GENERALIZED QUASI-LIKELIHOOD METHOD FOR LONGITUDINAL BINARY DATA WITH MEASUREMENT ERROR

YUHUA ZHANG

Generalized Quasi-likelihood Method for Longitudinal Binary Data with Measurement Error

by

©Yuhua Zhang

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

Master of Statistics

Department of Mathematics and Statistics

Memorial University of Newfoundland

December, 2011

St. John's Newfoundland, Canada

Abstract

In this thesis, we propose an approach to correct the estimation of the bias of the model parameters when using a generalized quasi-likelihood method to analyze longitudinal binary data with measurement errors. The measurement errors are assumed to follow a normal distribution with an unknown variance, which can be estimated by repeated observations or taken from previous similar studies. An approximation method proposed by Monahan and Stefanski (1992)is used to obtain the expectation of an unknown function involved in the calculation of the means and covariance, which will be used later to construct the estimating functions of the GQL. A simulation study is carried out in the aim of investigating the small sample performance of the proposed approach. The results of an intensive simulation study show that the proposed approach works very well in all configurations. The efficiency gain of the proposed method, as compared to the naive use of GQL is remarkable. The proposed method has great potential to be widely used to analyze data from social, economical and biomedical studies.

Keywords: Generalized quasi-likelihood, Longitudinal binary data, Measurement error.

Acknowledgements

I would like to thank my supervisor, Dr. Zhaozhi Fan, for his guidance, support, supervision and encouragement in my program of study, which was supported by the Department of Mathematics and Statistics and School of Graduate Studies.

Many thanks go to Dr. Brajendra C. Sutradhar, Dr. Concepcion Loredo-Osti and Dr. Asokan Variyath, who taught me courses for my program. I also want to thank Dr.Hong Wang for his kind support and help when I was a teaching assistant.

I also want to express my appreciation to Dr. Yunqi Ji, Mr. Yi Tao, Mr. Haiyan Yang and Mr. Derek Strong for their friendships and help during my study.

Special thanks go to my wife, Feng Gao, and my daughter, Tingzhi Zhang. Without their love and support, I would not be able to complete my study so smoothly.

I am also grateful to the School of Graduate Studies and the Department of Mathematics and Statistics for the financial support during my program. Also I want to thank all faculty members and staff in the department for their kindness and help.

Contents

List of Tables

Chapter 1

Introduction

Binary responses along with a set of multi-dimensional covariates are often collected repeatedly over a short period of time from a large number of independent individuals, which is called longitudinal binary data. Longitudinal binary data often appear in a wide range of areas such as public health, medicine, economics, sociology, and so on. The covariates in longitudinal data may be time-dependent or time-independent. Often times the main focus of the study is to evaluate the effects of covariates. The dynamics among the response variables over time is also of significant scientific interest. Actually, the repeated observations in a longitudinal study allow us to estimate both the effects of covariate variables and the pattern in the response variables over time, say the cohort effects. There is a considerable demand for adequate methods for the evaluation of data of this type in applications. Great attentions were drawn recently among the statisticians. Literatures include but not limited to Diggle et al. (2002), Dunlop (1994), Qu and Song (2002) and Ware (1985) among others.

Although most studies are well designed to obtain accurate informa-

tion, measurement errors in discrete data still occur due to many known and unknown reasons such as imperfect instruments and procedures, limited knowledge and experience of examiners and examinees, extremely high cost of getting exact measurement, and so on.

Measurement errors also occur in continuous data. Covariates in generalized linear models are frequently subject to measurement error, for instance in epidemiology studies where the effects of lifetime exposure to pollutants, alcohol, exercise and so on are often of scientific interest.

Explicit measurement error modeling is crucial for at least two reasons: First, neglecting measurement error will often lead to biased estimates for the regression parameters. For instance, it is well known that ordinary logistic regression can lead to biased estimates of odds ratios when the covariates are subject to measurement error (Rosner et al. (1990)), a phenomenon known as regression dilution in the simple case of a single covariate. Joint modeling of the response and measurement process allows estimation of a dis-attenuated odds ratio for the true covariate (see, for example, Carroll et al. (1995)). Secondly, measurement error modeling facilitates prediction of the true covariate or exposure for an individual unit, utilizing not only the exposure measurements for the unit but also information from the outcome as well as borrowing strength from the other units.

Much has been carried out on measurement error models for continuous data, for example, the classical additive measurement error models, the Berkson error model (Fuller (1987); Carroll et al. (2006); Buzas, Tosteson, and Stefanski (2003)), equation error model (Kipnis et al. (1999); Kipnis et al. (2003)), and regression calibration model (Mallick and Gelfand

(1996)) among others.

Rosychuk (1999) studied the estimation bias of the covariate effects when ignoring errors in response. Magder and Hughes (1997) proposed an EM approach to the inference of model effects. The model of Caroll, Maca and Ruppert (1999) handles complex models with a simple measurement error structure. Neuhaus (1999) proposed a computationally more efficient approximation to accommodate measurement error when estimating the model effects. Roy et al. (2009) proposed a model-based approach to the case of misclassification. Gustafson (2003), McGlothin et al. (2008), and Rosychuk et al. (2009) also proposed different approaches to correcting the bias of the estimation.

For correlated case such as longitudinal binary response data, the measurement error in covariates is even more difficult to handle, mainly due to the complex nature of the likelihood when complex correlation structure is involved. Ji (2011) and Tao (2010) investigated the GQL and MLE approaches for a kind of dynamical binary response model, taking measurement error and misclassification into account. The simulation results show remarkable efficiency gain by appropriately modeling the misclassification.

As mentioned previously, measurement errors frequently occur in covariates from studies in epidemiology, medicine, economics, and sociology. Simply ignoring measurement errors in covariates leads to biased estimation of model parameters and loss of power in detecting interesting association among variables. In order to improve the estimation of parameters, in this thesis we propose a new approach that combines the approximation method of Monahan and Stefanski (1992) (also see Roy, Baneriee

and Maiti (2005)) and the generalized quasi-likelihood approach by Sutradhar (2003). An approximate generalized quasi-likelihood method for longitudinal binary data is developed to correct the estimation bias of the regression parameters in the presence of measurement error in covariate.

The logistic mixed effects models (Sutradhar and Farrell (2007)) can be applied to analyze longitudinal binary data. When measurement errors are not ignorable, however, the likelihood function is usually very difficult to compute (Sutradhar and Mukerjee (2005)). The conditional inference approaches can also be applied to longitudinal binary data (Breslow and Clayton (1993), Sutradhar (2004)). The integration over the distinction of the measurement error is difficult for logistic cases, especially for multivariate measurement errors (Monahan and Stefanski (1992), Tao (2010)).

To avoid the complexity of the likelihood function and the short coming of conditional inferences, we exploit the generalized quasi-likelihood method (GQL), which has been proven to be almost as efficient as MLE for binary data modeling (Sutradhar and Farrell (2007)).

We focus on the unconditional generalized quasi-likelihood inference that involves unconditional moments of up to second order. The integrations are approximated by using the method of Monahan and Stefanski (1992).

By doing this we could avoid any extra distributional assumption on the correlated binary response. The method is hence widely applicable.

The thesis is organized as follows. We develop a regression model for longitudinal panel data with measurement error in Chapter 2. In Chapter 3, we provide the summarization of the logistic regression with logit link and probit link and we also introduce the covariance matrix and corrected

generalized quasi-likelihood (CGQL) method, which extends and unifies the previous work, generalized quasi-likelihood by Sutradhar (2003), and logit link approaches by Roy, Baneriee and Maiti (2005). The emphasis of this study is the bias correction of model effects when estimating equations are constructed on quasi-likelihood. The calculation and approximation are also detailed in Chapter 3. A simulation study is conducted in Chapter 4 to investigate the performance of the proposed method when the sample size is properly chosen. Conclusions and discussions are provided in Chapter 5.

Chapter 2

The Linear Dynamic Model for longitudinal Data

In this chapter, we briefly review the longitudinal models for binary data and then extend the models to include measurement errors.

Let $y_i = (y_{i1}, \cdots, y_{it}, \cdots, y_{iT})'$, be the vector of T repeated binary responses for $i = 1, \cdots, N$ and $\mathbf{x} = (\mathbf{x_1'}, \mathbf{x_2'})'$, where $\mathbf{x_1}(p_1 \times 1)$ and $\mathbf{x_2}(p_2 \times 1)$, are the covariates of interest. Further we assume that $\mathbf{x_1}$ is observed without measurement error and $\mathbf{x_2}$ is not observable, however its surrogate $\mathbf{z}$, is available. The effects of the error prone covariate $\mathbf{x_2}$ are of major scientific interests.

2.1 Regression model without measurement error

For a fixed $\mathbf{x_{it}} = (x_{1it}, x_{2it})'$, it is assumed that

$$P(y_{it} = 1 | x_{1it}, x_{2it}) = g(x_{1it}\beta_1 + x_{2it}\beta_2) \tag{2.1}$$

for $t = 1, \cdots, T$ and $i = 1, \cdots, N$, where $g(v) = (1 + exp(-v))^{-1}$.

We use μ_{it}^* to denote the mean of Y, with predictors without measurement error

$$\mu_{it}^* = E(Y_{it}|x_{1i}, x_{2i}) = P(y_{it} = 1|x_{1it}, x_{2it}) \tag{2.2}$$

Next suppose that y_{i1} has a Bernoulli distribution with mean parameter μ_{i1}^* denoted by $y_{i1} \sim bin(\mu_{i1}^*)$. In a dynamic linear model set up, we use the model (Tong (1990), Table 3.1, p113) given by

$$y_{it} = q_{1t}y_{it-1} + (1 - q_{1t})q_{2t} \tag{2.3}$$

where $q_{1t} \sim bin(p^*)$, p^* is the mixture probability parameter of the distribution of variable q_{1t}, $q_{2t} \sim bin(\mu_{it}^*)$, q_{1t}, q_{2t} are independent.

The conditional probabilities are given by

$$P(y_{it} = 1|y_{it-1}, x_{1it}, x_{2it}) = p^*y_{it-1} + (1 - p^*)\mu_{it}^* \tag{2.4}$$

for $t = 2, \cdots, T$, $i = 1, \cdots, N$ and $0 < p^* < 1$. It then follows that $y_{it} \sim bin(1, \mu_{it})$, where

$$\begin{aligned} \mu_{it} &= p^*\mu_{it-1} + (1 - p^*)\mu_{it}^* \\ &= p^{*t-1}\mu_{i1} + (1 - p^*)\sum_{j=2}^{t} p^{*t-j}\mu_{ij} \end{aligned} \tag{2.5}$$

For $t = 1$, the binary response y_{i1} has been assumed to have mean $\mu_{i1} = \mu_{i1}^*$.

