard errors of the estimators were computed. All computations were done in R version 3.01. We exploit the Grid search method with 25 Grid points to solve the minimization problem associated with the Generalized Quasi-likelihood estimating equations.

In each scenario, we study the bias by considering changes in one parameter over a reasonable range of values, while holding the other parameters fixed. The table below shows the range of values that were considered for the parameters in various scenarios.

Table 3.1: Range table for the model with measurement error

Parameter	Range	Step
α	(-3,3)	0.2
β	(-3,3)	0.2
σ_γ^2	(0,1)	0.1
$\mathbf{p}$	(0,1)	0.1
ρ	(0,1)	0.1
σ_u^2	(0,2)	0.1
n	(50,250)	50

3.1.1 Bias analysis for different values of α

Here, we study the bias of the naive estimator by fixing the coefficient of the true continuous covariate at $\beta = 1$, and varying the coefficient of the categorical predictor, α, from -3 to 3. The parameters σ_u^2, $\mathbf{p}$, σ_γ^2, and ρ, remain unchanged at 2, 0.3, 1 and 0.1, respectively. The summary of simulation results for this scenario are shown in Figure (3.1) and Table (3.2).

From Figure (3.1) and Table (3.2), we observe that:

- As the value of α decreases from 3 to 0, so does the magnitude of bias of the naive estimates of it, reaching its minimum at $\alpha = 0$ (as it was expected). The magnitude of the bias, then increases with the increase in α, in the opposite direction.

- The absolute bias of the naive estimates for β seems to follow a similar pattern as the one for α (except for the last two values).

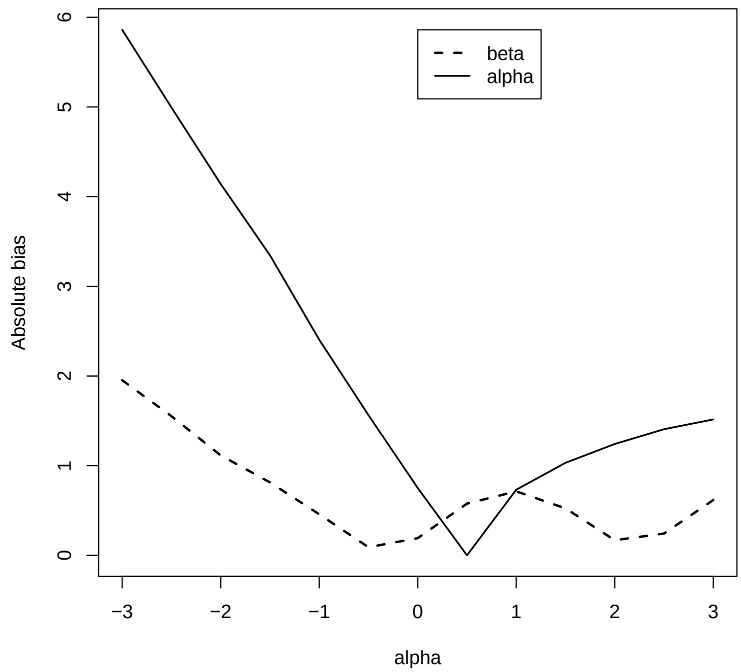

Figure 3.1: Absolute value of the bias of the naive estimators for different values of α

- The variabilities of the two estimators are more or less, unchanged. They may not represent the "real" variabilities, as the search method for estimators, was restricted to small intervals.

3.1.2 Bias analysis for different values of β

In this scenario, the coefficient of the categorical covariate, α, is fixed at 1 and that of the continuous covariate, β, is varied from -3 to 3. The parameters σ_u^2, $\mathbf{p}$, σ_γ^2, and

	$\hat{\alpha}$			$\hat{\beta}$		
α	Naive	Bias	SSE	Naive	Bias	SSE
−3.0	2.8604	5.8604	0.0392	2.9527	1.9527	0.0438
−2.5	2.4940	4.9940	0.0392	2.5521	1.5521	0.0437
−2.0	2.1413	4.1413	0.0403	2.1150	1.1150	0.0418
−1.5	1.8460	3.3460	0.0416	1.8143	0.8143	0.0389
−1.0	1.4048	2.4048	0.0434	1.4586	0.4586	0.0468
−0.5	1.0632	1.5632	0.0427	1.0909	0.0909	0.0451
0.0	0.7512	0.7512	0.0392	0.8080	−0.1920	0.0381
0.5	0.4996	−0.0004	0.0438	0.4228	−0.5772	0.0389
1.0	0.2680	−0.7320	0.0383	0.2862	−0.7138	0.0408
1.5	0.0480	−1.0320	0.0378	0.4755	−0.5245	0.0378
2.0	0.7579	−1.2421	0.0455	0.8309	−0.1691	0.0444
2.5	1.0930	−1.4069	0.0383	1.2437	0.2437	0.0378
3.0	1.4841	−1.5159	0.0414	1.6170	0.6170	0.0372

Table 3.2: Bias and standard error of the naive estimators for selected values of α

ρ, remain unchanged at 2, 0.3, 1 and 0.1, respectively. Shown in Figure (3.2) and Table (3.3) are the simulation results for this scenario.

These were noted from graph output in Figure (3.2) and Table (3.3) that

- The magnitude of the bias in naive estimate for β seems to have V shape, with the smallest value around 0.

- Similar to the last scenario, the absolute bias in the naive estimator of the other model coefficient parameter (α), follows a similar pattern in the one for β.

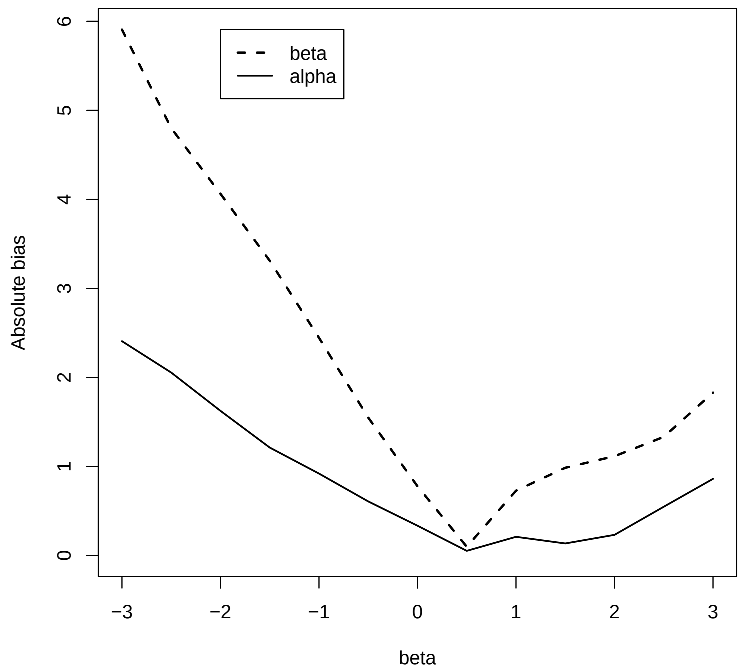

Figure 3.2: Absolute value of the bias of the naive estimators for different values of β

- This implies that although the two variables are independent, measurement error in one, highly affects the bias in the estimator of the other one.

3.1.3 Bias analysis for different values of p

In this case we study the bias in naive estimates when the true proportion of success of the categorical covariate, **p**, is varied from 0.1 to 1. The regression parameters, α and β are fixed at 1 and 0.5, respectively. The other parameters remain the same at

β	$\hat{\alpha}$			$\hat{\beta}$		
	Naive	Bias	SSE	Naive	Bias	SSE
−3.0	2.9071	2.4071	0.0442	2.9063	5.9064	0.0449
−2.5	2.5549	2.0549	0.0405	2.3000	4.8000	0.0395
−2.0	2.1248	1.6248	0.0475	2.0624	4.0624	0.0336
−1.5	1.7127	1.2127	0.0424	1.8110	3.3110	0.0430
−1.0	1.4205	0.9205	0.0421	1.4450	2.4450	0.0442
−0.5	1.1077	0.6077	0.0404	1.0490	1.5490	0.0424
0.0	0.8349	0.3349	0.0423	0.7778	0.7778	0.0414
0.5	0.5526	0.0526	0.0407	0.4013	−0.0987	0.0356
1.0	0.2897	−0.2103	0.0404	0.2719	−0.7281	0.0379
1.5	0.3643	−0.1357	0.0366	0.5145	−0.9855	0.0447
2.0	0.7323	0.2323	0.0439	0.8845	−1.1155	0.0376
2.5	1.0481	0.5481	0.0436	1.1673	−1.3327	0.0423
3.0	1.3613	0.8613	0.0402	1.7034	−1.8297	0.0338

Table 3.3: Bias and standard error of the naive estimators for selected values of β

$\sigma_u^2 = 2$, $\sigma_\gamma^2 = 1$, and $\rho = 0.1$. A graph and a table obtained from simulations under this scenario are given in Figure (3.3) and Table (3.4), respectively.

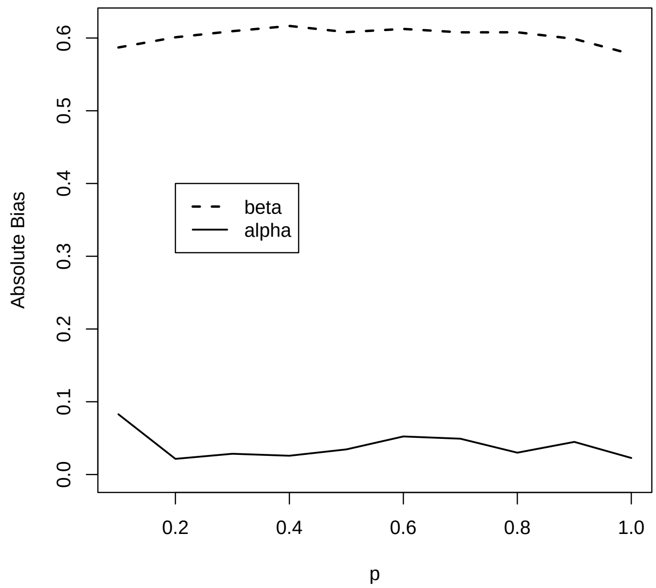

Figure 3.3: Absolute value of the bias of the naive estimators for different values of $\mathbf{p}$

. The major observations one can make from Figure (3.3) and Table (3.4) are

p	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
0.1	0.4172	−0.0828	0.0284	0.4130	−0.5870	0.0280
0.2	0.4785	−0.0215	0.0324	0.3990	−0.6010	0.0261
0.3	0.4715	−0.0285	0.0303	0.3906	−0.6094	0.0275
0.4	0.4742	−0.0258	0.0300	0.3834	−0.6166	0.0281
0.5	0.4655	−0.0345	0.0296	0.3919	−0.6081	0.0279
0.6	0.4477	−0.0523	0.0283	0.3875	−0.6125	0.0268
0.7	0.4508	−0.0492	0.0228	0.3922	−0.6078	0.0277
0.8	0.4700	−0.0300	0.0298	0.3931	−0.6069	0.0265
0.9	0.4552	−0.0448	0.0283	0.4012	−0.5988	0.0291
1.0	0.4773	−0.0227	0.0299	0.4221	−0.5779	0.0298

Table 3.4: Bias and standard error of the naive estimators for selected values of **p**

that:

- The changes in the true proportion **p** has very little impact on the absolute bias of the naive estimate of α
- Similarly, as **p** increases, the absolute bias in the naive estimate of β remains almost as a constant.
- So absolute bias in naive estimators is unaffected by changes in **p**.

3.1.4 Bias analysis for different values of σ_γ^2

Here, we study the bias pattern as variance of Gamma, the individual specific random effect, σ_γ^2 changes from 0.1 to 2 with all other parameters held constant as described before; $\alpha = 0.5$, $\beta = 1$, $\sigma_u^2 = 2$, **p**=0.3, and $\rho = 0.1$.

The results of this scenario are presented in Figure (3.4) and Table (3.5).

.

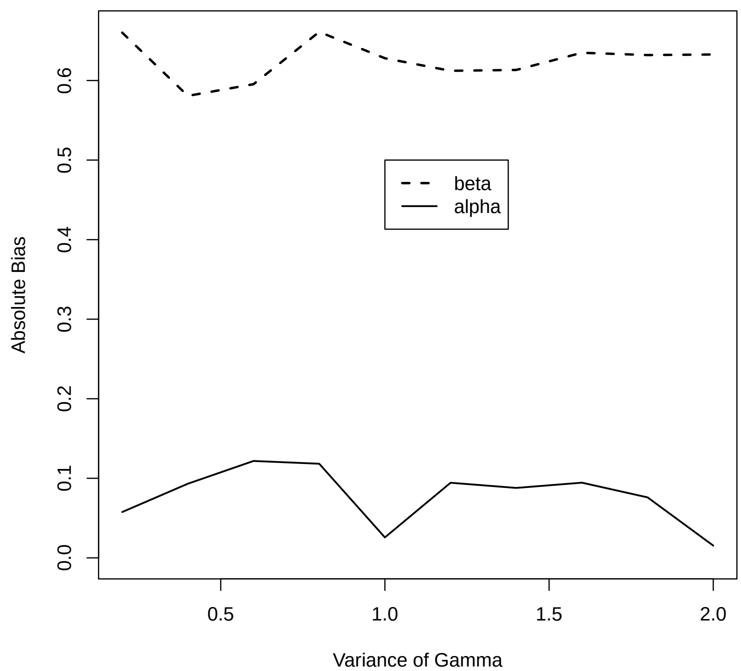

Figure 3.4: Absolute value of the bias of the naive estimators for different values of σ_γ^2

.

	$\hat{\alpha}$			$\hat{\beta}$		
σ_γ^2	Naive	Bias	SSE	Naive	Bias	SSE
0.2	0.4424	−0.0576	0.0381	0.3398	−0.6602	0.0419
0.4	0.4068	−0.0932	0.0392	0.4193	−0.5807	0.0449
0.6	0.3783	−0.1218	0.0362	0.4046	−0.5954	0.0340
0.8	0.3817	−0.1183	0.0394	0.3389	−0.6611	0.0414
1.0	0.4743	−0.0257	0.0452	0.3720	−0.6280	0.0406
1.2	0.4056	−0.0944	0.0428	0.3878	−0.6122	0.0415
1.4	0.4121	−0.0879	0.0447	0.3867	−0.6133	0.0382
1.6	0.4055	−0.0945	0.0375	0.3651	−0.6349	0.0391
1.8	0.4240	−0.0760	0.0398	0.3681	−0.6320	0.0393
2.0	0.4846	−0.0154	0.0369	0.3673	−0.6327	0.0346

Table 3.5: Bias and standard error of the naive estimators for selected values of σ_γ^2

From Figure (3.4) and Table (3.5), the following were observed concerning how changes in the variance of the individual random-effect affects the bias of parameter estimates;

- It is quite surprising that there is no remarkable increasing or decreasing trend in absolute bias of the naive estimates of both α and β.

- Absolute Bias in the estimators is unaffected by changes in the variance of random effect.

.

3.1.5 Bias analysis for different values of σ_u^2

In this scenario, we vary the variance of the measurement error, σ_u^2, from 0 to 2; then, keeping all other parameters unchanged at $\alpha = 0.5$, $\beta = 1$, $\sigma_\gamma^2 = 1$, $\mathbf{p} = 0.3$ and $\rho = 0.1$, we observe the patterns of bias in the naive estimator.

Presented in Figure (3.5) and Table (3.6) are the results of simulations studies for this scenario.

.