2.2 Regression model with measurement errors in covariates

Now we consider the dynamic model with measurement errors in covariates. Suppose that $\mathbf{x_2}$ is not observable, however its surrogate $\mathbf{z}$, say the observation of $\mathbf{x_2}$, is available, and they have the following relationship

$$x_{2i} = z_i + \epsilon_i \tag{2.6}$$

where ϵ_i, $i = 1, \cdots, N$, follow a normal distribution, i.e.

$$\epsilon_i \sim N(0, \sigma_i^2). \tag{2.7}$$

It is also assumed that the errors in the variables model are non-differential, that is,

$$f(y_{it}|x_{1it}, x_{2it}, z_{it}) = f(y_{it}|x_{1it}, x_{2it}) \tag{2.8}$$

In other words, z_{it} adds no additional information to the prediction of y if x_{2it} is known.

Regarding measurement error process, we assume

$$x_{2it}|z_{it} \sim N(z_{it}, \sigma^2) \quad for \quad i = 1, \cdots, I, \quad t = 1, \cdots, 4 \tag{2.9}$$

Where z_{it}, σ^2 are known.

We use μ_{it}^{*e} to denote the mean of Y, where the covariates are contam-

inated with measurement errors.

$$\begin{aligned}
\mu_{it}^{*e} &= E(Y_{it}|x_{1it}, z_{it}) \\
&= P(y_{it} = 1|x_{1it}, z_{it}) \\
&= E_{x_{2it}|z_{it}}(E(Y_{it}|x_{1it}, x_{2it}, z_{it})) \\
&= E_{x_{2it}|z_{it}}(E(Y_{it}|x_{1it}, x_{2it})) \\
&= \int_{-\infty}^{+\infty} g(x_{1it}\beta_1 + z_{it}\beta_2 + \beta_2\epsilon_i) f(\beta_2 x_{2it}|z_{it}) dx_{2it} \qquad (2.10)
\end{aligned}$$

for $t = 1, \cdots, T$ and $i = 1, \cdots, N$. The integral in (2.10) does not have a closed form solution. The approximation of this integral was considered by Monahan and Stefanski (1992). With some algebra they derived an approximation to the integral in (2.10) as

$$\begin{aligned}
\mu_{it}^{*e} &= E(Y_{it}|x_{1it}, z_{it}) \\
&= \int_{-\infty}^{+\infty} g(x_{1it}\beta_1 + z_{it}\beta_2 + \beta_2\epsilon_i) f(\beta_2 x_{2it}|z_{it}) dx_{2it} \\
&\approx g\left(\frac{x_{1it}\beta_1 + z_{it}\beta_2}{\sqrt{1 + \frac{\beta_2^2\sigma_i^2}{k^2}}}\right) \qquad (2.11)
\end{aligned}$$

for $t = 1, \cdots, T$ and $i = 1, \cdots, N$, where $g(\upsilon) = (1 + exp(-\upsilon))^{-1}$ and $k^2 = 1.70$. In most cases this gives a good approximation except when $\frac{\beta_2^2\sigma_i^2}{k^2}$ is large. The details will be discussed in the next subsection.

We still use the linear binary dynamic model given by

$$y_{it} = q_{1t}y_{it-1} + (1 - q_{1t})q_{2t} \qquad (2.12)$$

where $q_{1t} \sim bin(p^*)$, p^* is the mixture probability parameter of the distri-

bution of variable q_{1t}, $q_{2t} \sim bin(\mu_{it}^{*e})$, and q_{1t}, q_{2t} are independent.

The conditional probabilities are given by

$$P(y_{it} = 1|y_{it-1}, x_{1it}, z_{it}) = p^* y_{it-1} + (1 - p^*)\mu_{it}^{*e} \tag{2.13}$$

for $t = 2, \cdots, T$ and $i = 1, \cdots, N$. It then follows that $y_{it} \sim bin(1, \mu_{it}^e)$, where

$$\begin{aligned} \mu_{it}^e &= p^* \mu_{it-1}^e + (1 - p^*)\mu_{it}^{*e} \\ &= p^{*t-1}\mu_{i1}^e + (1 - p^*)\sum_{j=2}^{t} p^{*t-j}\mu_{ij}^{*e} \end{aligned} \tag{2.14}$$

For $t = 1$, the binary response y_{i1} has been assumed to have mean $\mu_{i1}^e = \mu_{i1}^{*e}$.

2.3 Computation of the probit-normal integral and logit-normal integral

With the normality assumption for the measurement error process we have

$$\begin{aligned} P(y_{it} = 1|x_{1it}, z_{it}) &= E_{x_{2it}|z_{it}}(E(Y_{it}|x_{1it}, x_{2it}, z_{it})) \\ &= E_{x_{2it}|z_{it}}(E(Y_{it}|x_{1it}, x_{2it})) \\ &= E_{x_{2it}|z_{it}} g(x_{1it}\beta_1 + x_{2it}\beta_2) \\ &= \int_{-\infty}^{+\infty} g(x_{1it}\beta_1 + z_{it}\beta_2 + \epsilon_i\beta_2) f(\epsilon_i\beta_2) d\epsilon_i \end{aligned} \tag{2.15}$$

where $f(\epsilon_i\beta_2)$ is the probability density function of $\epsilon_i\beta_2$.

We use μ_{it}^{*e} to denote the mean of Y, where the covariates have measurement errors, say,

$$\begin{aligned} \mu_{it}^{*e} &= E(Y_{it}|x_{1i}, z_i) = P(y_{it} = 1|x_{1it}, z_{it}) \\ &= \int_{-\infty}^{+\infty} g(x_{1it}\beta_1 + z_{it}\beta_2 + \epsilon_i\beta_2) f(\epsilon_i\beta_2) d\epsilon_i \end{aligned} \tag{2.16}$$

In this formula above, the link function $g(.)$ is nonlinear. The calculation of the involved integral is very hard. However the close relationship between logit link function and probit link function was already discovered in the early work (Eugene (2004), p334). There are several ways to approximate the logit link by probit link when the aforementioned integral is concerned, for instance, Roy, Banerjee, and Maiti (2005) used a method of approximation for probit link function and logit link function to deal with the measurement error model. Now we introduce this approach.

Let $g(.)$ be the logit link function, Φ be the distribution function of standard normal random variable. Suppose G is a function of g, as suggested by Eugene (2004, p335), then we have:

$$\begin{aligned} g(x_{1iu}\beta_1 + z_{iu}\beta_2 - \beta_2\epsilon_i) &= \frac{exp(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)}{1 + exp(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)} \\ &\approx G(\Phi(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)) \\ &= G(\Phi_1) \end{aligned} \tag{2.17}$$

where $u = 1, \cdots, T$. Now we can rewrite the function as below:

$$g(x_{1i}\beta_1 + x_{2i}\beta_1) = \frac{exp(x_{1i}\beta_1 + x_{2i}\beta_1)}{1 + exp(x_{1i}\beta_1 + x_{2i}\beta_1)} \approx G(\Phi(x_{1i}\beta_1 + x_{2i}\beta_1)) \tag{2.18}$$

2.3.1 Probit link

Let Φ be the cumulative distribution function of a standard normal distribution. We calculate the integral as follows

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \Phi(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) f(\beta_2\epsilon_i) d\epsilon_i \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i} \frac{1}{\sqrt{2\pi}} exp(-\frac{w_1^2}{2}) dw_1 \frac{1}{\beta_2\sigma_i\sqrt{2\pi}} exp(-\frac{\epsilon_i^2}{2\beta_2^2\sigma_i^2}) d\epsilon_1 \\
&= P(-\infty < \beta_2\epsilon_i < +\infty, w_1 < x_{1iv}\beta_1 + z_{iv}\beta_2 - \beta_2\epsilon_i) \\
&= P(w_1 - \beta_2\epsilon_i < x_{1iv}\beta_1 + z_{iv}\beta_2)
\end{aligned} \tag{2.19}$$

In the equation above, the two-dimensional integral may be viewed as the probability that the sum two independent normally distributed random variables w_1 and $\beta_2\epsilon_i$, where $w_1 \sim N(0,1)$ and $\beta_2\epsilon_i \sim N(0, \beta_2^2\sigma_i^2)$, hence $w_1 - \beta_2\epsilon_i \sim N(0, (1 + \beta_2^2\sigma_i^2))$ is less than $x_{1iv}\beta_1 + z_{iv}\beta_2$.

Therefore, we have

$$\begin{aligned}
&P(w_1 - \beta_2\epsilon_i < x_{1iv}\beta_1 + z_{iv}\beta_2) \\
&= \Phi(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{(1 + \beta_2^2\sigma_i^2)}})
\end{aligned} \tag{2.20}$$

Thus we obtain the equation

$$\int_{-\infty}^{+\infty} \Phi(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) f(\beta_2\epsilon_i) d\epsilon_i = \Phi(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{(1+\beta_2^2\sigma_i^2)}}) \tag{2.21}$$

2.3.2 Logit link

Define $\Phi = \Phi(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)$, $\Phi_0 = \Phi(x_{1iu}\beta_1 + z_{iu}\beta_2)$. Therefore, by using Taylor series expansion, we have

$$G(\Phi) \approx G(\Phi_0) + G(\Phi_0)'(\Phi - \Phi_0) \tag{2.22}$$

When logit link is used in the generalized linear model, we have the following term

$$\begin{aligned}
& \int_{-\infty}^{+\infty} G(\Phi) f(\beta_2\epsilon_i) d\epsilon_i \\
\approx\ & \int_{-\infty}^{+\infty} [G(\Phi_0) + G(\Phi_0)'(\Phi - \Phi_0)] f(\beta_2\epsilon_i) d\epsilon_i \\
=\ & G(\Phi_0) + G(\Phi_0)'|_{\Phi=\Phi_0} \int_{-\infty}^{+\infty} \Phi f(\beta_2\epsilon_i) d\epsilon_i - G(\Phi_0)'|_{\Phi=\Phi_0}\Phi_0 \\
=\ & G(\Phi_0) + G(\Phi_0)'|_{\Phi=\Phi_0}\Phi - G(\Phi_0)'|_{\Phi=\Phi_0}\Phi_0 \\
=\ & G(\Phi_0) + G(\Phi_0)'|_{\Phi=\Phi_0}(\Phi - \Phi_0) \\
\approx\ & G(\Phi) \\
=\ & \Phi(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
\approx\ & \frac{exp(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\frac{\beta_2^2\sigma_i^2}{k^2}}})}{1 + exp(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\frac{\beta_2^2\sigma_i^2}{k^2}}})}
\end{aligned} \tag{2.23}$$

The last approximation is due to Monahan and Stefanski(1992), where

$k^2 = 1.7$ and $\int_{-\infty}^{+\infty} \Phi f(\beta_2 \epsilon_i) d\epsilon_i = \Phi(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})$

In principle, the expectation of Y can be well approximated by

$$\begin{aligned} \mu_{it}^{*e} &= E(Y_{it}|x_{1it}, z_{it}) \\ &= \int_{-\infty}^{+\infty} g(x_{1it}\beta_1 + z_{it}\beta_2 + \beta_2\epsilon_i) f(\beta_2 x_{2it}|z_{it}) dx_{2it} \\ &= g(\frac{x_{1it}\beta_1 + z_{it}\beta_2}{\sqrt{1 + \frac{\beta_2^2\sigma_i^2}{k^2}}}) \end{aligned} \tag{2.24}$$

for $t = 1, \cdots, T$ and $i = 1, \cdots, N$, where $g(v) = (1 + exp(-v))^{-1}$ and $k^2 = 1.70$. In most cases this gives a good approximation except when $\beta_2^2\sigma_i^2$ is large.

These approximations will be used in the next chapter, when we use the generalized quasi-likelihood (GQL) method to estimate the regression coefficients. The method needs μ_{it}^{*e} and the second order moments to formulate the estimation equations.

Chapter 3

Estimation of the Parameters for the Regression Model

In order to correct the bias of regression parameters β caused by ignoring the covariate measurement errors, we apply a corrected generalized quasi-likelihood method (CGQL) to estimate the unknown parameters. The most important and challenging part of this chapter is the unconditional generalized quasi-likelihood inference which involves unconditional moments of up to the second order.

As previously mentioned, the goal of this thesis is to eliminate the estimation bias by using the CGQL. A comparison of the proposed estimates with the estimates from the naive generalized quasi-likelihood method (NGQL), which ignores the covariate measurement error will play an important role.

3.1 Naive generalized quasi-likelihood method

Let $y_i = (y_{i1}, \cdots, y_{iT})'$ denote the longitudinal observation of binary response, $\mu_i = (\mu_{i1}, \cdots, \mu_{iT})'$ be the vector of the means of Y_i and Σ_i be the $T \times T$ covariance matrix of Y_i. Sutrahar B. C. (2003) proposed a generalized quasi-likelihood (GQL) method to estimate the regression parameters β by solving the following estimating equations:

$$\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{-1} (y_i - \mu_i) = 0 \tag{3.1}$$

where $\theta = (\beta, p^*)'$, $\beta = (\beta_1', \beta_2')'$ and $\frac{\partial \mu_i'}{\partial \theta} = (\frac{\partial \mu_{i1}}{\partial \theta}, \cdots, \frac{\partial \mu_{iT}}{\partial \theta})$ is the $(p+1) \times T$ first derivative matrix of means μ_i, which is given by:

For $t = 1$,

$$\begin{aligned} \frac{\partial \mu_{i1}}{\partial \beta_p} &= \frac{\partial \mu_{i1}^*}{\partial \beta_p} \\ \frac{\partial \mu_{i1}}{\partial p^*} &= 0 \end{aligned} \tag{3.2}$$

where $p = 1, 2$.

For $t = 2, \cdots, T$,

$$\begin{aligned} \frac{\partial \mu_{it}}{\partial \beta_p} &= p^* \frac{\partial \mu_{it-1}}{\partial \beta_p} + (1 - p^*) \frac{\partial \mu_{it}^*}{\partial \beta_p} \\ \frac{\partial \mu_{it}}{\partial p^*} &= \mu_{it-1} - \mu_{it}^* \end{aligned} \tag{3.3}$$

where $i = 1, \cdots, N$ and $p = 1, 2$.

Σ_i be the $T \times T$ covariance matrix of Y_i, and have the form

$$\Sigma_i = \begin{bmatrix} var(Y_{i1}) & cov(Y_{i1}, Y_{i2}) & \cdots & cov(Y_{i1}, Y_{iT}) \\ & var(Y_{i2}) & \cdots & cov(Y_{i2}, Y_{iT}) \\ & & \ddots & \cdots \\ & & & var(Y_{iT}) \end{bmatrix} \tag{3.4}$$

The formulas for the components of this covariance matrix Σ_i is given as follows.

The diagonal elements of covariance matrix Σ_i are

$$\begin{aligned} var(Y_{it}) &= E(Y_{it}^2) - (E(Y_{it}))^2 \\ &= E(Y_{it}) - (E(Y_{it}))^2 \\ &= \mu_{it} - \mu_{it}^2 \\ &= \mu_{it}(1 - \mu_{it}) \end{aligned} \tag{3.5}$$

where $E(Y_{it}^2 = E(Y_{it})$ as Y_{it} follows Bernoulli distribution, for $t = 1, \cdots, T$, $i = 1, \cdots, N$.

$$\begin{aligned} cov(Y_{iu}, Y_{iv}) &= E(Y_{iu}Y_{iv}) - E(Y_{iu})E(Y_{iv}) \\ &= E(Y_{iu}Y_{iv}) - \mu_{iu}\mu_{iv} \end{aligned} \tag{3.6}$$

for $u = 1, \cdots, (T-1)$, $v = (u+1), \cdots, T$, $i = 1, \cdots, N$. where the second order moments have the formula, for $u < v$,

$$\begin{aligned} E(Y_{iu}Y_{iv}) &= E(E(\cdots(E(Y_{iu}Y_{iv}|Y_{iu},\cdots,Y_{iv-1})))) \\ &= \mu_{iu}[p^{*v-u} + (1-p^*)\sum_{j=u+1}^{v} p^{*v-j}\mu_{ij}^*] \end{aligned} \tag{3.7}$$

where $u = 1, \cdots, T$, $v = 2, \cdots, T$, and $i = 1, \cdots, N$.

Then the off-diagonal elements of the covariance matrix Σ_i are calculated as follows:

$$\begin{aligned} cov(Y_{iu}, Y_{iv}) &= E(Y_{iu}Y_{iv}) - E(Y_{iu})E(Y_{iv}) \\ &= E(Y_{iu}Y_{iv}) - \mu_{iu}\mu_{iv} \\ &= p^{*v-u}\mu_{iu}(1-\mu_{iu}) \end{aligned} \tag{3.8}$$

where $u < v$, $u = 1, \cdots, T$, $v = 2, \cdots, T$, and $i = 1, \cdots, N$.

The Newton-Raphson method was applied to solve the estimating equation in the following iteration formula:

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + COV^{-1}FUN|_{\theta=\hat{\theta}^{(n)}_{NGQL}} \tag{3.9}$$

where COV^{-1} denotes the variance-covariance matrix and FUN denotes the estimating function . In practice COV^{-1} can be estimated by

$$COV^{-1} = (\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \theta}\Sigma_i^{-1}\frac{\partial \mu_i}{\partial \theta})^{-1}|_{\theta=\hat{\theta}^{(n)}_{NGQL}} \tag{3.10}$$

FUN can be expressed as follows,

$$FUN = \sum_{i=1}^{N} \frac{\partial \mu_i^{'}}{\partial \theta} \Sigma_i^{-1} (y_i - \mu_i)|_{\theta=\hat{\theta}_{NGQL}^{(n)}} \tag{3.11}$$

3.2 Corrected generalized quasi-likelihood method

3.2.1 Regression model with measurement error

In order to correct the bias of the estimates of regression parameters, the measurement errors of the covariates should be taken into account.

Let $y_i = (y_{i1}, \cdots, y_{iT})'$ denote the longitudinal observation of a binary response, where $\mu_i^e = (\mu_{i1}^e, \cdots, \mu_{iT}^e)'$ is the vector of the means of Y_i and Σ_i^e is the $T \times T$ covariance matrix of Y_i. We use corrected generalized quasi-likelihood method (GQL) to estimate the regression parameters, by solving the following estimating equations:

$$\sum_{i=1}^{N} \frac{\partial \mu_i^{'e}}{\partial \theta} (\Sigma_i^e)^{-1} (y_i - \mu_i^e) = 0 \tag{3.12}$$

where $\theta = (\beta, p^*)'$ and $\frac{\partial \mu_i^{'e}}{\partial \theta} = (\frac{\partial \mu_{i1}^e}{\partial \theta}, \cdots, \frac{\partial \mu_{iT}^e}{\partial \theta})$ is the $(p+1) \times T$ first derivative matrix of the means μ_i^e, which is given by:

$$\begin{aligned} \frac{\partial \mu_{i1}^e}{\partial \beta_p} &= \frac{\partial \mu_{i1}^{*e}}{\partial \beta_p} \\ \frac{\partial \mu_{i1}^e}{\partial p^*} &= 0 \end{aligned} \tag{3.13}$$

for $t = 1$, where $p = 1, 2$

and

$$\begin{aligned}\frac{\partial \mu_{it}^{e}}{\partial \beta_p} &= p^{*}\frac{\partial \mu_{it-1}^{e}}{\partial \beta_p} + (1-p^{*})\frac{\partial \mu_{it}^{*e}}{\partial \beta_p} \\ \frac{\partial \mu_{it}^{e}}{\partial p^{*}} &= \mu_{it-1}^{e} - \mu_{it}^{*e} \end{aligned} \tag{3.14}$$

for $t = 2, \cdots, T$,where $i = 1, \cdots, N$ and $p = 1, 2$.

The covariance matrix Σ_i^e of Y_i has involved the covariates, which have measurement errors. It can be expressed as follows:

$$\Sigma_i^e = \begin{bmatrix} var(Y_{i1}) & cov(Y_{i1}, Y_{i2}) & \cdots & cov(Y_{i1}, Y_{iT}) \\ & var(Y_{i2}) & \cdots & cov(Y_{i2}, Y_{iT}) \\ & & \ddots & \cdots \\ & & & var(Y_{iT}) \end{bmatrix} \tag{3.15}$$

The formulas for the components of this covariance matrix Σ_i^e are given as follows:

The diagonal elements of covariance matrix Σ_i^e are

$$\begin{aligned} var(Y_{it}) &= E(Y_{it}^2) - (E(Y_{it}))^2 \\ &= E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)^2) - (E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)))^2 \\ &= E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)) - (E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)))^2 \\ &= \mu_{it}^e - (\mu_{it}^e)^2 \\ &= \mu_{it}^e(1-\mu_{it}^e) \end{aligned} \tag{3.16}$$

for $t = 1, \cdots, T$, $i = 1, \cdots, N$.