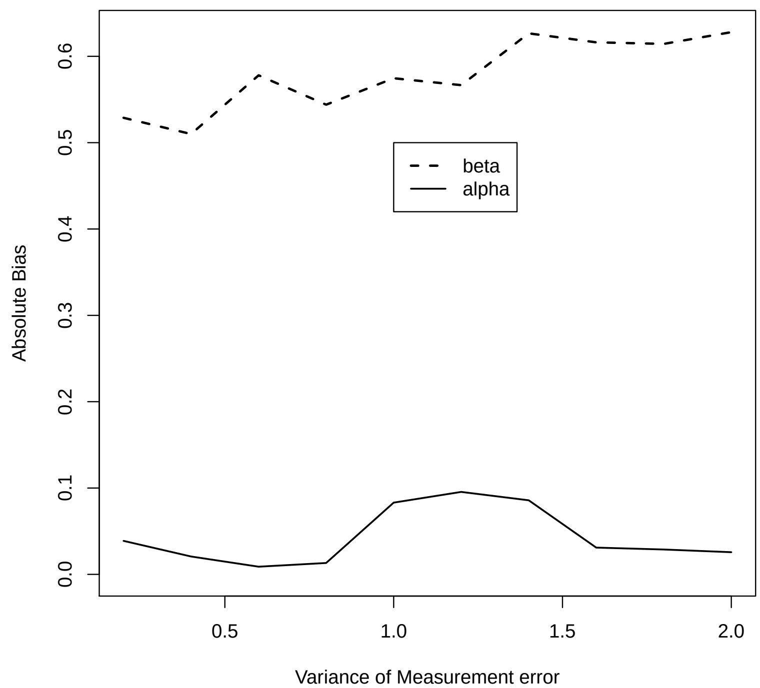

Figure 3.5: Absolute value of the bias of the naive estimators for different values of σ_u^2

σ_u^2	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
0.2	0.4612	−0.0388	0.0441	0.4712	−0.5288	0.0420
0.4	0.4793	−0.0207	0.0422	0.4900	−0.5100	0.0408
0.6	0.4911	−0.0089	0.0394	0.4220	−0.5780	0.0432
0.8	0.4898	−0.0132	0.0444	0.4560	−0.5440	0.0454
1.0	0.4169	−0.0831	0.0395	0.4254	−0.5746	0.0419
1.2	0.4045	−0.0955	0.0395	0.4334	−0.5666	0.0429
1.4	0.4142	−0.0858	0.0438	0.3734	−0.6266	0.0374
1.6	0.4690	−0.0310	0.0425	0.3838	−0.6162	0.0401
1.8	0.4712	−0.0288	0.0419	0.3856	−0.6144	0.0433
2.0	0.4743	−0.0257	0.0452	0.3720	−0.6280	0.0406

Table 3.6: Bias and standard error of the naive estimators for selected values of σ_u^2

From the Figure (3.5) and Table (3.6), one observes that

- Absolute bias in naive estimates of β increases with variance of measurement error.
- The Absolute bias in the naive estimates of α is not much affected by changes in measurement error.

3.1.6 Bias analysis for different values of n

In this scenario, we vary the number of individuals, n, keeping all other parameters unchanged, we observe the patterns of bias in the naive estimator. So in the set-up we have $\alpha = 0.5$, $\beta = 1$, $\sigma_u^2 = 2$, $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$ and $\rho = 0.1$.

The graph and table below are the outcomes of the simulation study in this scenario.

.

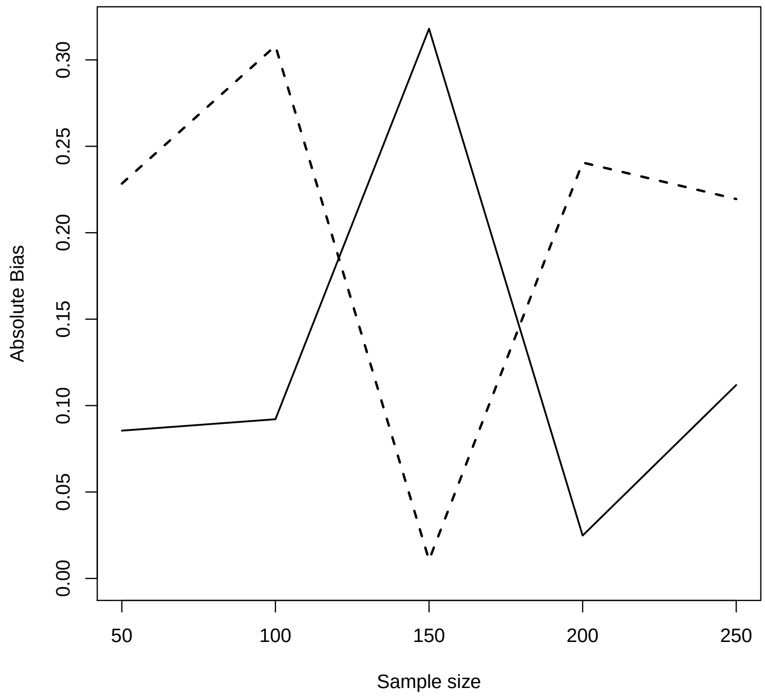

Figure 3.6: Absolute value of the bias of the naive estimators for different values of n

n	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
50	0.5855	0.0855	0.1280	1.2284	0.2284	0.1563
100	0.5921	0.0921	0.1491	1.3081	0.3081	0.1946
150	0.1820	−0.3180	0.1594	0.9894	-0.0106	0.1785
200	0.4752	−0.0248	0.1279	1.2406	0.2406	0.0867
250	0.6119	0.1119	0.2358	1.2195	0.2195	0.0918

Table 3.7: Bias and standard error of the naive estimators for selected values of n

From Figure (3.6) and Table (3.7);

- It is observed that the there is no increasing or decreasing trend in absolute bias of naive estimators for either α or β.
- However, the fluctuations in bias for the parameters go in opposite directions. When absolute bias of one parameter is falling, that of the other will be rising and vice versa.

3.1.7 Bias analysis for different values of ρ

In this study of the bias pattern, we varied the lag correlation coefficient, ρ, from -0.8 to 0.3 and kept all other parameters fixed as given before: $\alpha = 0.5$, $\beta = 1$, $\sigma_u^2 = 2$, $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$ and $\rho = 0.1$. The graphical and tabular results of simulation in this scenario are given in Figure (3.7) and Table (3.8).

.

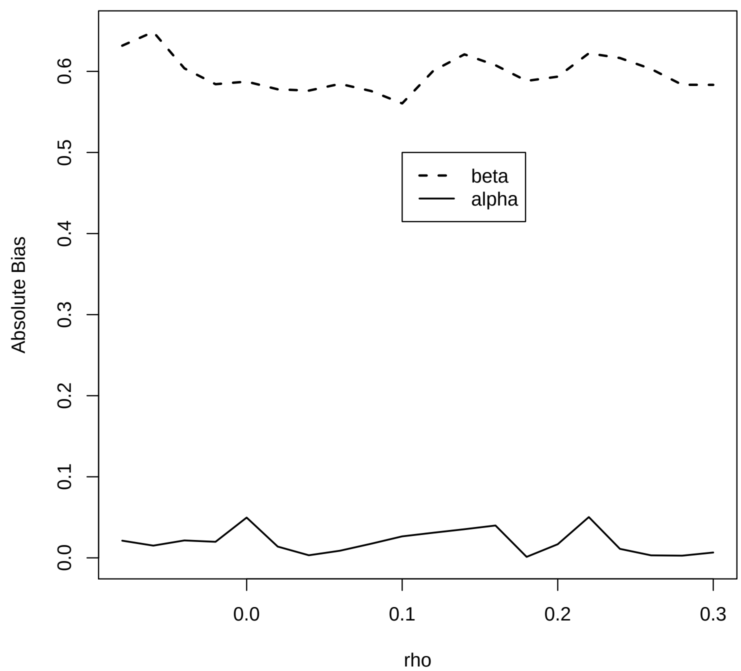

Figure 3.7: Absolute value of the bias of the naive estimators for different values of ρ

- From the results shown in Figure (3.7) and Table (3.8), it is clear that changes in the lag correlation coefficient, ρ has no impact on the absolute bias of the naive estimates.

ρ	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
−0.08	0.4788	−0.0212	0.0275	0.3683	−0.6317	0.0260
−0.06	0.4849	−0.0151	0.0281	0.3513	−0.6487	0.0247
−0.04	0.5215	−0.0215	0.0276	0.3963	−0.6037	0.0273
−0.02	0.4802	−0.0198	0.0284	0.4158	−0.5842	0.0295
0.00	0.4503	−0.0497	0.0268	0.4126	−0.5874	0.0297
0.02	0.4861	−0.0139	0.0283	0.4221	−0.5779	0.0309
0.04	0.4968	−0.0032	0.0295	0.4237	−0.5763	0.0296
0.06	0.4912	−0.0088	0.0311	0.4154	−0.5846	0.0284
0.08	0.4826	−0.0174	0.0312	0.4242	−0.5758	0.0291
0.10	0.4735	−0.0265	0.0304	0.4395	−0.5605	0.0302
0.12	0.4690	−0.0310	0.0321	0.3993	−0.6007	0.0281
0.14	0.4647	−0.0353	0.0296	0.3790	−0.6210	0.0252
0.16	0.4601	−0.0399	0.0304	0.3924	−0.6076	0.0278
0.18	0.4988	−0.0012	0.0293	0.4117	−0.5883	0.0272
0.20	0.4832	−0.0168	0.0286	0.4065	−0.5935	0.0278
0.22	0.4497	−0.0503	0.289	0.3777	−0.6223	0.0277
0.24	0.4889	−0.0111	0.0293	0.3835	−0.6165	0.0275
0.26	0.5130	0.0031	0.0290	0.3970	−0.6030	0.0287
0.28	0.4973	−0.0027	0.0299	0.4165	−0.5835	0.0289
0.30	0.4934	−0.0066	0.0295	0.4166	−0.5834	0.0285

Table 3.8: Bias and standard error of the naive estimators for selected values of ρ

Chapter 4

Simulation Studies on Longitudinal Binary Mixed-effect Model with Misclassification

In certain practical problems, the true categorical predictor, G, cannot be observed; instead, a time-independent variable G^* is observed, which is subject to misclassification.

We consider the case where the true covariate, G and the misclassified covariate, G^* are binary variables with values 0 or 1. Thus, observing value 1 is a success; and

0, a failure. The conditional distribution of G^* given G is defined as follows:

$$\theta_{ij} = P(G^* = i|G = j); \quad i = 0, 1, \quad j = 0, 1, \tag{4.1}$$

where,

$\sum_{i=0}^{1} \theta_{ij} = 1, \quad \text{for each} \;\; j = 0 \;\; \text{or} \;\; 1.$

In literature, the probability of the correct classification of success, θ_{11}, is called *sensitivity*, whereas the probability of correct classification of failure, θ_{00}, is called *specificity*.

Here,

$$\theta_{11} = P(G^* = 1|G = 1), \tag{4.2}$$

$$\theta_{00} = P(G^* = 0|G = 0) \tag{4.3}$$

Since the true predictor, G, is not observed in the model with misclassification, we can write the marginal moments of the response, in the terms of the observed covariate G^*. Therefore, by model assumptions and the law of iterative expectation, we have

$$
\begin{aligned}
& E(y_{it}|X_{it}, G_i^*) \\
= \;& E_{G|G^*}(E(y_{it}|X_{it}, G_i^*, G_i))|X_{it}, G_i^*) \\
= \;& E_{G|G^*}(E(y_{it}|X_{it}, G_i))|X_{it}, G_i^*) \qquad (4.4)\\
\approx \;& E_{G|G^*}\left(\frac{\exp\left(\frac{X_{it}'\beta+G_i\alpha}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}}\right)}{1+exp\left(\frac{X_{it}'\beta+G_i\alpha}{\sqrt{1+\sigma^2_{\gamma}/k^2}}\right)}, |X_{it}, G_i^*\right) \\
= \;& \left(\frac{\exp\left(\frac{X_{it}'\beta}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}}\right)}{1+\exp\left(\frac{X_{it}'\beta}{\sqrt{1+\sigma^2_{\gamma}/k^2}}\right)}\right).P(G=0|G^*) + \left(\frac{\exp\left(\frac{X_{it}'\beta+\alpha}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}}\right)}{1+\exp\left(\frac{X_{it}'\beta+\alpha}{\sqrt{1+\sigma^2_{\gamma}/k^2}}\right)}\right).P(G=1|G^*) \\
= \;& E_{G|G^*}(\mu_{it}|X_{it}, G^*) \qquad (4.5)\\
= \;& \mu^*_{2it} \qquad (4.6)
\end{aligned}
$$

where $k^2 = 1.7$ and equation (4.4) is true, since G^* is assumed to be surrogate.

Also, by independence assumption, $P(G|G^*, X) = P(G|G^*)$.

An approximation for(4.5) is given by equation (4.14) below. Now, we need to calculate $P(G_i = 1|G_i^*)$. Usually, in the application, $P(G)$, or the probability of success for the true variable is known. We define $P(G = 1)$ to be **p**. Hence by The Bayes'

Law,

$$\begin{aligned}
& P(G_i = 1|G_i^* = 0) \\
= & \frac{P(G_i^* = 0|G_i = 1)P(G_i = 1)}{P(G_i^* = 0|G_i = 0)P(G_i = 0) + P(G_i^* = 0|G_i = 1)P(G_i = 1)} \\
= & \frac{\theta_{01}\mathbf{p}}{\theta_{00}(1-\mathbf{p}) + \theta_{01}\mathbf{p}}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& P(G_i = 1|G_i^* = 1) \\
= & \frac{P(G_i^* = 1|G_i = 1)P(G_i = 1)}{P(G_i^* = 1|G_i = 0)P(G_i = 0) + P(G_i^* = 1|G_i = 1)P(G_i = 1)} \\
= & \frac{\theta_{11}\mathbf{p}}{\theta_{10}(1-\mathbf{p}) + \theta_{11}\mathbf{p}}
\end{aligned}$$

$$\begin{aligned}
P(G_i = 1|G_i^*) &= \begin{cases} P(G_i = 1|G_i^* = 0) \\ \\ P(G_i = 0|G_i^* = 1) \end{cases} \quad (4.7) \\
&= \lambda(1|G_i^*) \quad (4.8)
\end{aligned}$$

$P(G_i = 0|G_i^*) = 1 - \lambda(1|G_i^*)$.

The marginal variance and covariance of the response can be obtained in a similar fashion as follows:

$$
\begin{aligned}
& var(y_{it}|X_{it}, G_i^*) \\
= \; & var_{G|G^*}(E((y_{it}|X_{it}, G_i^*, G_i)|X_{it}, G_i^*)) + E_{G|G^*}(var((y_{it}|X_{it}, G_i^*, G_i)|X_{it}, G_i^*)) \\
= \; & var_{G|G^*}(E((y_{it}|X_{it}, G_i)|X_{it}, G_i^*)) + E_{G|G^*}(var((y_{it}|X_{it}, G_i)|X_{it}, G_i^*)) \quad (4.9) \\
= \; & E_{G|G^*}(\mu_{it}|G^*)(1 - E_{G|G^*}(\mu_{it}|G^*)) \quad (4.10)
\end{aligned}
$$

where equation (4.9) results from the surrogacy assumption on G^*, and (4.10) is obtained by the marginal properties of binary distribution, as illustrated in Appendix A. Moreover,

$$
\begin{aligned}
& cov(y_{it}, y_{iu}|X_{it}, G_i^*) \\
= \; & cov_{G|G^*}(E((y_{it}|X_{it}, G_i, G_i^*)|X_{it}, G_i^*), E((y_{iu}|X_{iu}, G_i, G_i^*)|X_{iu}, G_i^*)) \\
+ \; & E_{G|G^*}(cov(((y_{it}|X_{it}, G_i, G_i^*)|X_{it}, G_i^*), ((y_{iu}|X_{iu}, G_i, G_i^*)|X_{iu}, G_i^*))) \\
= \; & cov_{G|G^*}(E((y_{it}|X_{it}, G_i)|X_{it}, G_i^*), E((y_{iu}|X_{iu}, G_i)|X_{iu}, G_i^*)) \quad (4.11) \\
+ \; & E_{G|G^*}(cov(((y_{it}|X_{it}, G_i)|X_{it}, G_i^*), ((y_{iu}|X_{iu}, G_i)|X_{iu}, G_i^*))) \quad (4.12) \\
= \; & cov_{G|G^*}(\mu_{iu}, \mu_{it}|X_{it}, G_i^*)) + E_{G|G^*}(\rho^{t-u}[\mu_{iu} - \mu_{iuu}] + [\mu_{iut} - \mu_{iu}\mu_{it}]) \\
= \; & cov_{G|G^*}(\mu_{iu}, \mu_{it}|X_{it}, G_i^*)) + \rho^{t-u}[\mu_{2iu}^* - \mu_{2iuu}^*] + [\mu_{2iut}^* - \mu_{2iu}^*\mu_{2it}^*]
\end{aligned}
$$

where $\mu_{2iu}^*, \mu_{2it}^*, \mu_{2iuu}^*$ and μ_{2iut}^* are expectations are given in Appendix D. Equations (4.11) and (4.12) are true, as G^* is surrogate.