The off-diagonal elements of the covariance matrix Σ_i^e are expressed as follows:

$$
\begin{aligned}
& cov(Y_{iu}, Y_{iv}) \\
&= E(Y_{iu}Y_{iv}) - E(Y_{iu})E(Y_{iv}) \\
&= E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) - E_{x_{2i}|z_i}(E(Y_{iu}|x_{2i}|z_i))E_{x_{2i}|z_i}(E(Y_{iv}|x_{2i}|z_i)) \\
&= E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) - \mu^e_{iu}\mu^e_{iv} \qquad (3.17)
\end{aligned}
$$

for $u = 1, \cdots, T-1$, $v = u+1, \cdots, T$, $i = 1, \cdots, N$.

The second order moments are calculated as, for $u < v$,

$$
\begin{aligned}
& E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) \\
&= E_{x_{2i}|z_i}(E(E(\cdots(E(Y_{iu}Y_{iv}|Y_{iu}, \cdots, Y_{iv-1})))) \\
&= E_{x_{2i}|z_i}(\mu^e_{iu}[p^{*v-u} + (1-p^*)\sum_{j=u+1}^{v} p^{*v-j}\mu^{*e}_{ij}]) \\
&= \int(\mu^e_{iu}[p^{*v-u} + (1-p^*)\sum_{j=u+1}^{v} p^{*v-j}\mu^{*e}_{ij}])f(x_{2i}|z_i)dx_{2i} \qquad (3.18)
\end{aligned}
$$

where $u = 1, \cdots, T$, $v = 2, \cdots, T$, and $i = 1, \cdots, N$.

The integrals in the above formula have no closed form for the logit link. We discuss this further in details in the following subsection.

3.2.2 Approximation of second order moments

The approximation of the second order moment $E(Y_{iu}Y_{iv})$ is considered by Monahan and Steanski (1992) when the logit link function is involved in the integral with normal random effects. We use their approach to approximate a similar integral where the extra randomness is caused by normal measurement errors. The formula is given below.

$$E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i))$$
$$= \int_{-\infty}^{+\infty} (\mu_{iu}^{e}[p^{*v-u} + (1-p^{*}) \sum_{j=u+1}^{v} p^{*v-j}\mu_{ij}^{*e}])f(x_{2i}|z_i)dx_{2i} \quad (3.19)$$

where $u = 1, \cdots, T$, $v = 2, \cdots, T$, $u \neq v$, and $i = 1, \cdots, N$.

The integral involves the following terms:

$$\begin{aligned} m_{uv} &= \int_{-\infty}^{+\infty} g(x_{1iu}\beta_1 + x_{2iu}\beta_2)g(x_{1iv}\beta_1 + x_{2iv}\beta_2)f(x_{2i}|z_i)dx_{2i} \\ &= \int_{-\infty}^{+\infty} g(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)g(x_{1iu}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)f(\beta_2\epsilon_i)d\epsilon_i \end{aligned} \quad (3.20)$$

where $g(v) = (1 + exp(-v))^{-1}$ and $u = 1, \cdots, T$, $v = 1, \cdots, T$ and $u \neq v$, $i = 1, \cdots, N$. This integral includes one logit link function and one normal distribution function. There is no closed form solution for it. We use an approximation approach to handle this problem. This logistic-normal integral plays an important role in the analysis of binary data. In particular, this integral is essential to logistic regression with a normally distributed covariate measurement error.

Let $g(.)$ be the logit link function, Φ be the distribution function of standard normal random variable. Suppose G is a function of g, as suggested by Eugene (2004, p335). Then we have

$$\begin{aligned} g(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i) &= \frac{exp(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)}{1 + exp(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)} \\ &\approx G(\Phi(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)) \\ &= G(\Phi_1) \end{aligned} \tag{3.21}$$

for $u = 1, \cdots, T$ and $i = 1, \cdots, N$, and

$$\begin{aligned} g(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) &= \frac{exp(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)}{1 + exp(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)} \\ &\approx G(\Phi(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)) \\ &= G(\Phi_2) \end{aligned} \tag{3.22}$$

for $v = 1, \cdots, T$ and $i = 1, \cdots, N$.

We can rewrite the function as below:

$$\begin{aligned} & g(x_{1iu}\beta_1 + z_{iu}\beta_2 - \beta_2\epsilon_i) \\ &= \frac{exp(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)}{1 + exp(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)} \\ &\approx G(\Phi(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)) \\ &= G(\Phi_1) \end{aligned} \tag{3.23}$$

for $u = 1, \cdots, T$ and $i = 1, \cdots, N$, and

$$\begin{aligned} & g(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) \\ &= \frac{exp(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)}{1 + exp(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)} \\ &\approx G(\Phi(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)) \\ &= G(\Phi_2) \end{aligned} \tag{3.24}$$

for $v = 1, \cdots, T$ and $i = 1, \cdots, N$.

Therefore

$$\begin{aligned} m_{uv} &= \int_{-\infty}^{+\infty} g(x_{1iu}\beta_1 + x_{2iu}\beta_2) g(x_{1iv}\beta_1 + x_{2iv}\beta_2) f(x_{2i}|z_i) dx_{2i} \\ &= \int_{-\infty}^{+\infty} g(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i) g(x_{1iu}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) f(\beta_2\epsilon_i) d\epsilon_i \\ &\approx \int_{-\infty}^{+\infty} G(\Phi_1) G(\Phi_2) f(\beta_2\epsilon_i) d\epsilon_i \end{aligned} \tag{3.25}$$

Define $\Phi_1 = \Phi(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)$, $\Phi_{10} = \Phi(x_{1iu}\beta_1 + z_{iu}\beta_2)$, $\Phi_2 = \Phi(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i)$, $\Phi_{20} = \Phi(x_{1iv}\beta_1 + z_{iv}\beta_2)$.

By Taylor series expansion, we have following term

$$\begin{aligned} m_{uv} &= \int_{-\infty}^{+\infty} G(\Phi_1) G(\Phi_2) f(\beta_2\epsilon_i) d\epsilon_i \\ &\approx \int_{-\infty}^{+\infty} [G(\Phi_{10}) + G(\Phi_{10})'(\Phi_1 - \Phi_{10})] \\ &\quad [G(\Phi_{20}) + G(\Phi_{20})'(\Phi_2 - \Phi_{20})] f(\beta_2\epsilon_i) d\epsilon_i \end{aligned} \tag{3.26}$$

With some algebra, the terms are obtained as follow:

$$
\begin{aligned}
&\int_{-\infty}^{+\infty}[G(\Phi_{10})+G(\Phi_{10})'(\Phi_1-\Phi_{10})][G(\Phi_{20})+G(\Phi_{20})'(\Phi_2-\Phi_{20})]f(\beta_2\epsilon_i)d\epsilon_i \\
\approx\ & G(\Phi_{10})G(\Phi_{20})\int_{-\infty}^{+\infty}f(\beta_2\epsilon_i)d\epsilon_i \\
&+G(\Phi_{10})(G(\Phi_{20}))'\int_{-\infty}^{+\infty}\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i \\
&-G(\Phi_{10})(G(\Phi_{20}))'\Phi_{20}\int_{-\infty}^{+\infty}f(\beta_2\epsilon_i)d\epsilon_i \\
&+(G(\Phi_{10}))'G(\Phi_{20})\int_{-\infty}^{+\infty}\Phi_1 f(\beta_2\epsilon_i)d\epsilon_i \\
&+(G(\Phi_{10}))'(G(\Phi_{20}))'\int_{-\infty}^{+\infty}\Phi_1\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i \\
&-(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_{20}\int_{-\infty}^{+\infty}\Phi_1 f(\beta_2\epsilon_i)d\epsilon_i \\
&-(G(\Phi_{10}))'G(\Phi_{20})\Phi_{10}\int_{-\infty}^{+\infty}f(\beta_2\epsilon_i)d\epsilon_i \\
&-(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_{10}\int_{-\infty}^{+\infty}\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i \\
&+(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_{10}\Phi_{20}\int_{-\infty}^{+\infty}f(\beta_2\epsilon_i)d\epsilon_i
\end{aligned}
\tag{3.27}
$$

By the property of probability density functions and equation (2.20) we know that

$$
\begin{aligned}
&\int_{-\infty}^{+\infty}f(\beta_2\epsilon_i)d\epsilon_i=1 \\
&\int_{-\infty}^{+\infty}\Phi_1 f(\beta_2\epsilon_i)d\epsilon_i=\Phi_1(\frac{x_{1iu}\beta_1+z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
&\int_{-\infty}^{+\infty}\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i=\Phi_2(\frac{x_{1iv}\beta_1+z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})
\end{aligned}
\tag{3.28}
$$

Then we add and subtract a same amount

$(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_2(\frac{x_{1iu}\beta_1+z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})\Phi_2(\frac{x_{1iv}\beta_1+z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})$

Thus

$$
\begin{aligned}
m_{uv} &= \int_{-\infty}^{+\infty} G(\Phi_1)G(\Phi_2)f(\beta_2\epsilon_i)d\epsilon_i \\
&\approx G(\Phi_{10})G(\Phi_{20}) + G(\Phi_{10})(G(\Phi_{20}))'\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
&\quad -G(\Phi_{10})(G(\Phi_{20}))'\Phi_{20} + (G(\Phi_{10}))'G(\Phi_{20})\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
&\quad +(G(\Phi_{10}))'(G(\Phi_{20}))'\int_{-\infty}^{+\infty}\Phi_1\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i \\
&\quad -(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_{20}\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
&\quad -(G(\Phi_{10}))'G(\Phi_{20})\Phi_{10} \\
&\quad -(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_{10}\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
&\quad +(G(\Phi_{10}))'(G(\Phi_{20}))'\Phi_{10}\Phi_{20} \qquad (3.29)
\end{aligned}
$$

Simplifying the above function, we obtain

$$
\begin{aligned}
m_{uv} &= \int_{-\infty}^{+\infty} G(\Phi_1)G(\Phi_2)f(\beta_2\epsilon_i)d\epsilon_i \\
&\approx [G(\Phi_{10}) + (G(\Phi_{10}))'\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) - (G(\Phi_{10}))'\Phi_{10}] \\
&\quad [G(\Phi_{20}) + (G(\Phi_{20}))'\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) - (G(\Phi_{20}))'\Phi_{20}] \\
&\quad +(G(\Phi_{10}))'(G(\Phi_{20}))' \\
&\quad [\int_{-\infty}^{+\infty}\Phi_1\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i \\
&\quad -\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})] \qquad (3.30)
\end{aligned}
$$

Then by Taylor's theorem, we have

$$G(\Phi_{10}) + G(\Phi_{10})'(\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) - \Phi_{10}) = G(\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})) \tag{3.31}$$

$$G(\Phi_{20}) + G(\Phi_{20})'(\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) - \Phi_{20}) = G(\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})) \tag{3.32}$$