To estimate the parameters of the model with misclassification, we use the GQL estimating equation of the form:

$$\sum_{i=1}^{n} \frac{\partial \mu_{2i}^{*}{}'}{\partial \theta} \Omega_{2i}^{*-1}(y_i - \mu_{2i}^{*}) = 0 \tag{4.13}$$

where,

$Y_i = (Y_{i1} \ldots Y_{iT})'$ is the $T \times 1$ vector of the observed longitudinal binary responses, for $(i = 1 \ldots n)$;

$E(Y_i|X_i, G_i^*) = \mu_{2i}^* = (\mu_{2i1}^* \ldots \mu_{2iT}^*)'$ is the $T \times 1$ vector of marginal mean of response variable, $(i = 1 \ldots n)$;

$Cov(Y_i|X_i, G_i^*) = \Omega_{2i}^{*-1}$ is the $T \times T$ covariance matrix of response variable, and $\theta = (\beta, \alpha)'$ is the vector of parameters of interest.

$\frac{\partial \mu_{2i}^*}{\partial \theta'}$ is a $(p+1) \times T$ vector of first derivative of marginal mean response.

The marginal mean response is of the form:

$$E_{G|G^*}(\mu_{it}|X_{it}, G^*) = \mu_{2it}^* \approx \left(\frac{\exp(\frac{X_{it}'\beta}{\sqrt{1+\sigma_\gamma^2/k^2}})}{1 + \exp(\frac{X_{it}'\beta}{\sqrt{1+\sigma_\gamma^2/k^2}})} \right) (1-\lambda(1|G^*)) + \left(\frac{\exp(\frac{X_{it}'\beta+\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})}{1 + \exp(\frac{X_{it}'\beta+\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})} \right) \lambda(1|G^*) \tag{4.14}$$

Therefore, $\frac{\partial \mu_{2i}^*}{\partial \theta'} = (\frac{\partial \mu_{2i}^*}{\partial \beta}, \frac{\partial \mu_{2i}^*}{\partial \alpha})'$, where

$$\frac{\partial \mu_{2it}^*}{\partial \beta} \approx \left(\frac{\exp(\frac{X_{it}'\beta}{\sqrt{1+\sigma_\gamma^2/k^2}}) \frac{X_{it}\mathbf{1}_p}{\sqrt{1+\sigma_\gamma^2/k^2}}}{[1 + \exp(\frac{X_{it}'\beta}{\sqrt{1+\sigma_\gamma^2/k^2}})]^2} \right) (1 - \lambda(1|G^*)) + \left(\frac{\exp(\frac{X_{it}'\beta+\alpha^*}{\sqrt{1+\sigma_\gamma^2/k^2}}) \frac{X_{it}\mathbf{1}_p}{\sqrt{1+\sigma_\gamma^2/k^2}}}{[1 + \exp(\frac{X_{it}'\beta+\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})]^2} \right) \lambda(1|G^*)$$

and

$$\frac{\partial \mu_{2it}^*}{\partial \alpha} \approx \frac{\exp\left(\frac{X_{it}'\beta+\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}}\right)\frac{\lambda(1|G^*)}{\sqrt{1+\sigma_\gamma^2/k^2}}}{\left[1+\exp\left(\frac{X_{it}'\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}}\right)\right]^2}.$$

4.1 Simulation Studies

In this section, we perform the simulation studies by generating a categorical variable that is subject to misclassification. There is no ME in the simulation study. Misclassification probabilities are defined by sensitivity value, $\theta_{11} = 0.6$, and specificity value, $\theta_{00} = 0.7$.

Then bias of the naive estimator was examined in various scenarios. Table 4.1 shows the range of values that were considered for the parameters in the various scenarios.

Table 4.1: Range table for the model with misclassification

Parameter	Range	Step
α	(-3,3)	0.2
β	(-3,3)	0.2
$\mathbf{p}$	(0,1)	0.1
σ_γ^2	(0,1)	0.1
ρ	(0,1)	0.1
θ_{11}	(0,1)	0.1
θ_{00}	(0,1)	0.1
n	(50,250)	50

4.1.1 Bias analysis for different values of α

Here we study the bias of the naive estimator by fixing the coefficient of the true continuous covariate at $\beta = 0.5$, and varying the coefficient of the categorical predictor, α from -3 to 3. The other parameters $\mathbf{p}$, σ_γ^2, ρ, θ_{11}, and θ_{00} remain unchanged at 0.3, 1, 0.1 ,0.6, and 0.7, respectively. Presented in Figure (4.1) and Table (4.2) are the results of simulations studies for this scenario:

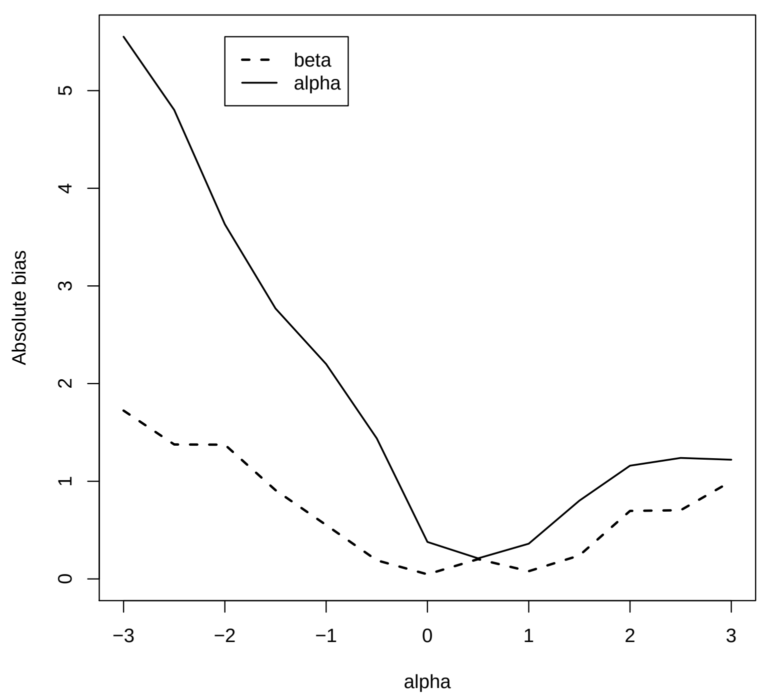

Figure 4.1: Absolute Bias of the Naive Estimators for different values of α

	$\hat{\alpha}$			$\hat{\beta}$		
α	Naive	Bias	SSE	Naive	Bias	SSE
-3.0	2.5523	5.5523	0.03880	2.2239	1.7239	0.0238
-2.5	2.3023	4.8023	0.0380	1.8767	1.3767	0.0365
-2.0	1.6311	3.6311	0.0472	1.8746	1.3746	0.0401
-1.5	1.2693	2.7693	0.0331	1.4090	0.9090	0.0410
-1.0	1.2002	2.2002	0.0303	1.0522	0.5522	0.0425
-0.5	0.9389	1.4389	0.0529	0.6905	0.1905	0.0414
0.0	0.3786	0.3786	0.0322	0.4528	-0.0472	0.0448
0.5	0.2893	-0.2107	0.0265	0.2951	-0.2028	0.0418
1.0	0.6392	-0.3608	0.0356	0.4209	-0.0791	0.0403
1.5	0.6986	-0.0714	0.0229	0.7401	-0.8014	0.0402
2.0	0.8403	-1.1597	0.0269	1.1966	0.6966	0.0412
2.5	1.2611	-1.2389	0.0395	1.2034	0.7034	0.0309
3.0	1.7793	-1.2207	0.0411	1.5003	1.0003	0.0174

Table 4.2: Performance of the Naive estimator for selected values of α

From Figure (4.1) and Table (4.2) the following observations were made:

- As the value of α increases from -3 to 3, the absolute bias in the naive estimates for both α and β decrease until they get to lower values, then increase again.
- For values of α is close to 0, the contribution of α to the model is small, and therefore, the absolute bias in naive estimates are observed to be low.
- As α increases away from 0, the impact of increase in α on the absolute bias in naive estimates of both parameters β is high. In general a kind of U shape is

observed for the bias in naive estimates of both parameters as α changes.

4.1.2 Bias analysis for different values of β

In this scenario, α is fixed at 1 and β is varied from -3 to 3. All other parameter remain the same at $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$, $\rho = 0.1$, $\theta_{11} = 0.6$, and $\theta_{00} = 0.7$. Results for this scenario are shown in Figure (4.2) and Table (4.3).

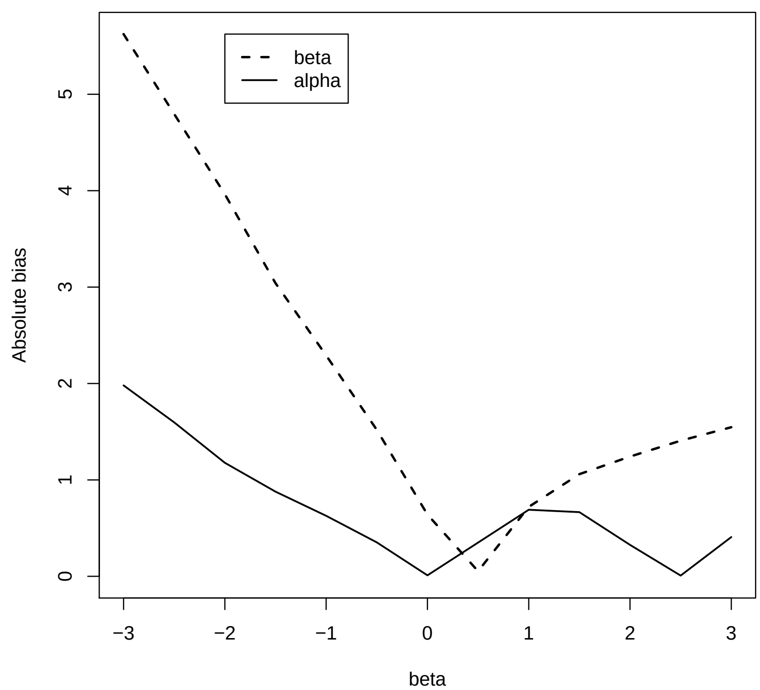

Figure 4.2: Bias of the Naive Estimators for different values of β

These were noted from Figure (4.2) and Table (4.3):

β	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
−3.0	2.9803	1.9803	0.0269	2.6243	5.6243	0.0207
−2.5	2.5967	1.5967	0.0293	2.2951	4.7951	0.0234
−2.0	2.1775	1.1775	0.0318	1.9632	3.9632	0.0266
−1.5	1.8783	0.8783	0.0349	1.5408	3.0408	0.0213
−1.0	1.6277	0.6277	0.0324	1.2914	2.2914	0.0274
−0.5	1.3522	0.3522	0.0217	1.0194	1.5194	0.0287
0.0	1.0102	0.0102	0.0125	0.6408	0.6408	0.0275
0.5	0.6504	−0.3500	0.0117	0.4470	−0.0530	0.0296
1.0	0.3095	−0.6905	0.0142	0.2795	−0.7205	0.0279
1.5	0.3343	−0.6657	0.0187	0.4404	−1.0596	0.0262
2.0	0.6737	−0.3263	0.0196	0.7580	−1.2420.	0.0290
2.5	1.0077	0.0077	0.0216	1.0931	−1.4069	0.0291
3.0	1.4072	0.4072	0.0252	1.4533	−1.5467	0.0279

Table 4.3: Performance of the Naive estimator for selected values of β

- Similar to the observations made in the case of changes in α values, the absolute bias in the naive estimates for both α and β increase as β increases away from 0, and becomes more important in the model.

- More specifically, the absolute bias in the naive estimates of parameters increase initially, for values of β below 0; attain a maximum values, for values of β close to 0; and increase afterward for values of β beyond 0.

- This give rise to the kind of U shape in the overall trend of the absolute bias in the naive estimates of the parameters.

4.1.3 Bias analysis for different values of p

In this case we study the bias in naive estimates when the true proportion of the categorical covariate, **p**, is varied from 0.1 to 1. The regression parameters, α and β, are fixed at 1 and 0.5, respectively. The other parameters remain the same as before; $\sigma_\gamma^2 = 1$, $\rho = 0.1$, $\theta_{11} = 0.6$ and $\theta_{00} = 0.7$. The observed graph and table are shown in Figure (4.3) and Table (4.4) respectively.

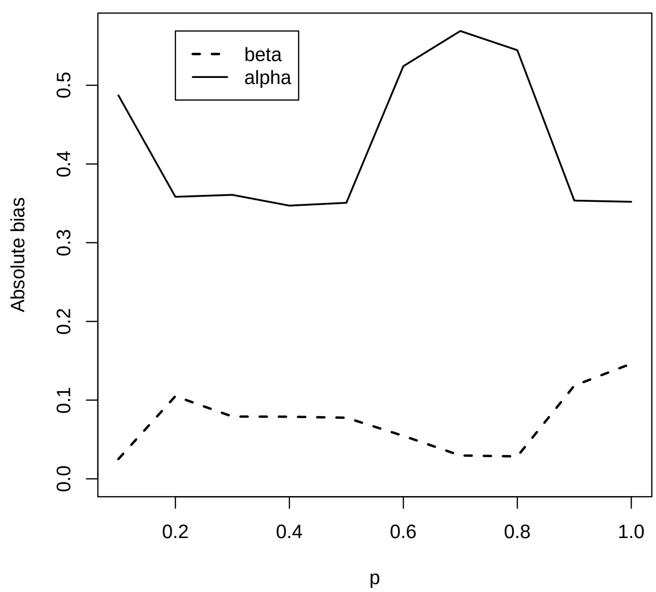

Figure 4.3: Bias of the Naive Estimators for different values of **p**

p	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
0.1	0.5128	−0.4872	0.0482	0.4749	−0.0251	0.0467
0.2	0.6417	−0.3583	0.0178	0.3950	−0.1050	0.0416
0.3	0.6392	−0.3608	0.0158	0.4209	−0.0791	0.0374
0.4	0.6529	−0.3471	0.0164	0.4211	−0.0789	0.0371
0.5	0.6493	−0.3507	0.0168	0.4224	−0.0776	0.0378
0.6	0.4757	−0.5243	0.0466	0.4454	−0.0546	0.0447
0.7	0.4310	−0.5690	0.0446	0.4704	−0.0296	0.0462
0.8	0.4554	−0.5446	0.0473	0.4715	−0.0285	0.0427
0.9	0.6465	−0.3535	0.0234	0.3816	−0.1184	0.0400
1.0	0.6480	−0.3520	0.0271	0.3537	−0.1463	0.0378

Table 4.4: Performance of the Naive estimator for selected values of **p**

Observations made from Figure (4.3) and Table (4.4) include the following:

- As **p** increases from 0.1 to 1, the absolute bias in naive estimates of parameters fluctuate in opposite directions; when the absolute bias for α is decreasing that for β is increasing, and vice versa.

- Therefore, when only one parameter is considered, the changes in the values of **p** does not give rise to a clear increasing or decreasing trend in the bias in naive estimates.

4.1.4 Bias analysis for different values of σ_γ^2

Here, we study the bias pattern as variance of γ, the individual specific random effect, changes from 0.1 to 2, with all other parameters held constant as described before: α and β, are fixed at 1 and 0.5, respectively, $\mathbf{p} = 0.3$, $\rho = 0.1$, $\theta_{11} = 0.6$ and $\theta_{00} = 0.7$. The observed graph and table are shown below in Figure (4.4) and Table (4.5);

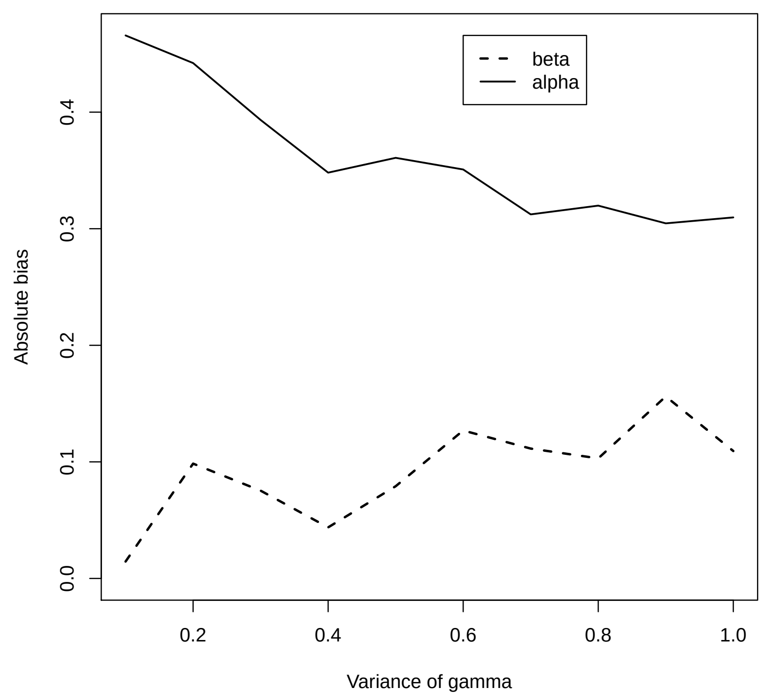

Figure 4.4: Bias of the Naive Estimators for different values of σ_γ^2

From Figure (4.4) and Table (4.5), we made the following observations;

	$\hat{\alpha}$			$\hat{\beta}$		
σ_γ^2	Naive	Bias	SSE	Naive	Bias	SSE
0.2	0.5342	−0.4658	0.0239	0.4855	−0.0145	0.0419
0.4	0.5579	−0.4421	0.0198	0.4015	−0.0985	0.0412
0.6	0.6067	−0.3933	0.0198	0.4247	−0.0753	0.0413
0.8	0.6519	−0.3481	0.0154	0.4562	−0.0438	0.0396
1.0	0.6392	−0.3608	0.0158	0.4209	−0.0791	0.0374
1.2	0.6492	−0.3508	0.0235	0.3731	−0.1269	0.0394
1.4	0.6876	−0.3123	0.0176	0.3886	−0.1114	0.0412
1.6	0.6802	−0.3198	0.0173	0.3970	−0.1030	0.0414
1.8	0.6954	−0.3046	0.0246	0.3439	−0.1561	0.0383
2.0	0.6903	−0.3097	0.0183	0.3907	−0.1093	0.0355

Table 4.5: Performance of the Naive estimator for selected values of σ_γ^2

- There is a remarkable increasing trend in the absolute bias of naive estimates of β.
- Contrary to the above observation, the absolute bias in the naive estimates of α exhibited a decreasing trend.
- Therefore, while the estimates of α improves, as variance of random effect increases, those of β become worse.

4.1.5 Bias analysis for different values of ρ

In this study of the bias pattern, we varied the lag correlation coefficient, ρ, from -0.8 to 0.3 and kept all other parameters fixed as given before: α and β, are fixed at 1 and 0.5, respectively, $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$, $\theta_{11} = 0.6$ and $\theta_{00} = 0.7$. The observed graph and table are shown below; Shown in Figure 4.5 and Table 4.6 are the summary of simulation results for this scenario.

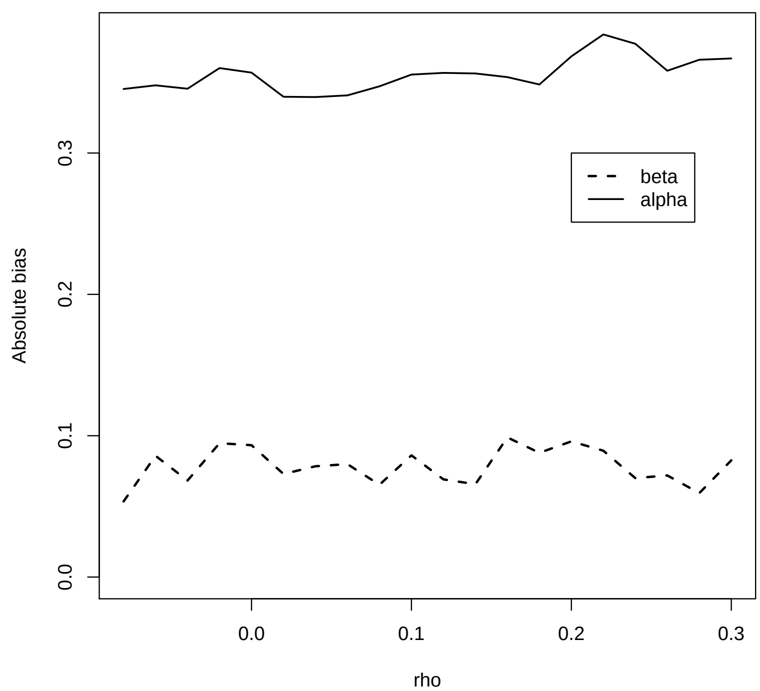

Figure 4.5: Bias of the Naive Estimators for different values of ρ

- From the graph and table presented in Figure 4.5 and Table 4.6, we observe that bias in the naive estimate for α and β show a slight increasing trend as the value of ρ changes.

4.1.6 Bias analysis for different values of *Sensitivity*

In this scenario, we observe the patterns of bias in the naive estimator with varying Sensitivity, θ_{11}, but keeping all other parameters fixed. That is; α and β, are fixed at 1 and 0.5, respectively, $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$, $\rho = 0.1$ and $\theta_{00} = 0.7$. The observed graph and table are shown in Figure 4.6 and Table 4.7 respectively.

- In this set-up, specificity is set high enough for good classification of failure, but the true proportion of success is set quite low.

- From Figure 4.6 and Table 4.7, bias in the naive estimate of α, coefficient of the misclassified variable, is insensitive to the levels of sensitivity. This behaviour may be influenced by the value of $\mathbf{p}$, which seems small.

- Bias in the naive estimate of β is not affected by the changes in the level of sensitivity.

- Bias seems to be insensitive to the values of sensitivity for small value of $\mathbf{p}$.

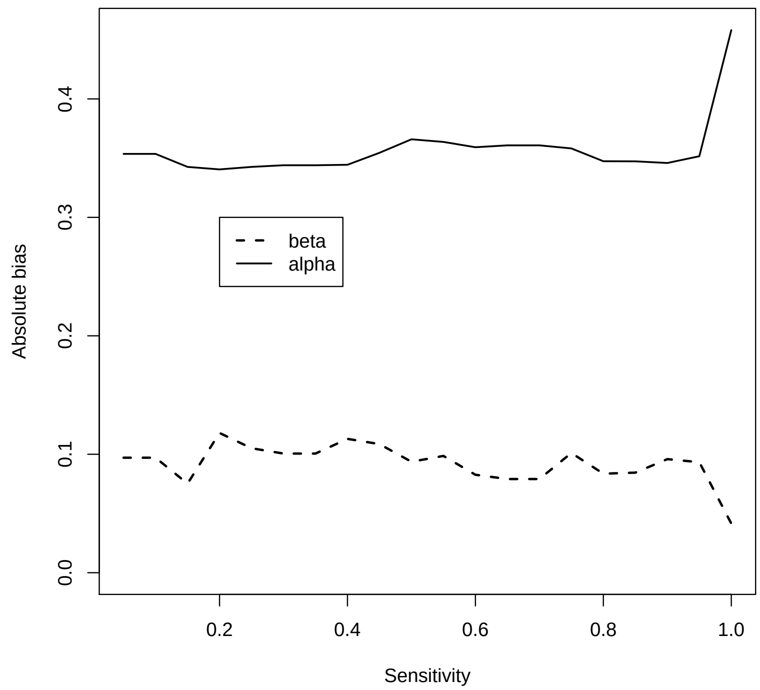

Figure 4.6: Bias of the Naive Estimators for different values of *Sensitivity*

4.1.7 Bias analysis for different values of ***Specificity***

In this scenario, we observe the patterns of bias in the naive estimator with varying Specificity, θ_{00}, but keeping all other parameters fixed. In the set-up we have α and β fixed at 1 and 0.5, respectively, $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$, $\rho = 0.1$, and $\theta_{11} = 0.6$. The observed graph and table are shown in Figure 4.7 and Table 4.8.

- In this scenario, sensitivity is relatively high, meaning that we do well with correct classification of success.

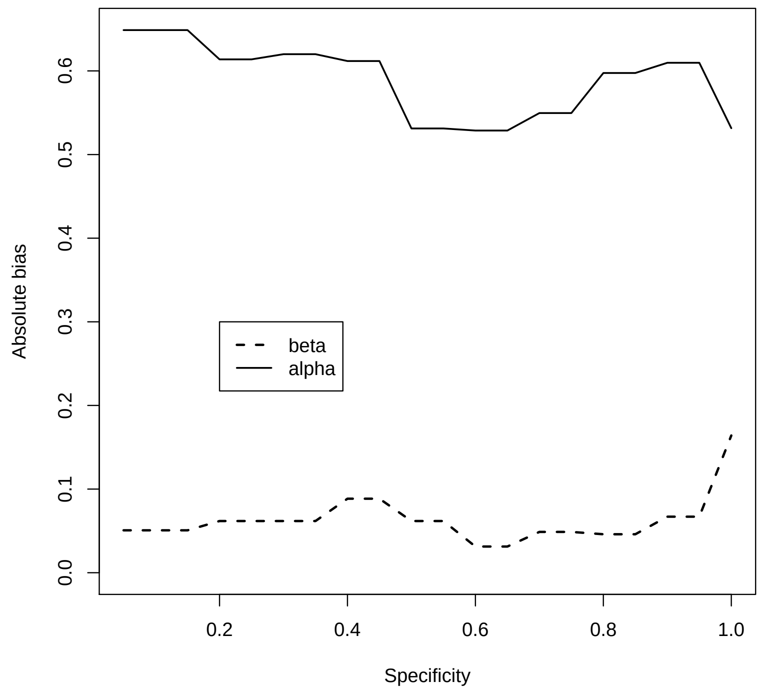

Figure 4.7: Bias of the Naive Estimators for different values of *Specificity*

- As specificity increases, classification with failure improves. Thus from Figure 4.7 and Table 4.8, bias in the coefficient of misclassified variable shows a slight decreasing trend as specificity increases.

- For the coefficient of the continuous variable, the bias in the naive estimate remain quite unaffected as specificity increases.

4.1.8 Bias analysis for different values of sample size

In this scenario, we observe the patterns of bias in the naive estimator by varying the number of individuals, n, but keeping all other parameters fixed. In the set-up we have α and β fixed at 1 and 0.5, respectively, $\mathbf{p} = 0.3$, $\sigma_\gamma^2 = 1$, $\rho = 0.1$, $\theta_{00} = 0.7$, and $\theta_{11} = 0.6$. Figure 4.8 and Table 4.9 summarize the simulation results for this scenario.

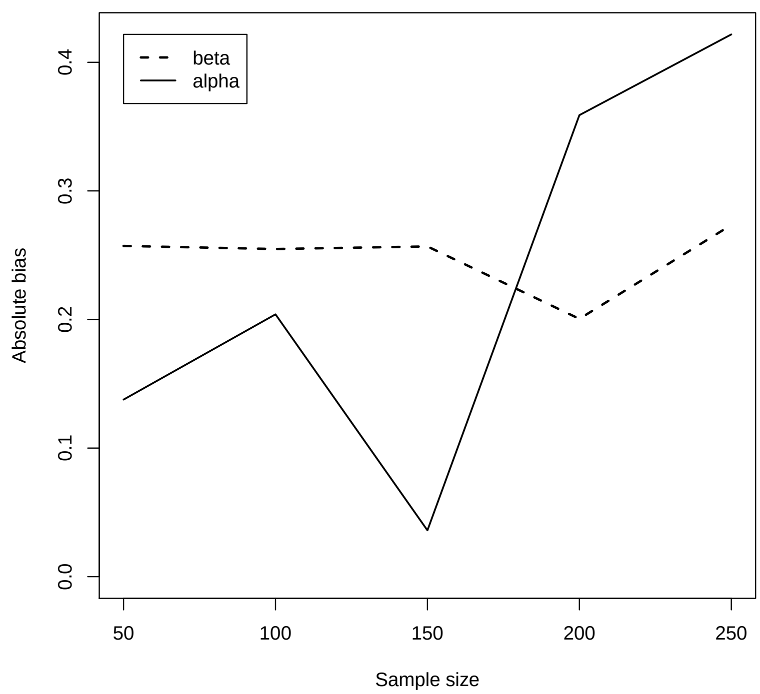

Figure 4.8: Bias of the Naive Estimators for different values of n

- From Figure 4.8 and Table 4.9, it is observed that the there is an increasing trend in bias of naive estimator of α.

- On the other hand, it appears the changes in sample size does not strongly affect the bias in the naive estimate of β.