From the two equations above, we obtain

$$\begin{aligned} m_{uv} &= \int_{-\infty}^{+\infty} G(\Phi_1)G(\Phi_2)f(\beta_2\epsilon_i)d\epsilon_i \\ &\approx G(\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}))G(\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})) \\ &\quad +(G(\Phi_{10}))'(G(\Phi_{20}))' \\ &\quad [\int_{-\infty}^{+\infty} \Phi_1\Phi_2 f(\beta_2\epsilon_i)d\epsilon_i - \Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})] \end{aligned} \tag{3.33}$$

From the equation (2.17) and (2.19) in Chapter 2 for Φ_1 and Φ_2, we have

$$\begin{aligned} &\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\ &= \int_{-\infty}^{+\infty} \Phi_1(x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i)f(\beta_2\epsilon_i)d\epsilon_i \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i} \frac{1}{\sqrt{2\pi}}exp(-\frac{w_1^2}{2})dw_1 \frac{1}{\beta_2\sigma_i\sqrt{2\pi}}exp(-\frac{\epsilon_i^2}{2\beta_2^2\sigma_i^2})d\epsilon_1 \\ &= P(-\infty < \beta_2\epsilon_i < +\infty, w_1 < x_{1iu}\beta_1 + z_{iu}\beta_2 - \beta_2\epsilon_i) \\ &= P(w_1 - \beta_2\epsilon_i < x_{1iu}\beta_1 + z_{iu}\beta_2), \end{aligned} \tag{3.34}$$

$$
\begin{aligned}
&\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}) \\
&= \int_{-\infty}^{+\infty} \Phi_2(x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) f(\beta_2\epsilon_i) d\epsilon_i \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{x_{1iv}\beta_1 + z_{iv}\beta_2 - \beta_2\epsilon_i} \frac{1}{\sqrt{2\pi}} exp(-\frac{w_2^2}{2}) dw_2 \frac{1}{\beta_2\sigma_i\sqrt{2\pi}} exp(-\frac{\epsilon_i^2}{2\beta_2^2\sigma_i^2}) d\epsilon_1 \\
&= P(-\infty < \beta_2\epsilon_i < +\infty, w_2 < x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) \\
&= P(w_2 - \beta_2\epsilon_i < x_{1iv}\beta_1 + z_{iv}\beta_2). \qquad (3.35)
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} \Phi_1\Phi_2 f(\beta_2\epsilon_i) d\epsilon_i \\
&= \int_{-\infty}^{+\infty} \Phi_1\Phi_2 \frac{1}{\beta_2\sigma_i\sqrt{2\pi}} exp(-\frac{\epsilon_i^2}{2\beta_2^2\sigma_i^2}) d\epsilon_1 \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i} \frac{1}{\sqrt{2\pi}} exp(-\frac{w_1^2}{2}) dw_1 \\
&\int_{-\infty}^{x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i} \frac{1}{\sqrt{2\pi}} exp(-\frac{w_2^2}{2}) dw_2 \frac{1}{\beta_2\sigma_i\sqrt{2\pi}} exp(-\frac{\epsilon_i^2}{2\beta_2^2\sigma_i^2}) d\epsilon_1 \\
&= P(-\infty < \beta_2\epsilon_i < +\infty, w_1 < x_{1iu}\beta_1 + z_{iu}\beta_2 + \beta_2\epsilon_i, w_2 < x_{1iv}\beta_1 + z_{iv}\beta_2 + \beta_2\epsilon_i) \\
&= P(w_1 - \beta_2\epsilon_i < x_{1iu}\beta_1 + z_{iu}\beta_2, w_2 - \beta_2\epsilon_i < x_{1iv}\beta_1 + z_{iv}\beta_2). \qquad (3.36)
\end{aligned}
$$

Now $G(\Phi_{10})'$ and $G(\Phi_{20})'$ need to be computed, with some algebra the results are given as below

$$
\begin{aligned}
G(\Phi_{10})' = \quad & \frac{exp(x_{1iu}\beta_1 + z_{iu}\beta_2)}{(1 + exp(x_{1iu}\beta_1 + z_{iu}\beta_2))^2} \\
& (\frac{1}{\sqrt{2\pi}} exp(-\frac{(x_{1iu}\beta_1 + z_{iu}\beta_2)^2}{2}))^{-1}, \qquad (3.37)
\end{aligned}
$$

$$G(\Phi_{20})' = \frac{exp(x_{1iv}\beta_1 + z_{iv}\beta_2)}{(1+exp(x_{1iv}\beta_1 + z_{iv}\beta_2))^2} (\frac{1}{\sqrt{2\pi}} exp(-\frac{(x_{1iv}\beta_1 + z_{iv}\beta_2)^2}{2}))^{-1}. \quad (3.38)$$

By the approximation exploited by Monahan and Stefanski (1992), combining the results above, we have

$$\begin{aligned} m_{uv} &= \int_{-\infty}^{+\infty} G(\Phi_1)G(\Phi_2)f(\beta_2\epsilon_i)d\epsilon_i \\ &\approx G(\Phi_1(\frac{x_{1iu}\beta_1 + z_{iu}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}}))G(\Phi_2(\frac{x_{1iv}\beta_1 + z_{iv}\beta_2}{\sqrt{1+\beta_2^2\sigma_i^2}})) \\ &+ \frac{exp(x_{1iu}\beta_1 + z_{iu}\beta_2)}{(1+exp(x_{1iu}\beta_1 + z_{iu}\beta_2))^2}(\frac{1}{\sqrt{2\pi}} exp(-\frac{(x_{1iu}\beta_1 + z_{iu}\beta_2)^2}{2}))^{-1} \\ &\frac{exp(x_{1iv}\beta_1 + z_{iv}\beta_2)}{(1+exp(x_{1iv}\beta_1 + z_{iv}\beta_2))^2}(\frac{1}{\sqrt{2\pi}} exp(-\frac{(x_{1iv}\beta_1 + z_{iv}\beta_2)^2}{2}))^{-1} \\ &[P(w_1 - \beta_2\epsilon_i < x_{1iu}\beta_1 + z_{iu}\beta_2, w_2 - \beta_2\epsilon_i < x_{1iv}\beta_1 + z_{iv}\beta_2) \\ &-P(w_1 - \beta_2\epsilon_i < x_{1iu}\beta_1 + z_{iu}\beta_2) \\ &P(w_2 - \beta_2\epsilon_i < x_{1iv}\beta_1 + z_{iv}\beta_2)] \end{aligned} \quad (3.39)$$

Based on the above results, we can calculate the variance matrix of CGQL. The details will be given in the next subsection.

3.2.3 Computation of the covariance matrix of CGQL

Now we use the formulas of the second order moment to calculate the covariance matrix of CGQL, and provide the formulas for the components of this covariance matrix as follows:

$$
\begin{aligned}
var(Y_{it}) &= E(Y_{it}^2) - (E(Y_{it}))^2 \\
&= E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)^2) - (E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)))^2 \\
&= E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)) - (E_{x_{2i}|z_i}(E(Y_{it}|x_{2i}|z_i)))^2 \\
&= \mu_{it}^e - (\mu_{it}^e)^2 \\
&= \mu_{it}^e(1 - \mu_{it}^e), \qquad (3.40)
\end{aligned}
$$

for $t = 1, \cdots, T$, $i = 1, \cdots, N$.

The off-diagonal elements of covariance matrix as follows:

$$
\begin{aligned}
& cov(Y_{iu}, Y_{iv}) \\
&= E(Y_{iu}Y_{iv}) - E(Y_{iu})E(Y_{iv}) \\
&= E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) - E_{x_{2i}|z_i}(E(Y_{iu}|x_{2i}|z_i))E_{x_{2i}|z_i}(E(Y_{iv}|x_{2i}|z_i)) \\
&= E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) - \mu_{iu}^e\mu_{iv}^e, \qquad (3.41)
\end{aligned}
$$

for $u = 1, \cdots, T-1$, $v = u+1, \cdots, T$, $i = 1, \cdots, N$.

The second moment have the formula, for $u < v$,

$$
\begin{aligned}
& E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) \\
&= E_{x_{2i}|z_i}(E(E(\cdots(E(Y_{iu}Y_{iv}|Y_{iu}, \cdots, Y_{iv-1}))))) \\
&= E_{x_{2i}|z_i}(\mu_{iu}^e[p^{*v-u} + (1-p^*)\sum_{j=u+1}^{v} p^{*v-j}\mu_{ij}^{*e}]) \\
&= \int(\mu_{iu}^e[p^{*v-u} + (1-p^*)\sum_{j=u+1}^{v} p^{*v-j}\mu_{ij}^{*e}])f(x_{2i}|z_i)dx_{2i}, \qquad (3.42)
\end{aligned}
$$

where $u = 1, \cdots, T-1$, $v = 2, \cdots, T$, and $i = 1, \cdots, N$.

For $u < v$, $u = 1, \cdots, T-1$, $v = 2, \cdots, T$, and $i = 1, \cdots, N$,

$$\begin{aligned} E(Y_{iu}Y_{iv}) &= E_{x_{2i}|z_i}(E(Y_{iu}Y_{iv}|x_{2i}|z_i)) \\ &= E_{x_{2i}|z_i}(E(E(\cdots(E(Y_{iu}Y_{iv}|Y_{iu}, \cdots, Y_{iv-1}|x_{2i}|z_i))))) \\ &= p^{*(v-u)}\mu_{iu}^{*e} + (1-p^*)\sum_{j=2}^{v} p^{*(v-j)}m_{ju}. \end{aligned} \tag{3.43}$$

Now the diagonal elements and off-diagonal elements of covariance matrix have been calculated from equations (3.40)-(3.43). This completes the construction of the covariance matrix.

We can then use the Newton-Raphson method to solve the estimating equations, as addressed in the iterative formula below:

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + (COV^e)^{-1}FUN^e|_{\theta=\hat{\theta}^{(n)}_{CGQL}}, \tag{3.44}$$

where $(COV^e)^{-1}$ denotes the variance-covariance matrix and FUN^e denotes the estimating function. In practice $(COV^e)^{-1}$ can be estimated by

$$(COV^e)^{-1} = (\sum_{i=1}^{N} \frac{\partial \mu_i'^e}{\partial \theta}(\Sigma_i^e)^{-1}\frac{\partial \mu_i^e}{\partial \theta})^{-1}|_{\theta=\hat{\theta}^{(n)}_{CGQL}}, \tag{3.45}$$

and FUN^e can be expressed as follows,

$$FUN^e = \sum_{i=1}^{N} \frac{\partial \mu_i'^e}{\partial \theta}(\Sigma_i^e)^{-1}(y_i - \mu_i^e)|_{\theta=\hat{\theta}^{(n)}_{CGQL}}. \tag{3.46}$$

3.2.4 Asymptotic distribution of the GQL estimator

For $\theta = (\beta, p^*)'$, let us define $\hat{\theta}_{CGQL}$ to be the CGQL estimators of θ, which is obtained by solving the CGQL estimating equation (3.12). Under some mild regularity conditions, for example, those from the Theorem 3.4 of Newey and McFadden (1993), it follows from estimating equation (3.12) that as $N \to \infty$, by Central Limit Theorem,

$$N^{\frac{1}{2}}(\hat{\theta}_{CGQL} - \theta) \sim N(0, N(\sum_{i=1}^{N} \frac{\partial \mu_i'^e}{\partial \theta}(\Sigma_i^e)^{-1}\frac{\partial \mu_i^e}{\partial \theta})^{-1}) \tag{3.47}$$

Consequently, one may estimate the asymptotic standard errors of the CGQL of θ by using the above equation.

In the next chapter, we compare the relative performance of the proposed CGQL estimators obtained from estimating equation (3.12) and the NGQL estimators obtained from estimating equation (3.1). This will be done through a simulation study to be reported in the next chapter. The asymptotic variance of NGQL estimators and CGQL estimators are computed by (3.49) and (3.50).

Chapter 4

Simulation Study

In this chapter, we investigate the small sample performance of the CGQL method through simulation studies. The comparison between the CGQL and the NGQL is made at a variety of model settings, which reflect many practical situations. The simulation results are summarized in the form of tables, which are followed by a brief discussion for conclusion of this chapter.

4.1 Designs

In repeated binary data analysis, the dynamic dependence parameter is also of primary interest along with the regression effects. The main objective of the simulation study is to examine the performance of the NGQL and the CGQL in estimating these parameters. For this purpose, we choose a set of values for the components of θ. A set of repeated binary observations will be generated following the linear dynamic model. The covariate will be selected as in the following simulation design. The parameter values will then be estimated by solving the NGQL estimation

equation (3.1) and the CGQL estimating equation (3.12). The data generation and estimation of parameters will be repeated 500 times. Finally, these 500 NGQL and CGQL estimates will be summarized into tables. We will consider the logit link only. With the probit link, the proposed method performs similarly.

Now we conduct 500 simulations each time with the sample size $I = 100, 500$ under the assumption that the variance of the covariate measurement errors is known. Each independent individual with $t = 4$ repeated observation is generated following the linear dynamic model. The true parameter values are: the regression coefficients $\beta = (1, 1)$, the measurement errors are $\sigma^2 = 0.04, 0.25, 0.64$ and $p^* = 0.2, 0.5, 0.8$ for each set respectively. Finally 500 simulated data sets are yielded under the longitudinal model. The setting of the parameter values is presented as follows:

$$\beta = (1,1) \begin{cases} p^* = 0.2 & \begin{cases} \sigma^2 = 0.04 \\ \sigma^2 = 0.25 \\ \sigma^2 = 0.64 \end{cases} \\ p^* = 0.5 & \begin{cases} \sigma^2 = 0.04 \\ \sigma^2 = 0.25 \\ \sigma^2 = 0.64 \end{cases} \\ p^* = 0.8 & \begin{cases} \sigma^2 = 0.04 \\ \sigma^2 = 0.25 \\ \sigma^2 = 0.64 \end{cases} \end{cases} \tag{4.1}$$

We use x_{pit} to denote the time-dependent covariate for the ith individual at time t, where $p = 1, 2$, and z_{it} the observation of x_{2it}. x_{1it} follows

normal distribution, x_{2it} follows uniform distribution and z_{it} follows normal distribution. The covariate changes over time and subject. It can be generated as follows

Design:

The first covariate ,which does not have measurement errors, is considered to be,

$$x_{1it} \sim \begin{cases} N(0.1*(t-1), 1.5) & for \quad i = 1, \cdots, I/2, \quad t = 1, \cdots, 4 \\ N(0.1*t, 1.5) & for \quad i = I/2+1, \cdots, I, \quad t = 1, \cdots, 4 \end{cases} \tag{4.2}$$

The surrogate of second covariate, say z_{it}, is considered to be,

$$z_{it} \sim U(-4, 4) \quad for \quad i = 1, \cdots, I, \quad t = 1, \cdots, 4 \tag{4.3}$$

The second covariate, which has measurement error, is generated as follow:

$$x_{2it}|z_{it} \sim N(z_{it}, \sigma^2) \quad for \quad i = 1, \cdots, I, \quad t = 1, \cdots, 4 \tag{4.4}$$

The asymptotic covariance matrices for the NGQL and the CGQL estimators are calculated by

$$\sum_{i=1}^{N} (\frac{\partial \mu_i^{'}}{\partial \theta} (\Sigma_i)^{-1} \frac{\partial \mu_i}{\partial \theta})^{-1} \tag{4.5}$$

and

$$\sum_{i=1}^{N} (\frac{\partial \mu_i^{'e}}{\partial \theta} (\Sigma_i^e)^{-1} \frac{\partial \mu_i^e}{\partial \theta})^{-1} \tag{4.6}$$

respectively. The diagonal elements of the these covariance matrices, which are the variance of these estimators, are reported in the simulation results.

Moreover, we use the following formula to approximate the integral

$$\begin{aligned} \mu_{it}^{*e} &= E(Y_{it}|x_{1it}, z_{it}) \\ &= \int_{-\infty}^{+\infty} g(x_{1it}\beta_1 + z_{it}\beta_2 - \beta_2\epsilon_i) f(\beta_2 x_{2it}|z_{it}) dx_{2it} \\ &\approx g(\frac{x_{1it}\beta_1 + z_{it}\beta_2}{\sqrt{1 + \frac{\beta_2^2\sigma_i^2}{k^2}}}) \end{aligned} \tag{4.7}$$

for all $t = 1, \cdots, T$ and all $i = 1, \cdots, N$, where $g(v) = (1 + exp(-v))^{-1}$ and $k^2 = 1.70$. In most cases, the denominator in the above formula is very close to 1, and the regression estimation is a good approximation (Raymond J. Carroll and David Ruppert, 2006, pp91).

4.2 Results

In this section, we report the simulation results and discuss the performance of the CGQL and the NGQL approaches in estimating the parameters β. Simulation studies were conducted for regression coefficient β. For each of two estimation approaches we compute the simulated mean of the estimated β (SM), simulated standard errors (SSE), estimated standard error (ESE) and the coverage probability of 90% confidence interval (CPr). The simulation results are shown in the following tables.

Table 4.1: Sample size is 100, Dependence parameter is $p^* = 0.2$, Measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and Regression coefficient is $\beta_1 =$

$1, \beta_2 = 1$.

Table 4.2: Sample size is 100, Dependence parameter is $p^* = 0.5$, Measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and Regression coefficient is $\beta_1 = 1, \beta_2 = 1$.

Table 4.3: Sample size is 100, Dependence parameter is $p^* = 0.8$, Measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and Regression coefficient is $\beta_1 = 1, \beta_2 = 1$.

Table 4.4: Sample size is 500, Dependence parameter is $p^* = 0.2$, Measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and Regression coefficient is $\beta_1 = 1, \beta_2 = 1$.

Table 4.5: Sample size is 500, Dependence parameter is $p^* = 0.5$, Measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and Regression coefficient is $\beta_1 = 1, \beta_2 = 1$.

Table 4.6: Sample size is 500, Dependence parameter is $p^* = 0.8$, Measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and Regression coefficient is $\beta_1 = 1, \beta_2 = 1$.

Table 4.1: Comparison of simulated mean values (SM), simulated standard errors (SSE), estimated standard errors (ESE) and 90% coverage probability (CPr) for the estimation of model parameters under NGQL and CGQL, with dependence parameter is $p^* = 0.2$, measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and $\beta_1 = 1, \beta_2 = 1$, $n = 100$ and 500 simulations

Dependence parameter(p^*)	*Measurement error*(σ^2)	*Method*	*Quantity*	*Estimates*		
				$\hat{\beta}_1$	$\hat{\beta}_2$	p^*
0.2	0.04	NGQL	SM	0.9823	0.9425	0.1664
			SSE	0.1334	0.0689	0.1045
			ESE	0.1341	0.0650	0.1141
			CPr	0.8620	0.5230	0.620
		CGQL	SM	1.0020	1.0015	0.1799
			SSE	0.1958	0.0804	0.1128
			ESE	0.1921	0.0881	0.1085
			CPr	0.9080	0.9032	0.650
	0.25	NGQL	SM	0.9593	0.9051	0.1579
			SSE	0.2361	0.1844	0.1430
			ESE	0.2549	0.1862	0.1470
			CPr	0.8740	0.6260	0.7310
		CGQL	SM	1.0464	1.0296	0.1682
			SSE	0.2930	0.2864	0.1668
			ESE	0.2847	0.2872	0.1824
			CPr	0.8998	0.9026	0.8560
	0.64	NGQL	SM	0.8664	0.8537	0.1145
			SSE	0.3422	0.2491	0.3084
			ESE	0.3352	0.2485	0.3025
			CPr	0.8740	0.4080	0.5920
		CGQL	SM	1.0652	1.0491	0.1364
			SSE	0.3649	0.3202	0.3593
			ESE	0.3615	0.3260	0.3545
			CPr	0.9020	0.9008	0.6840

Table 4.2: Comparison of simulated mean values (SM), simulated standard errors (SSE), estimated standard errors (ESE) and 90% coverage probability (CPr) for the estimation of model parameters under NGQL and CGQL, with dependence parameter is $p^* = 0.5$, measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and $\beta_1 = 1, \beta_2 = 1$, $n = 100$ and 500 simulations

Dependence parameter(p^*)	*Measurement error*(σ^2)	*Method*	*Quantity*	*Estimates*		
				$\hat{\beta}_1$	$\hat{\beta}_2$	p^*
0.5	0.04	NGQL	SM	0.9317	0.9145	0.4167
			SSE	0.0994	0.0587	0.0909
			ESE	0.1025	0.0583	0.1052
			CPr	0.8240	0.6160	0.3800
		CGQL	SM	1.0151	1.0019	0.4318
			SSE	0.1178	0.0805	0.1466
			ESE	0.1211	0.0835	0.1318
			CPr	0.9120	0.9058	0.6400
	0.25	NGQL	SM	0.8854	0.8894	0.4253
			SSE	0.1240	0.1750	0.1754
			ESE	0.1303	0.1785	0.1765
			CPr	0.7380	0.4920	0.5240
		CGQL	SM	0.9743	1.0233	0.4428
			SSE	0.1419	0.1951	0.2124
			ESE	0.1397	0.1834	0.2138
			CPr	0.8992	0.9018	0.6840
	0.64	NGQL	SM	0.8125	0.7948	0.4164
			SSE	0.2152	0.2565	0.2216
			ESE	0.2200	0.2654	0.2389
			CPr	0.7660	0.4980	0.6840
		CGQL	SM	1.0442	1.0524	0.4395
			SSE	0.3981	0.3174	0.2932
			ESE	0.3929	0.3014	0.2848
			CPr	0.9180	0.8998	0.7540