ρ	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
−0.08	0.6547	−0.3453	0.0200	0.4465	−0.0535	0.2422
−0.06	0.6521	−0.3479	0.0158	0.4142	−0.0858	0.0402
−0.04	0.6545	−0.3455	0.0152	0.4317	−0.0683	0.0416
−0.02	0.6399	−0.3601	0.0190	0.4053	−0.0947	0.0409
0.0	0.6431	−0.3569	0.0197	0.4067	−0.0933	0.0040
0.02	0.6602	−0.3398	0.0146	0.4271	−0.0729	0.0414
0.04	0.6604	−0.3396	0.0184	0.4216	−0.0784	0.0427
0.06	0.6592	−0.3408	0.0190	0.4201	−0.0799	0.0413
0.08	0.6528	−0.3472	0.0189	0.4347	−0.0653	0.0410
0.10	0.6445	−0.3555	0.0188	0.4139	−0.0861	0.0404
0.12	0.6433	−0.3567	0.0182	0.4309	−0.0691	0.0434
0.14	0.6437	−0.3563	0.0147	0.4343	−0.0657	0.0414
0.16	0.6463	−0.3537	0.0160	0.4012	−0.0988	0.0407
0.18	0.6515	−0.3485	0.0159	0.4121	−0.0879	0.0433
0.20	0.6316	−0.3684	0.0173	0.4040	−0.0960	0.0415
0.22	0.6161	−0.3839	0.0174	0.4106	−0.0894	0.0406
0.24	0.6227	−0.3773	0.0184	0.4300	−0.0700	0.0435
0.26	0.6418	−0.3582	0.0183	0.4581	−0.0719	0.0444
0.28	0.6340	−0.3660	0.0147	0.4406	−0.0594	0.0450
0.30	0.6331	−0.3669	0.0164	0.4174	−0.0826	0.0452

Table 4.6: Performance of the Naive estimator for selected values of ρ

	$\hat{\alpha}$			$\hat{\beta}$		
θ_{11}	Naive	Bias	SSE	Naive	Bias	SSE
0.05	.6464	−0.3536	0.0441	0.4029	−0.0971	0.0591
0.10	0.6464	−0.3536	0.0441	0.4029	−0.0971	0.0591
0.15	0.6574	−0.3426	0.0245	0.4248	−0.0752	0.0600
0.20	0.6595	−0.3405	0.0245	0.3820	−0.1180	0.0600
0.25	0.6574	−0.3426	0.0245	0.3949	−0.1051	0.0600
0.30	0.6560	−0.3440	0.0251	0.3994	−0.1006	0.0599
0.35	0.6560	−0.3440	0.0251	0.3994	−0.1006	0.0599
0.40	0.6556	−0.3444	0.0251	0.3870	−0.1130	0.0599
0.45	0.6455	−0.3545	0.0233	0.3914	−0.1086	0.0491
0.50	0.6341	−0.3659	0.0233	0.4063	−0.0937	0.0491
0.55	0.6363	−0.3637	0.0233	0.4014	−0.0986	0.0491
0.60	0.6408	−0.3592	0.0233	0.4173	−0.0827	0.0491
0.65	0.6392	−0.3608	0.0233	0.4209	−0.0791	0.0491
0.70	0.6392	−0.3608	0.0233	0.4209	−0.0791	0.0491
0.75	0.6418	−0.3582	0.0239	0.3989	−0.1011	0.0542
0.80	0.6526	−0.3474	0.0239	0.4164	−0.0836	0.0542
0.85	0.6527	−0.3473	0.0239	0.4155	−0.0836	0.0542
0.90	0.6541	−0.3459	0.0244	0.4041	−0.0959	0.0608
0.95	0.6484	−0.3516	0.0244	0.4067	−0.0933	0.0608
1.00	0.5419	−0.4581	0.0720	0.4585	−0.0415	0.0645

Table 4.7: Performance of the Naive estimator for selected values of *sensitivity*

	$\hat{\alpha}$			$\hat{\beta}$		
θ_{00}	Naive	Bias	SSE	Naive	Bias	SSE
0.05	0.3512	−0.6488	0.0775	0.4493	−0.0507	0.0686
0.10	0.3512	−0.6488	0.0775	0.4493	−0.0507	0.0685
0.15	0.3512	−0.6488	0.0775	0.4493	−0.0507	0.0685
0.20	0.3862	−0.6138	0.0781	0.4381	−0.0618	0.0639
0.25	0.3862	−0.6138	0.07811	0.4381	−0.0618	0.0639
0.30	0.3800	−0.6200	0.0771	0.4352	−0.0618	0.0637
0.35	0.3800	−0.6200	0.0771	0.4352	−0.0618	0.0637
0.40	0.3882	−0.6200	0.0765	0.4115	−0.0618	0.0673
0.45	0.3882	−0.6118	0.0765	0.4115	−0.0885	0.0673
0.50	0.4688	−0.6118	0.0794	0.4381	−0.0885	0.0664
0.55	0.4713	−0.5312	0.0741	0.4686	−0.0618	0.0706
0.60	0.4713	−0.5312	0.0741	0.4686	−0.0618	0.0706
0.65	0.4688	−0.5287	0.0794	0.4381	−0.0313	0.0664
0.70	0.4504	−0.5287	0.0788	0.4511	−0.0313	0.0627
0.75	0.4504	−0.5496	0.0788	0.4511	−0.0488	0.0627
0.80	0.4025	−0.5975	0.0778	0.4539	−0.0460	0.0627
0.85	0.4025	−0.5975	0.0778	0.4539	−0.0460	0.0627
0.90	0.3903	−0.5975	0.0680	0.4330	−0.0460	0.0663
0.95	0.3903	−0.6097	0.0680	0.4330	−0.0670	0.0663
1.0	0.4685	−0.5315	0.0680	0.3359	−0.1641	0.0657

Table 4.8: Performance of the Naive estimator for selected values of θ_{00}

n	$\hat{\alpha}$ Naive	$\hat{\alpha}$ Bias	$\hat{\alpha}$ SSE	$\hat{\beta}$ Naive	$\hat{\beta}$ Bias	$\hat{\beta}$ SSE
50	0.8623	−0.1377	0.1936	0.7572	0.2572	0.1040
100	1.2040	0.2040	0.2309	0.7548	0.2548	0.0909
150	0.9640	−0.0360	0.2270	0.7568	0.2568	0.0534
200	1.3587	0.3589	0.0564	0.7005	0.2005	0.0817
250	1.4217	0.4217	0.0588	0.7735	0.2735	0.0375

Table 4.9: Performance of the Naive estimator for selected values of n

Chapter 5

Simulation Studies on Longitudinal Binary Mixed-effects Model with Measurement Error and Misclassification

In this chapter, we consider a model where the true covariate, X_{it}, is not observed; instead, W_{it} is observed with some measurement error as described in Chapters 2. Also, for unobserved covariate G_i, another variable G_i^* is observed with misclassification as discussed in Chapter 3.

By combining the two model assumptions and the law of iterative expectation, we can write the marginal moments of the response, as follows.

$$
\begin{aligned}
& E(y_{it}|W_{it}, G_i^*) \\
= \;& E_{G|G^*}(E_{X|W}(E(y_{it}|X_{it}, W_{it}, G_i, G_i^*)|W_{it}, G_i^*)|W_{it}, G_i^*) \\
= \;& E_{G|G^*}(E_{X|W}(E(y_{it}|X_{it}, G_i)|W_{it}, G_i^*)|W_{it}, G_i^*) && (5.1) \\
\approx \;& E_{G|G^*}(E_{X|W}(\mu_{it}|W_{it}, G_i^*)|W_{it}, G_i^*) && (5.2) \\
= \;& E_{G|G^*}(\mu_{1it}^*|W_{it}, G_i^*) && (5.3) \\
= \;& \mu_{3it}^* && (5.4)
\end{aligned}
$$

Equation (5.1) comes from the assumption that both G^* and W are surrogates. The inner expectations of equation (5.2) and (5.3) follow the approximations explained in Chapter 2 and Chapter 3 respectively.

Using similar iterative techniques, the marginal variance and covariance of the response can be obtained as follows:

$$
\begin{aligned}
& var\left(y_{it}|W_{it},G_i^*\right) \\
= \; & E_{G|G^*}(E_{X|W}(E\left(y_{it}^2|X_{it},W_{it},G,G_i^*\right)|W_{it},G_i^*)|W_{it},G_i^*) \\
- \; & \left(E_{G|G^*}(E_{X|W}(E(y_{it}|X_{it},W_{it},G,G_i^*)|W_{it},G_i^*)|W_{it},G_i^*)\right)^2 \\
= \; & E_{G|G^*}(E_{X|W}\left(E(y_{it}|X_{it},G_i)|W_{it},G_i^*)|W_{it},G_i^*\right)) \\
- \; & \left(E_{G|G^*}(E_{X|W}(E(y_{it}|X_{it},G_i)|W_{it},G_i^*|W_{it},G_i^*)\right)^2 \qquad (5.5) \\
\approx \; & \mu_{3it}^* - (\mu_{3it}^*)^2 \qquad (5.6) \\
= \; & \mu_{3it}^*(1-\mu_{3it}^*)
\end{aligned}
$$

Equation (5.5) is true, since W is assumed to be surrogate. Moreover, equation (5.6) follows from the marginal properties of binary distribution and derivation from 5.1 to 5.3. Now, when $u < t$, we have

$$
\begin{aligned}
& cov(y_{it},y_{iu}|W_{it},G_i^*) \\
= \; & cov(E_{G|G^*}E_{X|W}(E(y_{it}|X_{it},W_{it},G_i,G_i^*)|W_{it},G_i^*), E_{G|G^*}E_{X|W}(E(y_{iu}|X_{iu},W_{iu},G_i,G_i^*)|W_{iu},G_i^*) \\
+ \; & E_{G|G^*}E_{X|W}(cov(((E(y_{it}|X_{it},W_{it},G_i,G_i^*)|W_{it},G_i^*),((E(y_{iu}|X_{iu},W_{iu},G_i,G_i^*)|W_{iu},G_i^*))) \\
= \; & cov(E_{G|G^*}E_{X|W}(E(y_{it}|X_{it},G_i)|W_{it},G_i^*), E_{G|G^*}E_{X|W}(E(y_{iu}|X_{iu},G_i)|W_{iu},G_i^*) \\
+ \; & E_{G|G^*}E_{X|W}(cov(((E(y_{it}|X_{it},G_i)|W_{it},G_i^*),((E(y_{iu}|X_{iu},G_i)|W_{iu},G_i^*))) \qquad (5.7) \\
= \; & cov((\mu_{3it}^*,\mu_{3iu}^*|W_{it},W_{iu},G_i^*) + E_{G|G^*}E_{X|W}(\rho^{t-u}[\mu_{iu}-\mu_{iuu}]+[\mu_{iut}-\mu_{iu}\mu_{it}]|W_{it},G_i^*)
\end{aligned}
$$

Equation (5.7) is true, since W and G^* are assumed to be surrogate.

To estimate the parameters of the model based on the observed variables, we use the GQL estimating equation of the form:

$$\sum_{i=1}^{n} \frac{\partial {\mu^*_{3i}}'}{\partial \theta^*} \Omega^{*-1}_{3i}(\lambda, \theta, \rho)(y_i - \mu^*_{3i}) = 0, \tag{5.8}$$

where,

$Y_i = (Y_{i1} \dots Y_{iT})'$ is the $T \times 1$ vector of the observed longitudinal binary responses, for $(i = 1 \dots n)$;

$E(Y_i|W_i, G^*_i) = \mu^*_{3i} = (\mu^*_{3i1} \dots \mu^*_{3iT})'$ is the marginal mean of response variable, $(i = 1 \dots n)$;

$Cov(Y_i|W_i, G^*_i) = \Omega^{-1}_{3i}$ is the $T \times T$ covariance matrix of response variable,

$\theta = (\beta', \alpha)'$ to be the vector of parameters of interest.

$\frac{\partial \mu^*_{3i}}{\partial \theta'}$ is a $(p + 1) \times T$ vector of first derivative of marginal mean response.

Therefore, $\frac{\partial \mu^*_{3i}}{\partial \theta'} = (\frac{\partial \mu^*_{3i}}{\partial \beta}, \frac{\partial \mu^*_{3i}}{\partial \alpha})'$,

5.1 Simulation Studies

In this final simulation studies, both the error-prone and misclassified variables were generated as described in the previous studies. Then similar to what was done before, bias of the naive estimator was examined from various scenarios.

Table 5.1 given below shows the range of values that were considered for the parameters for the various scenarios.

Table 5.1: Range table for the model with measurement error and misclassification

Parameter	Range	Step
α	(-3,3)	0.2
β	(-3,3)	0.2
$\mathbf{p}$	(0,1)	0.1
σ^2_γ	(0,1)	0.1
σ^2_u	(0,2)	0.1
θ_{11}	(0,1)	0.1
θ_{00}	(0,1)	0.1
n	(50,250)	50
ρ	(0,1)	0.1

5.1.1 Bias analysis for different values of α

Here we study the bias of the naive estimator by fixing the coefficient of the true continuous covariate at $\beta = 0.5$ and varying the coefficient of the categorical predictor, α from -3 to 3. The parameters σ_u^2 , $\mathbf{p}$, σ_γ^2, ρ, θ_{11}, θ_{00} remain unchanged at 2, 0.3, 1, 0.1, 0.6, and 0.7, respectively with $\alpha = 1$ and $\beta = 0.5$.

Summary of the simulation results are presented in Figure 5.1 and Table 5.2;

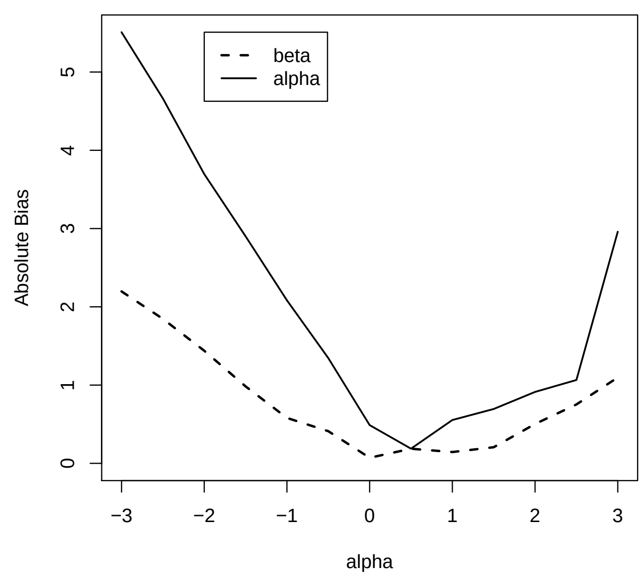

Figure 5.1: Bias of the Naive Estimators for different values of α

From Figure 5.1 and Table 5.2, the following observations were made:

- As the value of α increases from -3 to 3, the bias in the naive estimates for

α	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
-3.0	2.5089	5.5089	0.0461	2.6970	2.1970	0.0229
-2.5	2.1624	4.6624	0.0473	2.3463	1.8463	0.0231
-2.0	1.6962	3.6962	0.0479	1.9388	1.4388	0.0353
-1.5	1.4012	2.9012	2.9012	1.4836	0.9836	0.0439
-1.0	1.0830	2.0830	0.0398	1.0804	0.5804	0.0405
-0.5	0.8452	1.3452	0.0477	0.9111	0.4111	0.0391
0.0	0.4884	0.4884	0.0440	0.5714	0.0714	0.0238
0.5	0.3125	-0.1875	0.0450	0.3133	-0.1867	0.0293
1.0	0.4464	-0.5536	0.0469	0.3556	-0.1444	0.0366
1.5	0.8052	-0.6948	0.0473	0.7061	0.2061	0.0447
2.0	1.0872	-0.9128	0.0466	1.0029	0.5029	0.0409
2.5	1.4349	-1.0651	0.0501	1.2528	0.7528	0.0297
3.0	1.7759	-1.2241	0.0497	1.5968	1.0968	0.0254

Table 5.2: Performance of the Naive estimator for selected values of α

both α and β decrease, get to lower values and eventually rise.

- For points where α is close to 0, the contribution of α to the model is little and the bias in naive estimates are observed to be low.

- However, as α increases away from 0, the impact of increase in α on the bias in naive estimates of both parameters β is high. In general a kind of U shape is observed for the bias in naive estimates of both parameters, as α changes.

5.1.2 Bias analysis for different values of β

In this scenario, α is fixed at 1 and β is varied from -3 to 3. All other parameters remain the same as given before: the parameters σ_u^2 , $\mathbf{p}$, σ_γ^2, ρ, θ_{11}, θ_{00} remain unchanged at 2, 0.3, 1, 0.1 , 0.6 and 0.7, respectively. Figure 5.2 and Table 5.3 are the results of the simulations for this scenario.