Table 4.3: Comparison of simulated mean values (SM), simulated standard errors (SSE), estimated standard errors (ESE) and 90% coverage probability (CPr) for the estimation of model parameters under NGQL and CGQL, with dependence parameter is $p^* = 0.8$, measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and $\beta_1 = 1, \beta_2 = 1$ and n=100 under 500 simulations

Dependence parameter(p^*)	*Measurement error*(σ^2)	*Method*	*Quantity*	*Estimates*		
				$\hat{\beta}_1$	$\hat{\beta}_2$	p^*
0.8	0.04	NGQL	SM	0.9820	0.9410	0.7474
			SSE	0.1034	0.0562	0.0991
			ESE	0.0935	0.0549	0.1048
			CPr	0.7920	0.7600	0.7140
		CGQL	SM	1.0178	1.0083	0.7514
			SSE	0.1371	0.1073	0.1418
			ESE	0.1451	0.1072	0.1431
			CPr	0.9180	0.9160	0.7500
	0.25	NGQL	SM	0.8813	0.8746	0.7034
			SSE	0.1640	0.1259	0.1806
			ESE	0.1615	0.1127	0.1759
			CPr	0.8660	0.4960	0.670
		CGQL	SM	0.9519	1.0149	0.7428
			SSE	0.2285	0.1957	0.2127
			ESE	0.2303	0.1990	0.2146
			CPr	0.8940	0.9060	0.7800
	0.64	NGQL	SM	0.7744	0.7547	0.6970
			SSE	0.2360	0.2298	0.2999
			ESE	0.2328	0.2191	0.2987
			CPr	0.8940	0.5120	0.7640
		CGQL	SM	1.0344	1.0290	0.7569
			SSE	0.3594	0.3144	0.3438
			ESE	0.3500	0.3215	0.3415
			CPr	0.9102	0.9108	0.6380

Table 4.4: Comparison of simulated mean values (SM), simulated standard errors (SSE), estimated standard errors (ESE) and 90% coverage probability (CPr) for the estimation of model parameters under NGQL and CGQL, with dependence parameter is $p^* = 0.2$, measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and $\beta_1 = 1, \beta_2 = 1$, $n = 500$ and 500 simulations

Dependence parameter(p^*)	*Measurement error*(σ^2)	*Method*	*Quantity*	*Estimates*		
				$\hat{\beta}_1$	$\hat{\beta}_2$	p^*
0.2	0.04	NGQL	SM	0.9647	0.9721	0.1584
			SSE	0.1046	0.0247	0.0952
			ESE	0.1024	0.0263	0.1038
			CPr	0.7200	0.6200	0.7800
		CGQL	SM	1.0080	1.0017	0.1629
			SSE	0.1303	0.0329	0.1228
			ESE	0.1330	0.0310	0.1346
			CPr	0.9040	0.9140	0.8500
	0.25	NGQL	SM	0.8934	0.8343	0.1476
			SSE	0.1368	0.0353	0.1231
			ESE	0.1347	0.0375	0.1246
			CPr	0.8960	0.5630	0.6700
		CGQL	SM	1.0017	1.0078	0.1609
			SSE	0.1614	0.0404	0.1436
			ESE	0.1654	0.0406	0.1412
			CPr	0.9160	0.9200	0.7800
	0.64	NGQL	SM	0.7764	0.7466	0.1382
			SSE	0.1624	0.0416	0.1506
			ESE	0.1665	0.0327	0.1576
			CPr	0.6500	0.7500	0.7400
		CGQL	SM	1.0199	1.0173	0.1516
			SSE	0.2059	0.0742	0.2027
			ESE	0.1974	0.0804	0.2048
			CPr	0.8990	0.9100	0.7200

Table 4.5: Comparison of simulated mean values (SM), simulated standard errors (SSE), estimated standard errors (ESE) and 90% coverage Probability (CPr) for the estimation of model parameters under NGQL and CGQL, with dependence parameter is $p^* = 0.5$, measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and $\beta_1 = 1, \beta_2 = 1$, $n = 500$ and 500 simulations

Dependence parameter(p^*)	*Measurement error*(σ^2)	*Method*	*Quantity*	*Estimates* $\hat{\beta}_1$	$\hat{\beta}_2$	p^*
0.5	0.04	NGQL	SM	0.9563	0.9343	0.4381
			SSE	0.1032	0.0275	0.0910
			ESE	0.1092	0.0291	0.0891
			CPr	0.8860	0.6960	0.3020
		CGQL	SM	1.0028	1.0026	0.4516
			SSE	0.1219	0.0332	0.1208
			ESE	0.1277	0.0329	0.1274
			CPr	0.9200	0.9106	0.640
	0.25	NGQL	SM	0.8777	0.8464	0.4246
			SSE	0.1251	0.0383	0.1315
			ESE	0.1274	0.0374	0.1248
			CPr	0.6560	0.5140	0.5920
		CGQL	SM	1.0187	1.0112	0.4634
			SSE	0.1436	0.0570	0.1592
			ESE	0.1478	0.0573	0.1508
			CPr	0.9180	0.8996	0.6080
	0.64	NGQL	SM	0.7946	0.7843	0.4257
			SSE	0.1688	0.0476	0.1480
			ESE	0.1644	0.0448	0.1390
			CPr	0.5820	0.6220	0.300
		CGQL	SM	1.0226	1.0168	0.4536
			SSE	0.1930	0.0697	0.1884
			ESE	0.1989	0.0668	0.1872
			CPr	0.9160	0.9076	0.580

Table 4.6: Comparison of simulated mean values (SM), simulated standard errors (SSE), estimated standard errors (ESE) and 90% coverage Probability (CPr) for the estimation of model parameters under NGQL and CGQL, with dependence parameter is $p^* = 0.8$, measurement error $\sigma^2 = 0.04, 0.25, 0.84$ and $\beta_1 = 1, \beta_2 = 1$, $n = 500$ and 500 simulations

Dependence parameter(p^*)	*Measurement error*(σ^2)	*Method*	*Quantity*	*Estimates*		
				$\hat{\beta}_1$	$\hat{\beta}_2$	p^*
0.8	0.04	NGQL	SM	0.9331	0.9223	0.7481
			SSE	0.0789	0.0231	0.0923
			ESE	0.0781	0.0235	0.0946
			CPr	0.7720	0.6250	0.350
		CGQL	SM	1.0012	1.0016	0.7546
			SSE	0.0914	0.0326	0.0986
			ESE	0.0920	0.0352	0.0991
			CPr	0.8994	0.9044	0.5600
	0.25	NGQL	SM	0.8983	0.8783	0.7525
			SSE	0.0975	0.0317	0.1226
			ESE	0.0956	0.0324	0.1218
			CPr	0.6100	0.6150	0.780
		CGQL	SM	1.0188	1.0104	0.7615
			SSE	0.1141	0.0450	0.1436
			ESE	0.1165	0.0448	0.1445
			CPr	0.9040	0.8998	0.860
	0.64	NGQL	SM	0.7871	0.7946	0.7569
			SSE	0.1522	0.0448	0.1341
			ESE	0.1539	0.0445	0.1336
			CPr	0.6040	0.6740	0.670
		CGQL	SM	1.0388	1.0261	0.762
			SSE	0.1921	0.0646	0.1815
			ESE	0.1971	0.0660	0.1846
			CPr	0.9080	0.9032	0.580

4.3 Comparison

From the simulation results, we can observe that the means of the estimates of parameters $\hat{\beta}_2$ from the CGQL method are much closer to the true parameters value than that of the NGQL, which are known to be biased. In the Tables 4.1-4.6, the performance of the CGQL method and the NGQL method is compared in terms of the simulated mean value (SM), simulated standard error (SSE), estimated standard errors (ESE) and the coverage probability (CPr) of the confidence interval. These indicators are reported in Tables 4.1-4.3 for the case when $\beta_1 = 1, \beta_2 = 1$ and $n = 100$ with measurement error 0.04, 0.25, and 0.64 respectively and Tables 4.4-4.6 for the case when $\beta_1 = 1, \beta_2 = 1$ and $n = 500$ with measurement error 0.04, 0.25, and 0.64 respectively. In this subsection we discuss the simulation results by analyzing the bias, standard deviation and coverage probability.

From the Tables 4.1-4.6 we can see that the NGQL estimates are biased for most cases. The effect of measurement error on the estimation of regression parameters is ignorable, when the variance of the measurement error is very small. The simulated mean of $\hat{\beta}_2$ attenuates towards zero as the measurement error variance σ^2 increases. For instance, from Tables 4.1-4.6, for $\sigma^2 = 0.04$, $\hat{\beta}_2 = 0.9425, 0.9145, 0.9410$ while $\sigma^2 = 0.64$, $\hat{\beta}_2 = 0.8537, 0.7948, 0.7547$. We also see that the estimates of regression parameters $\hat{\beta}_1$ are affected. For example, from Tables 4.1-4.6, $\sigma^2 = 0.04$, $\hat{\beta}_1 = 0.9823, 0.9317, 0.9820$ while $\sigma^2 = 0.64$, $\hat{\beta}_2 = 0.8664, 0.8125, 0.7744$. Generally speaking, the estimation for the dynamic dependence parameter p^* is not very much affected by the measurement errors.

The CGQL method performs well in correcting the attenuation ef-

fect caused by measurement errors. The simulated means of the CGQL are closer to the true parameter values compared to those of the NGQL method. When the sample size increases, the efficiency gets better. The simulation results in Tables 4.1-4.6 indicate that all of the biases of the CGQL estimates are small. Hence the estimates from the CGQL method can be treated as unbiased. For example, in Table 4.4 when $p^* = 0.2$ and $\sigma^2 = 0.25$, the biases for β_2 are -0.1657 and 0.0078 for the NGQL and the CGQL method and in Table 4.6 when $p^* = 0.8$ and $\sigma^2 = 0.25$, the biases for β_2 are -0.1217 and 0.0104 for the NGQL and the CGQL method, respectively. So we can conclude that the improvement of the CGQL method is remarkable.