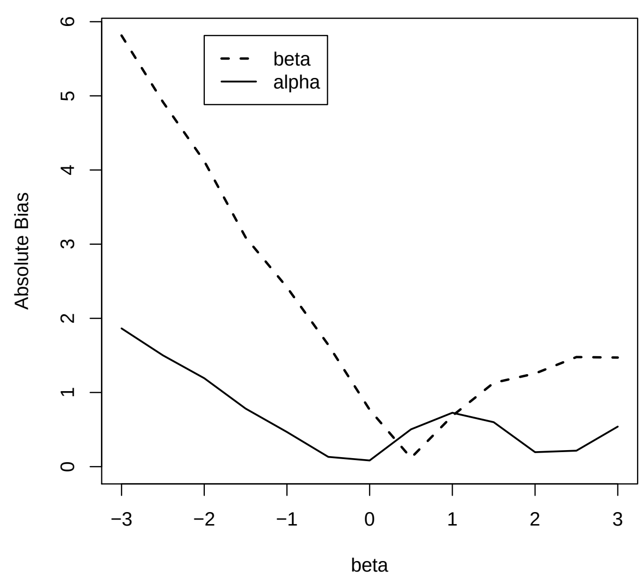

Figure 5.2: Bias of the Naive Estimators for different values of β

These were noted from Figure 5.2 and Table 5.3:

- On the average, the bias in naive estimate for β has a kind of U shape.
- For values of β around 0,the bias in naive estimate for β is low.

β	$\hat{\alpha}$			$\hat{\beta}$		
	Naive	Bias	SSE	Naive	Bias	SSE
-3.0	2.8643	1.8643	0.0475	2.8136	5.8136	0.0482
-2.5	2.5012	1.5012	0.0412	2.4188	4.9188	0.0438
-2.0	2.1925	1.1925	0.0378	2.1194	4.1194	0.0447
-1.5	1.7831	0.7831	0.0438	1.5926	3.0926	0.0203
-1.0	1.4671	0.4671	0.0420	1.4191	2.4191	0.0457
-0.5	1.1318	0.1318	0.0433	1.1422	1.6422	0.0433
0.0	0.9166	-0.0834	0.0382	0.7678	0.7678	0.0417
0.5	0.4967	-0.5033	0.0448	0.3820	-0.1180	0.0327
1.0	0.2733	-0.7267	0.0389	0.3154	-0.6846	0.0389
1.5	0.4005	-0.5995	0.0416	0.3695	-1.1305	0.0354
2.0	0.8047	-0.1953	0.0431	0.7441	-1.2559	0.0486
2.5	1.2153	0.2153	0.0396	1.0222	-1.4778	0.0437
3.0	1.5401	0.5401	0.0441	1.5288	-1.4712	0.0507

Table 5.3: Performance of the Naive estimator for selected values of β

- As β increases away from 0, the bias in naive estimate for β increases.
- As β increases, it is observed that the bias in naive estimate for α decreases.

5.1.3 Bias analysis for different values of p

In this case we study the bias in naive estimates when the true proportion of the categorical covariate, **p**, is varied from 0.1 to 1. The regression parameters, α and β are fixed at 1 and 0.5, respectively. The other parameters remain the same. Thus σ_u^2, σ_γ^2, ρ, θ_{11}, θ_{00} remain unchanged at 2, 1, 0.1 ,0.6 and 0.7, respectively. Simulation results for this scenario are shown in Figure 5.3 and Table 5.4.

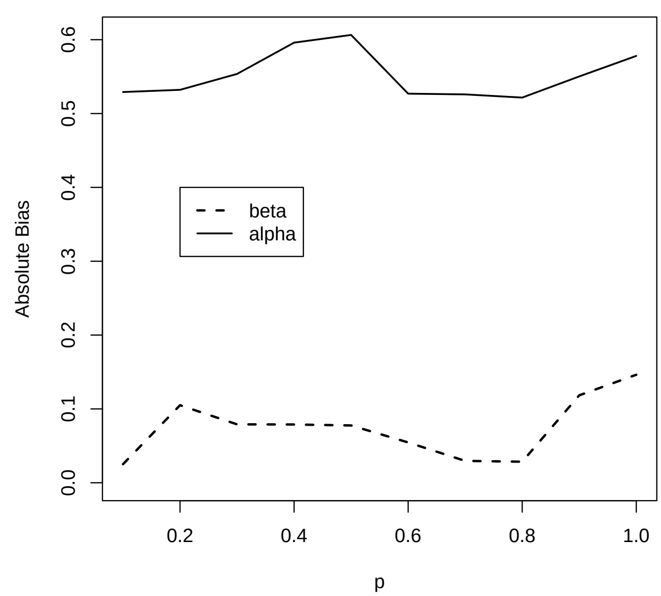

Figure 5.3: Bias of the Naive Estimators for different values of π

Observations that were made from Figure 5.3 and Table 5.4 include the following:

- The bias of the naive estimate of α is not much affected by the increase in **p**.

p	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
0.1	0.4709	−0.5291	0.0440	0.4749	−0.0251	0.0364
0.2	0.4679	−0.5321	0.0484	0.3950	−0.1050	0.0293
0.3	0.4464	−0.5536	0.0469	0.4209	−0.0791	0.0366
0.4	0.4040	−0.5960	0.0461	0.4211	−0.0789	0.0398
0.5	0.3936	−0.6064	0.0455	0.4224	−0.0776	0.0376
0.6	0.4731	−0.5269	0.0425	0.4454	−0.0546	0.0379
0.7	0.4741	−0.5259	0.0372	0.4704	−0.0296	0.0380
0.8	0.4784	−0.5216	0.0377	0.4715	−0.0285	0.0412
0.9	0.4498	−0.5502	0.0460	0.3816	−0.1184	0.0413
1.0	0.4220	−0.5780	0.0433	0.3537	−0.1463	0.0403

Table 5.4: Performance of the Naive estimator for selected values of **p**

The fluctuations in bias does not show any significant increasing or decreasing pattern.

- For the bias in the naive estimate of β, there is an observed slight increasing trend as **p** increases.

5.1.4 Bias analysis for different values of σ_γ^2

Here, we study the bias pattern as variance of Gamma, the individual specific random effect, σ_γ^2 changes from 0.1 to 2 with all other parameters held constant as described before. So we have the parameters σ_u^2 , $\mathbf{p}$, ρ, θ_{11}, θ_{00} fixed at 2, 0.3, 0.1 ,0.6 and 0.7, respectively with $\alpha = 1$ and $\beta = 0.5$. Summary of simulation results are presented in Figure 5.4 and Table 5.5.

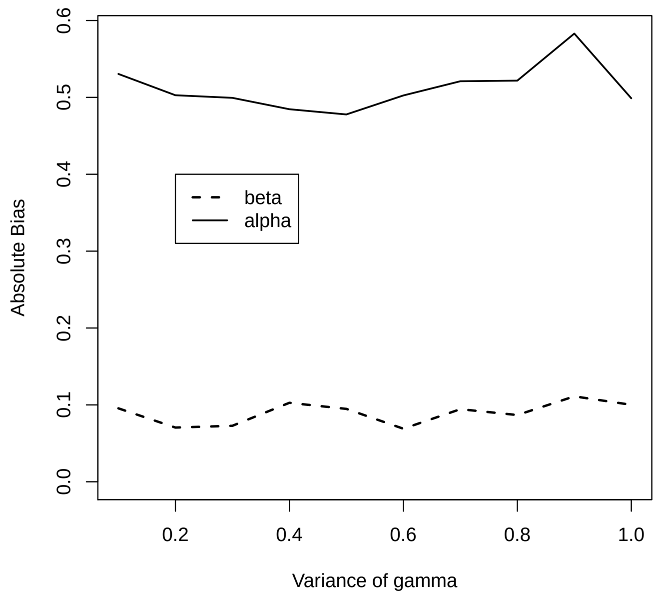

Figure 5.4: Bias of the Naive Estimators for different values of γ

Below are comments related to graph and table in Figure 5.4 and Table 5.5:

- As variance of the individual specific random effect increases, the bias in naive estimates of both parameter are not much affected by the changes. There is no observed increasing or decreasing pattern in bias.

	$\hat{\alpha}$			$\hat{\beta}$		
σ_γ^2	Naive	Bias	SSE	Naive	Bias	SSE
0.2	0.4695	−0.5305	0.0301	0.4045	−0.0955	0.0282
0.4	0.4972	−0.5028	0.0311	0.4295	−0.0705	0.0249
0.6	0.5006	−0.4994	0.0309	0.4272	−0.0728	0.0254
0.8	0.5154	−0.4846	0.0307	0.3973	−0.1027	0.0246
1.0	0.5228	−0.4778	0.0312	0.4052	−0.0948	0.0240
1.2	0.4975	−0.5025	0.0282	0.4310	−0.0690	0.0259
1.4	0.4790	−0.5210	0.0297	0.4056	−0.0944	0.0266
1.6	0.4782	−0.5218	0.0296	0.4132	−0.0868	0.0259
1.8	0.4170	−0.5830	0.0302	0.3889	−0.1111	0.0234
2.0	0.5012	−0.4988	0.0302	0.3998	−0.1002	0.0254

Table 5.5: Performance of the Naive estimator for selected values of σ_γ^2

5.1.5 Bias analysis for different values of σ_u^2

In this scenario, we observe the patterns of bias in the naive estimator by varying the variance of the measurement error, σ_u^2 while keeping all other parameters fixed. Thus, the parameters $\mathbf{p}$, σ_γ^2, ρ, θ_{11}, θ_{00} remain unchanged at 0.3, 1, 0.1, 0.6, and 0.7, respectively, with $\alpha = 1$ and $\beta = 0.5$. Presented in Figure 5.5 and Table 5.6 are the simulation results for this case.

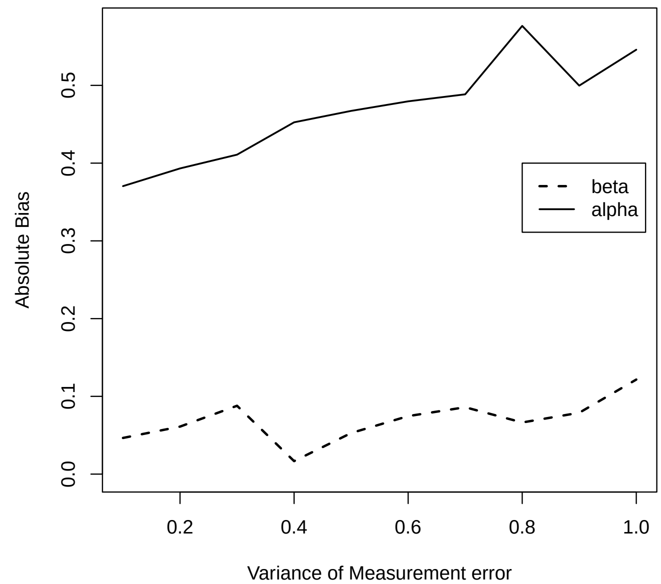

Figure 5.5: Bias of the Naive Estimators for different values of σ_u^2

- Results from Figure 5.5 and Table 5.6 indicate that when covariate is subject to both measurement error and misclassification, bias in naive estimator for β increase with variance of measurement error.

- The bias of naive estimate of α, also increase with variance of measurement error.

	$\hat{\alpha}$			$\hat{\beta}$		
σ_u^2	Naive	Bias	SSE	Naive	Bias	SSE
0.2	0.6296	−0.3704	0.0139	0.4535	−0.0464	0.0280
0.4	0.6068	−0.3932	0.0244	0.4389	−0.0611	0.0283
0.6	0.5891	−0.4109	0.0253	0.4121	−0.0879	0.0289
0.8	0.5475	−0.4525	0.0246	0.4833	−0.0167	0.0297
1.0	0.5328	−0.4672	0.0281	0.4472	−0.0528	0.0306
1.2	0.5205	−0.4795	0.0303	0.4254	−0.0746	0.0305
1.4	0.5115	−0.4885	0.0289	0.4140	−0.0860	0.0311
1.6	0.4235	−0.5765	0.0280	0.4336	−0.0664	0.0310
1.8	0.5003	−0.4997	0.0281	0.4211	−0.0789	0.0301
2.0	0.4541	−0.5459	0.0298	0.3784	−0.1216	0.0255

Table 5.6: Performance of the Naive estimator for selected values of σ_u^2

5.1.6 Bias analysis for different values of *Sensitivity*

In this scenario, we vary Sensitivity, θ_{11}, while keeping all other parameters unchanged. Then we observe the patterns of bias in the naive estimator. So we have the parameters σ_u^2 , $\mathbf{p}$, σ_γ^2, ρ, θ_{00} fixed at 2, 0.3, 1, 0.1 and 0.7, respectively, with $\alpha = 1$ and $\beta = 0.5$.

Figure 5.6 and Table 5.7 give summary of the results from simulations.

- It is observed from Figure 5.6 and Table 5.7 that as sensitivity increases, bias in the naive estimate of α shows a slight decreasing trend.
- Also, bias in the naive estimate of β shows a similar decreasing pattern as

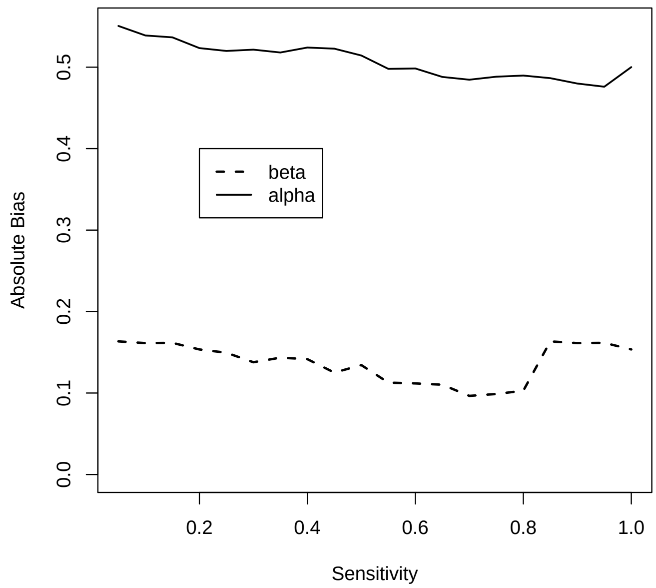

Figure 5.6: Bias of the Naive Estimators for different values of *sensitivity*

sensitivity changes.

- It is surprising that in the presence of both errors, the estimates appear to be improving.

5.1.7 Bias analysis for different values of $Specificity$

In this scenario, we vary Specificity, θ_{00}, then keeping all other parameters unchanged, we observe the patterns of bias in the naive estimator. Thus, parameters σ_u^2 , $\mathbf{p}$, σ_γ^2, ρ, θ_{11} remain unchanged at 2, 0.3, 1, 0.1 and 0.6, respectively, with $\alpha = 1$

and $\beta = 0.5$. Figure 5.7 and Table 5.8 present summary of the simulation results.

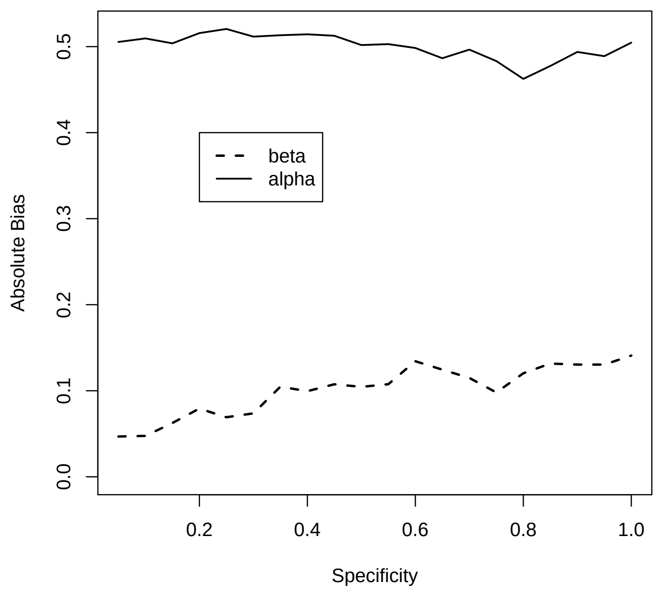

Figure 5.7: Bias of the Naive Estimators for different values of $specificity$

- The simulation results shown in Figure 5.7 and Table 5.8 indicate that bias in the naive estimate of β increases with specificity while that of α remain quite stable.

5.1.8 Bias analysis for different values of Sample size

In this scenario, we vary the number of individuals, n, while keeping all other parameters unchanged. Then we observe the patterns of bias in the naive estimator. Thus, the parameters σ_u^2, $\mathbf{p}$, σ_γ^2, ρ, θ_{11}, θ_{00} remain unchanged at 2, 0.3, 1, 0.1, 0.6, and 0.7, respectively, with $\alpha = 1$ and $\beta = 0.5$.

The graph and table in Figure 5.8 and Table 5.9 are the outcomes of the simulation study in this scenario;

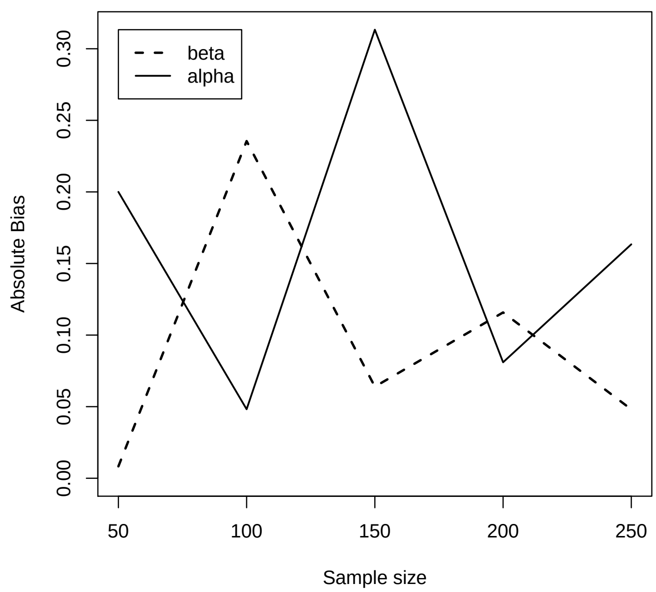

Figure 5.8: Bias of the Naive Estimators for different values of n

- From Figure 5.8 and Table 5.9 we observed that there is no remarkable increasing or decreasing pattern in bias on naive estimate of either parameter.

5.1.9 Bias analysis for different values of ρ

In this study of the bias pattern, we varied the lag correlation coefficient, ρ, from -0.8 to 0.3, and kept all other parameters fixed as given before. The parameters σ_u^2, $\mathbf{p}$, σ_γ^2, θ_{11}, θ_{00} remain unchanged at 2, 0.3, 1, 0.6, and 0.7 respectively, with $\alpha = 1$ and $\beta = 0.5$.

For this scenario, we obtained from simulations the graph and table shown in Figure 5.9 and Table 5.10.

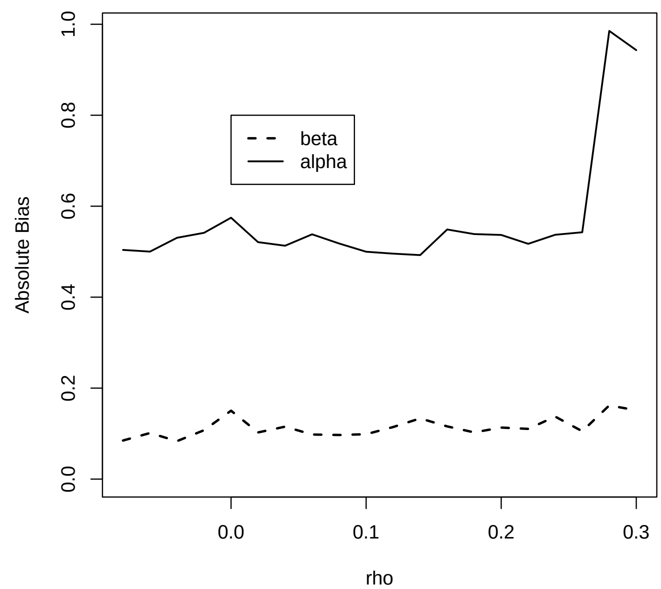

Figure 5.9: Bias of the Naive Estimators for different values of ρ

- From the results in Figure 5.9 and Table 5.10, it is clear that changes in the lag correlation coefficient, ρ, has no impact on the bias of the naive estimates. Absolute bias for both parameters remain the same for almost all the values of ρ.

θ_{11}	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
0.05	0.4494	−0.5506	0.0292	0.3366	−0.1634	0.0215
0.10	0.4611	−0.5389	0.0288	0.3387	−0.1613	0.0221
0.15	0.4634	−0.5366	0.0289	0.3384	−0.1616	0.0228
0.20	0.4766	−0.5234	0.0296	0.3465	−0.1535	0.0233
0.25	0.4801	−0.5199	0.0300	0.3505	−0.1495	0.0239
0.30	0.4785	−0.5215	0.0300	0.3622	−0.1378	0.0234
0.35	0.4820	−0.5180	0.0302	0.3566	−0.1434	0.0243
0.40	0.4759	−0.5241	0.0298	0.3584	−0.1416	0.0245
0.45	0.4773	−0.5227	0.0295	0.3749	−0.1251	0.0237
0.50	0.4857	−0.5143	0.0285	0.3657	−0.1343	0.0238
0.55	0.5021	−0.4979	0.0284	0.3872	−0.1128	0.0254
0.60	0.5016	−0.4984	0.0290	0.3882	−0.1118	0.0243
0.65	0.5120	−0.4880	0.0291	0.3898	−0.1102	0.0262
0.70	0.5154	−0.4846	0.0288	0.4035	−0.0965	0.0274
0.75	0.5117	−0.4883	0.0286	0.4012	−0.0988	0.0273
0.80	0.5103	−0.4897	0.0290	0.3971	−0.1029	0.0264
0.85	0.5135	−0.4865	0.0290	0.3366	−0.1634	0.0271
0.90	0.5201	−0.4799	0.0295	0.3387	−0.1613	0.0276
0.95	0.5240	−0.4770	0.0298	0.3384	−0.1616	0.0277
1.00	0.4999	−0.5001	0.0323	0.3465	−0.1535	0.0253

Table 5.7: Performance of the Naive estimator for selected values of *Sensitivity*

θ_{00}	$\hat{\alpha}$			$\hat{\beta}$		
	Naive	Bias	SSE	Naive	Bias	SSE
0.05	0.4946	−0.5054	0.0300	0.4532	−0.0468	0.0298
0.10	0.4005	−0.5995	0.0302	0.4525	−0.0475	0.0299
0.15	0.4962	−0.5038	0.0310	0.4375	−0.0625	0.0294
0.20	0.4843	−0.5157	0.0306	0.4207	−0.0793	0.0293
0.25	0.4795	−0.5205	0.0300	0.4308	−0.0692	0.0290
0.30	0.4884	−0.5116	0.0300	0.4263	−0.0737	0.0278
0.35	0.4867	−0.5132	0.0307	0.3953	−0.1047	0.0267
0.40	0.4857	−0.5143	0.03006	0.4045	−0.0995	0.0266
0.45	0.4874	−0.5126	0.0303	0.3925	−0.1075	0.0264
0.50	0.4982	−0.5018	0.0302	0.3955	−0.1045	0.0269
0.55	0.4971	−0.5029	0.0300	0.3924	−0.1076	0.0263
0.60	0.5016	−0.4984	0.0290	0.3657	−0.1343	0.0243
0.65	0.5135	−0.4865	0.0283	0.3755	−0.1245	0.0269
0.70	0.5035	−0.4965	0.0283	0.3851	−0.1149	0.0268
0.75	0.5168	−0.4832	0.0288	0.4020	−0.0980	0.0275
0.80	0.5375	−0.4625	0.0288	0.3797	−0.1203	0.0268
0.85	0.5225	−0.4775	0.0290	0.3684	−0.1316	0.0250
0.90	0.5062	−0.4938	0.0300	0.3696	−0.1304	0.0242
0.95	0.5111	−0.4889	0.0297	0.3696	−0.1304	0.0242
1.0	0.4954	−0.5046	0.0291	0.3591	−0.1409	0.0232

Table 5.8: Performance of the Naive estimator for selected values of *Specificity*

n	$\hat{\alpha}$			$\hat{\beta}$		
	Naive	Bias	SSE	Naive	Bias	SSE
50	0.8000	−0.2000	0.0829	0.5083	0.0083	0.1046
100	1.0481	0.0482	0.2464	0.7355	0.2355	0.1147
150	0.6866	−0.3133	0.0757	0.4360	−0.0640	0.1880
200	0.9190	−0.0810	0.1736	0.6158	0.1158	0.2020
250	0.8366	−0.1634	0.1710	0.5481	0.0481	0.0971

Table 5.9: Performance of the Naive estimator for selected values of n

ρ	$\hat{\alpha}$ Naive	Bias	SSE	$\hat{\beta}$ Naive	Bias	SSE
−0.08	0.4961	−0.5039	0.0429	0.4152	−0.0848	0.0413
−0.06	0.4998	−0.5002	0.04439	0.3987	−0.1013	0.0415
−0.04	0.4694	−0.5306	0.0436	0.4165	−0.0835	0.0416
−0.02	0.4585	−0.5415	0.0450	0.3925	−0.1075	0.0363
0.0	0.4253	−0.5747	0.0466	0.0350	−0.1504	0.0030
0.02	0.4790	−0.5210	0.0455	0.3974	−0.1026	0.0349
0.04	0.4869	−0.5131	0.0455	0.3847	−0.1153	0.0348
0.06	0.4618	−0.5382	0.0448	0.4020	−0.0980	0.0389
0.08	0.4820	−0.5180	0.0444	0.4030	−0.0970	0.0414
0.10	0.5001	−0.4999	0.0445	0.4013	−0.0987	0.0432
0.12	0.5043	−0.4957	0.0423	0.3855	−0.1145	0.0390
0.14	0.5075	−0.4925	0.0420	0.3665	−0.1325	0.0330
0.16	0.4512	−0.5488	0.0445	0.3841	−0.1159	0.0370
0.18	0.4613	−0.5387	0.0426	0.3973	−0.1027	0.0401
0.20	0.4632	−0.5368	0.0415	0.3867	−0.1133	0.0400
0.22	0.4827	−0.5173	0.04133	0.3990	−0.1104	0.0408
0.24	0.4628	−0.5372	0.0394	0.3626	−0.1374	0.0377
0.26	0.4573	−0.5427	0.0402	0.3948	−0.1052	0.0450
0.28	0.4638	−0.9854	0.0416	0.3380	−0.1620	0.0386
0.30	0.4177	−0.9431	0.0421	0.3487	−0.1513	0.0364

Table 5.10: Performance of the Naive estimator for selected values of ρ

Chapter 6

Discussion

The focus of this research is to study the bias pattern in the naive estimator of a Longitudinal Binary Mixed-effect model with Measurement error and Misclassification in covariates. The prevalence of binary outcomes in some areas of studies like Genetics, Environmental and Behaviorial sciences serves as a motivation for this study. In such areas, the use of error-prone variables instead of unobserved variables, is very common; and that often leads to bias in naive estimators. In our study we aim at how changes in each of the parameters in our model affect the bias in the naive estimators, in the situation where the others are held constant.

The model we considered in the study has several advantages. First, it is a contribution to statistical modelling and analysis involving categorical responses. It also enabled us to study the bias pattern using a model that incorporates both categorical and continuous covariates; this is not a common feature in most literature. Furthermore, by including the individual specific random effect in the model, we are able to account for subject specific variations in the mean response. This is not so with marginal models; as they only account for changes in overall mean response. Finally, the model also takes lag correlation into account; this is very necessary feature of any Longitudinal Studies. We studied the bias patterns, first, for a model with measurement error in covariate; then we did the same for model with Misclassification in covariate; and finally, for a model with both measurement error and misclassification in covariates.

For the model with measurement error, we confirmed that the bias in naive estimators increase as the variance of measurement error increases. Most of the well known measurement error literature such as Fuller (1987) has drawn similar conclusion. In particular, we have worse estimates for β, the coefficient of the continuous variable, as the variance of measurement error increases.

For the model with misclassification, it was observed that for ordinary events where the probability of success for the categorical covariate is quite large and sensitivity is fixed, the bias in α, the coefficient of the categorical variable reduces as specificity increases. So by improving specificity in such scenario, we can improve the estimates of α. However, for rare events where the probability of success is very small and specificity is fixed, bias in the naive estimates remain unaffected for all levels of sensitivity, which implies, in such case, we cannot improve estimates by varying sensitivity. Buonaccorsi (2010) has similar discussion for simple linear regression with misclassification in the covariate.

In the final case where we have both measurement error and misclassification in the model, we observed that bias in the naive estimates are higher for higher values of α and β. So in effect the more parameters become important in the model, the greater the effect of measurement error and misclassification on the bias of naive estimators. Surprisingly, the changes in other parameters such as $\mathbf{p}$, ρ, σ_{γ}^{2}, and the sample size seem not to have much effect on the bias of the naive estimators. This we suggest must be subject to further investigation.

Some challenges we encountered in this research are as follows: First, the model is highly restrictive. The feature of conditional probability model forces the lag correlation coefficient to fall in a specific range in order for the model to hold. Due to the restrictive nature of the model, we were limited to the use of small values of parameters and even covariates. We also encountered some computational challenges. By incorporating all the nice features of the model, our estimating equation became quite complicated. We could no more depend on the widely used Newton Raphson's iterative equation to solve the equation numerically. Therefore, we had to resort to Grid search method which was slow and time consuming.

Apart from the above mentioned challenges, we like to acknowledge the following limitations: First, we used fewer covariates- one continuous and one categorical covariate. This consideration was made in order to avoid confusing trend patterns. For the sake of time, we made use of independence assumptions in measurement errors to simplify computations where necessary. Also, could not perform iterations for higher numbers. Furthermore, we were not able to repeat the simulations for different values of the parameters apart from those given in the various scenarios. Therefore the results are limited to values used in the simulations and cannot be generalized for all numbers. In conclusions, we hope this research will serve as a resource for further statistical analysis and help in the development of methodologies for corrections of bias in naive estimators.

Appendices

Appendix A

Marginal moments of Binary response

The unconditional mean of the Binary response is given by:

$$\pi_{it} = \frac{exp(X_{it}'\beta + \gamma_i + G_i\alpha)}{1 + exp(X_{it}'\beta + \gamma_i + G_i\alpha)}, \quad t = 1, \ldots, T,$$

where $\gamma_i \sim N(0, \sigma_\gamma^2)$.

Assuming $f(\gamma_i)$ to be the probability density function of γ_i, then under normallity assumption we have

$f(\gamma_i) = \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} e^{-\frac{\gamma_i^2}{2\sigma_\gamma^2}}, -\infty < \gamma_i < \infty.$

$$\begin{aligned}
E(Y_{it}|X_{it}, G_i) &= E_{\gamma_i}(E(Y_{it}|X_{it}, G_i, \gamma_i)) = \mu_{it} = \int \pi_{it} f(\gamma_i) d\gamma_i \\
Var(Y_{it}|X_{it}, G_i) &= E_{\gamma_i}(Var(Y_{it}|X_{it}, G_i, \gamma_i)) + Var_{\gamma_i}(E(Y_{it}|X_{it}, G_i, \gamma_i)) \\
&= E_{\gamma_i}[\pi_{it} - (\pi_{it})^2] + Var_{\gamma_i}[\pi_{it}] \\
&= E_{\gamma_i}[\pi_{it}] - E_{\gamma_i}[(\pi_{it})^2] + E_{\gamma_i}[(\pi_{it})^2] - (E_{\gamma_i}[\pi_{it}])^2 \\
&= \mu_{it} - (\mu_{it})^2 \\
&= \mu_{it}(1 - \mu_{it})
\end{aligned}$$

Therefore, the marginal response has binary distribution as it is reflected in all the chapters.

Appendix B

Approximation of integrals for unconditional moments

Some authors such as Monahan and Stefanski (1992), and Tao and Fan (2010) have used the probit and logit link functions in Generalized linear Models to approximate integrals associated with the marginal moments of binary response. The moments that are applied in our model are presented here.

Let Φ be cumulative distribution function of standard normal variable, $Z \sim N(0, 1)$, and $\gamma_i \sim N(0, \sigma_\gamma^2)$.

1. By probit approximation:

$$\begin{aligned}
& \int \Phi(X_{it}'\beta + G_i\alpha + \gamma_i) f(\gamma_i) d\gamma_i \\
= \;& P(Z - \gamma_i < x_{it}'\beta + g_i\alpha) \\
= \;& \Phi\left(\frac{X_{it}'\beta + G_i\alpha}{\sqrt{1+\sigma_\gamma^2}}\right)
\end{aligned}$$

where $Z - \gamma_i \sim N(0, \sqrt{1+\sigma_\gamma^2})$

2. If g is the logit link function, then a logit approximation is as follows:

$$\begin{aligned}
\mu_{it} &= \int g(X_{it}'\beta + G_i\alpha + \gamma_i) f(\gamma_i) d\gamma_i \\
&= \int \mathbf{G}(\Phi(X_{it}'\beta + G_i\alpha + \gamma_i)) f(\gamma_i) d\gamma_i \\
&\approx \Phi\left(\frac{X_{it}'\beta + G_i\alpha + \gamma_i}{\sqrt{(1+\sigma_\gamma^2)}}\right) \\
&\approx \frac{\exp\left(\frac{X_{it}'\beta + G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}}\right)}{1+\exp\left(\frac{X_{it}'\beta + G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}}\right)},
\end{aligned}$$

where $k^2 = 1.7$

The approximation uses $\mathbf{G}(.)$ as the 1st order Taylor expansion of $g(.)$ around $\phi = \phi_0$. The final approximation is due to Monahan and Stefanski (1992).

3. Using both the probit and logit approximations above, the unconditional second

order moments are also approximated as:

$$\begin{aligned}
\mu_{iut} &= \int \pi_{iu}\pi_{it} f(\gamma_i) d\gamma_i \\
&= \int \frac{\exp(X'_{iu}\beta + \gamma_i + G_i\alpha)}{1+\exp(X'_{iu}\beta + \gamma_i + G_i\alpha)} \frac{\exp(X'_{it}\beta + \gamma_i + G_i\alpha)}{1+\exp(X'_{it}\beta + \gamma_i + G_i\alpha)} f(\gamma_i) d\gamma_i \\
&\approx \frac{\exp(\frac{X'_{iu}\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})}{1+\exp(\frac{X'_{iu}\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})} \frac{\exp(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})}{1+\exp(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})} \\
&+ \frac{\exp(X'_{iu}\beta + G_i\alpha)}{(1+\exp(X'_{iu}\beta + G_i\alpha))^2} [\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{iu}\beta + G_i\alpha)^2}{2}))]^{-1} \\
&\quad \frac{\exp(X'_{it}\beta + G_i\alpha)}{(1+\exp(X'_{it}\beta + G_i\alpha))^2} [\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{it}\beta + G_i\alpha)^2}{2}))]^{-1} \\
&\quad [P(Z_1 - \gamma_i < X'_{iu}\beta + G_i\alpha, Z_2 - \gamma_i < X'_{it}\beta + G_i\alpha) \\
&\quad -P(Z_1 - \gamma_i < X'_{iu}\beta + G_i\alpha) P(Z_2 - \gamma_i < X'_{it}\beta + G_i\alpha)]
\end{aligned}$$

$$\begin{aligned}
\mu_{iuu} &= \int \pi_{iu}\pi_{iu} f(\gamma_i) d\gamma_i \\
&= \int [\frac{\exp(X'_{iu}\beta + \gamma_i + G_i\alpha)}{1+\exp(X'_{iu}\beta + \gamma_i + G_i\alpha)}]^2 f(\gamma_i) d\gamma_i \\
&\approx [\frac{\exp(\frac{X'_{iu}\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})}{1+\exp(\frac{X'_{iu}\beta+G_i\alpha}{\sqrt{1+\sigma_\gamma^2/k^2}})}]^2 \\
&+ [\frac{\exp(X'_{iu}\beta + G_i\alpha)}{(1+\exp(X'_{iu}\beta + G_i\alpha))^2} [\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{iu}\beta + G_i\alpha)^2}{2}))]^{-1}]^2 \\
&\quad P(Z - \gamma_i < X'_{iu}\beta + G_i\alpha) - P(Z - \gamma_i < X'_{iu}\beta + G_i\alpha)^2
\end{aligned}$$

μ_{it}, μ_{iut} *and* μ_{iuu} were used in Chapter 2.

Appendix C

Approximations for model in Chapter 3

When the variable W is observed instead of X, we can obtain approximations similar to those in Chapter 1 as follows:

$$\begin{aligned}\mu^*_{1it} &= \int g(W'_{it}\beta + G_i\alpha + \gamma_i)f(\gamma_i)d\gamma_i \\ &\approx \Phi(\frac{W'_{it}\beta + G_i\alpha + \gamma_i}{\sqrt{(1+\sigma^2_\gamma)}})\end{aligned}$$

Alternatively, by computing the expectations iteratively we have the terms as in Chapter 2:

$$\begin{aligned}
\mu^*_{1it} &= E_{X|W}\left(\frac{\exp(\frac{X'_{it}\beta+G_i\alpha}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}})}{1+exp(\frac{X'_{it}\beta+G_i\alpha)}{\sqrt{1+\sigma^2_{\gamma}/k^2}})}, |W_{it}, G_i\right) \qquad (C.1)\\
&= E(\mu_{it}|W_{it}, G_i)
\end{aligned}$$

However, these expressions do not have closed forms.

Appendix D

Approximations for model in Chapter 4

By taking iterative expectations of $\mu_{it}, \mu_{iu}, \mu_{iut}$ *and* μ_{iuu} from Chapter 1, we can get $\mu^*_{2it}, \mu^*_{2iu}, \mu^*_{2iut}$ *and* μ^*_{2iuu} for the model in chapter 3.

$$\begin{aligned}
\mu^*_{2it} &= E_{G|G^*}(\mu_{it}|X_{it}, G^*_i) \\
&\approx \left(\frac{\exp(\frac{X'_{it}\beta}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}})}{1+\exp(\frac{X'_{it}\beta}{\sqrt{1+\sigma^2_{\gamma}/k^2}})}\right).(1-\lambda(1|G^*)) \\
&+ \left(\frac{\exp(\frac{X'_{it}\beta+\alpha}{\sqrt{1+\sigma^2_{\gamma_i}/k^2}})}{1+\exp(\frac{X'_{it}\beta+\alpha}{\sqrt{1+\sigma^2_{\gamma}/k^2}})}\right).\lambda(1|G^*) \qquad \text{(D.1)}
\end{aligned}$$

Similarly, μ^*_{2iu} is obtained by substituting X_{iu} in (D.1)

$$
\begin{aligned}
\mu^*_{2iuu} &= E_{G|G^*}(\mu_{iuu}|X_{iu}, G^*_i) \\
&\approx [\frac{\exp(\frac{X'_{iu}\beta}{\sqrt{1+\sigma^2_\gamma/k^2}})}{1+\exp(\frac{X'_{iu}\beta}{\sqrt{1+\sigma^2_\gamma/k^2}})}]^2 \\
&+ ([\frac{\exp(X'_{iu}\beta)}{(1+\exp(X'_{iu}\beta))^2}[\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{iu}\beta)^2}{2}))]^{-1}]^2 \\
&\quad P(Z-\gamma_i < X'_{iu}\beta) - P(Z-\gamma_i < X'_{iu}\beta)^2).(1-\lambda(1|G^*) \\
&+ ([\frac{\exp(\frac{X'_{iu}\beta+\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})}{1+\exp(\frac{X'_{iu}\beta+\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})}]^2 \\
&+ [\frac{\exp(X'_{iu}\beta+\alpha)}{(1+\exp(X'_{iu}\beta+\alpha))^2}[\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{iu}\beta+\alpha)^2}{2}))]^{-1}]^2 \\
&\quad P(Z-\gamma_i < X'_{iu}\beta+\alpha) - P(Z-\gamma_i < X'_{iu}\beta+\alpha)^2)(\lambda(1|G^*))
\end{aligned}
$$

$$
\begin{aligned}
\mu^*_{2iut} &= E_{G|G^*}(\mu_{iut}|X_i, G^*_i) \\
&\approx (\frac{\exp(\frac{X'_{iu}\beta}{\sqrt{1+\sigma^2_\gamma/k^2}})}{1+\exp(\frac{X'_{iu}\beta}{\sqrt{1+\sigma^2_\gamma/k^2}})}\frac{\exp(\frac{X'_{it}\beta}{\sqrt{1+\sigma^2_\gamma/k^2}})}{1+\exp(\frac{X'_{it}\beta}{\sqrt{1+\sigma^2_\gamma/k^2}})} \\
&+ \frac{\exp(X'_{iu}\beta)}{(1+\exp(X'_{iu}\beta))^2}[\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{iu}\beta)^2}{2}))]^{-1} \\
&\quad \frac{\exp(X'_{it}\beta)}{(1+\exp(X'_{it}\beta))^2}[\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{it}\beta)^2}{2}))]^{-1} \\
&\quad [P(Z_1-\gamma_i < X'_{iu}\beta, Z_2-\gamma_i < X'_{it}\beta) \\
&\quad -P(Z_1-\gamma_i < X'_{iu}\beta)P(Z_2-\gamma_i < X'_{it}\beta)]).(1-\lambda(1|G^*)^2 \\
&+ (\frac{\exp(\frac{X'_{iu}\beta+\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})}{1+\exp(\frac{X'_{iu}\beta+\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})}\frac{\exp(\frac{X'_{it}\beta+\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})}{1+\exp(\frac{X'_{it}\beta+\alpha}{\sqrt{1+\sigma^2_\gamma/k^2}})} \\
&+ \frac{\exp(X'_{iu}\beta+\alpha)}{(1+\exp(X'_{iu}\beta+\alpha))^2}[\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{iu}\beta+\alpha)^2}{2}))]^{-1} \\
&\quad \frac{\exp(X'_{it}\beta+\alpha)}{(1+\exp(X'_{it}\beta+\alpha))^2}[\frac{1}{2\sqrt(\pi)}(\exp(\frac{-(X'_{it}\beta+\alpha)^2}{2}))]^{-1} \\
&\quad [P(Z_1-\gamma_i < X'_{iu}\beta+\alpha, Z_2-\gamma_i < X'_{it}\beta+\alpha) \\
&\quad -P(Z_1-\gamma_i < X'_{iu}\beta+\alpha)P(Z_2-\gamma_i < X'_{it}\beta+\alpha)])(\lambda(1|G^*))^2
\end{aligned}
$$

Bibliography

[1] Abarin, T., Li H., Wang L. and Briollais L. (2012). On Method of Moments Estimation in Linear Mixed Effects Models with Measurement Error on Covariates and Response with Application to a Longitudinal Study of Gene-Environment Interaction. *Stat Biosci. DOI* 10.1007/s12561-012-9074-5.

[2] Batistatou E.and McNamee R. (2012). Performance of bias correction methods for exposure measurement error using repeated measurements with and without missing data. *Statistics in medicine*, **31**, (28), 3467-3480.

[3] Breslow N.E., and Clayton D.G.,(1993) Approximate inference in generalized linear mixed models. *Journal of American Statistical Assoc.*, **88**, 9-25.

[4] Buonaccorsi J.P.(2010) *Measurement Error*, Chapman and Hall, New York.

[5] Buzas, Jeffrey S., Leonard A., Stefanski L.A., Tosteson T.D., (2014) *Measurement error." Handbook of Epidemiology* 1241-1282.

[6] Carroll R.J. and Stefanski L.A.(1985) Covariate measurement error in logistic regression. *Ann. Stat.*, **13**, 1335-1351.

[7] Cook J. and Stefanski L.A. (1994) A simulation extrapolation method for parametric measurement error model. *Journal of the American Statistical Association*, **80**, 1314-1328.

[8] Diggle P.J., Heagety P., Liang K.Y., and Zeger S.L(2002) *Analysis of Longitudinal Data*, Oxford Clarendon Press.

[9] Eisenhower D., Mathiowetz N. A. and Morganstein D. (1991). Recall error: Sources and bias reduction techniques. *Measurement errors in surveys*, **27**, 125-144.

[10] Fan Z., Sutradhar B.C. and Rao R.P.(2012) Bias corrected generalised method of moment and generalised quasi-likelihood inference in linear models for panel data with measurement error, *Sankhyia B*, **74**, 126-148.

[11] Fitzmaurice G.,Davadian M., Verberke G. and Molenberghs G. (2009) *Longitudinal Data Analysis*, London, Chapman and Hall/CRC.

[12] Fuller W.A. (1987) *Measurement Error Models*, Willey series, New York.

[13] Goldstein(1991) Nonlinear Multi-level models, with application to discrete response data, *Biometrika*, **78**, 45-51.

[14] Ji T. and Fan Z.(2010)*Analysis of Longitudinal categorical count data subject to measurement error*. Doctorial Thesis, Memorial University.

[15] Laird N.M. and Ware J.H. (1982) Random effects models for longitudinal data. *Biometrics* , **38**, 963-974.

[16] Liang K.Y. and Zeger S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

[17] Monahan and Stefanski L.A.(1991) Normal Scale mixture distrributions and computation of logistic normal Integral. *In handbook of logistic distribution*, Balakrisbose 529-540, Marcel Dekker, New York.

[18] Nelder,J.A. and Wedderburn R.W.M. (1972). Generalized Linear Models.*Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

[19] McCullagh P. and Nelder J. A. (1983)*Generalized Linear Models*. London: Chapman and Hall

[20] Raymond J., David R., Stefanski L. A. and Ciprian M. C.(2005) *Measurement Error in Nonlinear Model*, 2nd Ed, Chapman and Hall, New York.

[21] Spiegelman D., McDermott A., and Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American journal of clinical nutrition, 1179S-1186S.*

[22] Sutradhar, Brajendra C (2012). *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers.* New York: Springer New York, LLC.

[23] Sutradhar B.C.(2011) *Dynamic Mixed Models for Familiar Longitudinal Data*,Springer, New York.

[24] Thomas L.E., Stefanski L.A. and Davidian M.(2011) Moment adjusted imputation for measurement error models. *Biometrics*, **67**, 1461-1470.

[25] Wedderburn R.W.M (1974) Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* , **61**, 439-447.

[26] Greene W. and Cai J. (2004) Measurement error in the Marginal Hazard model for Multivariate Failure Time data.*Biometrics*, **60**, (4), 787-946.

[27] Xihong L. and Raymond J. (1999) SIMEX Variance component in Generalized Mixed Measurement Error Model, *Biometrics*, **55**, (2), 613-619.

[28] Tao Y. and Fan Z. (2010) *Generalized Quasilikelihood method for Misclassified Longitudinal Binary Data.* Master Thesis, Memorial University.