Once an estimate, either by the NGQL or the CGQL approach, is obtained for the true parameter value, in practice, one has to compute the standard error of the estimate for the construction of a confidence interval at a desired level of confidence and test of null hypothesis versus its complete alternative, as well. For this purpose, we have computed the asymptotic standard errors of the estimates for the parameters by using variance equation (3.15). The average of those standard errors for each of the three estimates were computed. From Table 4.1-4.6, we can see that the estimated standard errors are almost unbiased in the sense that the estimated standard errors are very close to the simulated standard error. As an example, when $p^* = 0.5$ and $\sigma^2 = 0.25$ in Table 4.5, $SSE = 0.0570$ and $ESE = 0.0573$ for β_2 of CGQL method. The variance estimation is meaningless for the NGQL approach except the case when the measurement error variance is very small.

From the simulation results, we also see that the standard error of es-

timated regression parameter increases with the variance of the measurement errors. For instance, when the measurement error variance changes from $\sigma^2 = 0.04$ to $\sigma^2 = 0.64$ in Table 4.7, The $ESE = 0.0352$ for β_2 change to $ESE = 0.0660$. This result is just as what we expected.

For the concern of possible testing of hypotheses, we have computed the coverage probability of the 90% confidence intervals for each of the three parameters. To be specific, by using a simulation based on the CGQL estimate of a parameter , say $\beta_2 = 1$, we have calculated the test statistics $z_{2,s} = (\hat{\beta}_{2,CGQL} - 1)/ESE(\hat{\beta}_{2,CGQL})$. The coverage probability is calculated as the proportion of situations that the 90% confidence interval include the true β_2. It is clear from Tables 4.1-4.6 that the coverage probabilities for the CGQL method are much closer to the nominal level of 90 percent than that of the NGQL method. Since the NGQL method provides biased estimates of the model parameters, the corresponding confidence interval is already meaningless. The proposed CGQL method successfully corrects the estimation bias caused by the measurement errors. It is more efficient as compared with the naive use of GQL.

Chapter 5

Conculsion

Due to the widely existing measurement errors in practical data, it is of great interest to examine the adverse effects of measurement errors on the underlying statistical inference. In this thesis, we have considered a linear dynamic conditional probability based model to analyze the longitudinal binary data, where the covariates are measurement error prone. This dynamic model allows the expected response to be related with the history of the covariates, which is more appropriate in many biomedical studies dealing with non-curable type diseases. For this model, the likelihood function is very hard to deal with. Thus the maximum likelihood method is difficult to apply. So we utilize a generalized quasi-likelihood method to conduct statistical inference of the model parameters. When the measurement errors of the covariates are not appropriately handled, the estimates of the regression parameters of the model are attenuated.

In order to rectify the attenuation caused by measurement errors in covariates, we have developed an approach that efficiently corrects the estimation bias. Our focus is on the unconditional generalized quasi-likelihood

inference that involves unconditional moments of up to the second order. We assumed a generalized dynamic linear model with logit link and probit link. We also use the method proposed by Monahan and Stefanski (1992) to approximate the expectation of a function involved in the calculation of the expectations and covariance.

Simulation studies were conducted in the aim of investigating the small sample properties of the proposed method. The simulation results show that the naive generalized quasi-likelihood method create remarkable estimation bias while the CGQL approach provides much better estimates.

The GQL method has very good potential in econometrics and biomedical science. In particular, this approach also has been broadly used for analyzing continuous binary data. We believe that this study should be useful for analyzing similar binary data in biological or medical sciences. This is to be taken up in a future study.

Bibliography

[1] Alan Agresti. *A Model for Repeated Measurements of a Multivariate Binary Response*, Journal of the American Statistical Association, Volume 92, No. 437, Theory and Method, 315-321, March, 1997.

[2] Breslow N.E. and Clayton D.G. *Approximate Inference in Generalized Linear Mixed Models*, Journal of the American Statistical Association, 88, 9-25, 1993.

[3] Buzas J.S. and Tosteson T.D. and Stefanski L.A. *Measurement Error*, Institute of Statistics Mimeo Series, paper No. 2544, 2003.

[4] Carroll, R.J., Maca, J., and Ruppert D. *Nonparametric Regression in the Presence of Measurement Error*, Biometrika, 86, 541-554, 1999.

[5] Carroll R.J., Spiegelman C.H., Lan K.K. and Bailey K.T. *On Errors-in-variable for Binary Regression Models*, Biometrika, 7, 19-26, 1984.

[6] Carroll R.J., Ruppert D., Stefanski L.A. and Crainiceanu M. *Measure-ment Error in Non Linear Models: A Modern Perspective*, Chapman and Hall, New York., 2006.

[7] Diggle P.J., Heagerty, P., Liang, K.-Y.and Zeger, S. L. *Longitudinal Data Analysis*, 2nd ed. Oxford, UK, Oxford University Press, 2002.

[8] Dunlop D.D. *Regression for Longitudinal Data: A Bridge from Least Squares Regression*, The American Statostocian, 48, 299-303, 1994.

[9] Eugene Demidenko. *Mixed Models: Theory and Applications*, John Wiley Sons, Inc., 2004

[10] Fuller W. *Measurement Error Models*, Wiley, New York, 1987.

[11] Garrett M. Fitzmaurice, Nan M. Laird and James H. Ware. *Applied Longitudinal Analysis*, John Wiley Sons, Inc., 2004.

[12] Gustafson P. *Measurement Error and Misclassification in Statistics and Impacts and Bayesian Adjustments*, Chapman and Hall/CRC, Boca Raton, 2003.

[13] Helmut Küchenhoff, Samuel M. Mwalili, and Emmanuel Lesaffre. *A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX*, Biometrics 62:85-96, March, 2006.

[14] Imbens G.W. *Generalized Method of Moments and Emperical Likelihood*, Journal of Business and Economic Statistics, 20, 493-506, 2002.

[15] Ji Yunqi. *Analysis of Longitudinal Categorical and Count Data Subject to Measurement Error*, Doctorial Thesis, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, 2011.

[16] Jiang J. and Zhang W. *Robust Estimation in Generalized Linear Mixed Models*, Biometrika, 88, 753-765, 2001.

[17] Kipnis V., Carroll R.J., Freedman L.S. and Li L. *A New Dietary Measurement Error Modem and Its Application to the Estimation of Relatie Risk: Application to Four Validation Studies*, American Journal of Epidemiology, 150, 642-51, 1999.

[18] Kipnis V., Midthune D., Freedman L.S., Bingham S., Day N.E., Riboli E. and Carroll R.J. *Bias in Dietary-report Instruments and Its Implications for Nutritional Epidemiology*, Public Health Nutrition, 5, 915-23, 2003.

[19] Kung-Yee Liang and Scott L. Zeger. *Longitudinal Data Analysis Using Generalized Linear Models*, Biometrika, Vol.73, No.1:13-22, April, 1986.

[20] Larid N.M. and Ware J.H. *Random-effects Models for Longitudinal Data*, Biometrics, 38, 963-974, 1982.

[21] Lee L.F and Sepanski J.H. *Estimation of Linear and Nonlinear Errors-in-variables Models Using Validation Data*, Journal of the American Statistical Association, 90, 130-140, 1995.

[22] Liu X. and Liang K. *Efficacy of Repeated Measures in Regression Models with Measurement Error*, Biometrics, 48, 645-654, March, 1992.

[23] Magder L.S. and Hughes J.P. *Logistic Regression when the Outcome is Measured with Uncertainty*, American Journal of Epidemiology, 146, 195-203, 1997.

[24] Mallick B.K., and Gelfand A.E. *Semiparametric Error-in-variables Models: A Bayesian Approach*, Journal of Statistical Planning and Inference, 52, 307-21, 1996.

[25] McGlothlin A., Stamey J.D. and Seaman J.W. *Binary Regression with Misclassified Response and Covariate Subject to Measurement Error: A Bayesian Approach*, Biometrical Journal, 50, 123-34, 2008.

[26] McCullagh P. and Nelder J.A. *Generalized Linear Models*, Chapman and Hall, London, 1989.

[27] Monahan John F. and Stefanski Leonard A. *Normal Scale Mixture Approximations to F(z) and Computation of the Logistic-normal Integral*, In Handbook of the Logistic Distribution, N. Balakrisshnan, eds, 529-540,1992.

[28] Neuhaus, J.M. *Bias and Effciency Loss due to Misclassified Responses in Binary Regression*, In handbook of Economics, 4, eds. D. McFadden and R. Engler, Amsterdam, North Holland, 1993.

[29] Newey W. K. and McFadden D. *Estimation in Large Sample*, In handbook of Economics, 4, eds. D. McFadden and R. Engler, Amsterdam, North Holland, 1993.

[30] Pierce D.A., Stram D.O., Vaeth M. and Schafer D.W. *The Error-in-variables Problem: Considerations Provided by Radiation Dose-response Analysis of the A-bomb Survivor Data*, Journal of the American Statistical Association, 87, 351-359, 1992.

[31] Qu A. and Song P. X.-K. *Testing Ignorable Missingness in Estimating Equation Approaches for Longitudinal Data*, Biometrika, 89, 823-836, 2002.

[32] Rosner B. Spiegelman D. and Willet W.C. *Correction of Logistics Regression Relative Risk Estimates and Confidence Intervals for Measurement Error: the Case of Multiple Covariates Measured with Error*, American Journal of Epidemiology, 132, 743-745, 1990.

[33] Raymond J. Carroll, David Ruppert, Leonard A.Stefanski and Ciprian M. Crainiceanu. *Measurement Error in Linear Models, A Modern Perspective*, Second Edition. Chapman and Hall/CRC, 1995.

[34] Rosychuk, R.J. *Accounting for Misclassification in Binary Longitudinal Data*, Doctorial Thesis, Waterloo University, Waterloo, Ontario, Canada, 1999.

[35] Rosychuk R.J. and Thompson M.E. *A Semi-Markov Model for Binary Longitudinal Responses Subject to Misclassification*, The Canadian Journal of Statistics, 29, 395-404, 2001.

[36] Rosychuk R.J. and Islam, S. *Parameter Estimation in a Model For Misclassified Markov Data - a Bayesian Approach*, Computational Statistics and Data Analysis, 53, 3805-16, 2009.

[37] Roy S. and Banerjee T. *Analysis of Misclassified Correlated Binary Data Using a Multivariate Probit Model when Covariates are subject to measurement error*, Biometrical Journal, 51, 420-32, 2009.

[38] Surupa Roy, T. Banerjee and Tapabrata Maiti. *Measurement Error Model for Misclassified Binary Responses*, Statist. Med. 24:269-283, 2005. Available at `www.interscience.wiley.com`

[39] Sutradhar, B.C. *Generalized Quasi-likelihood(GQL) Inference*, Version 8, StaProb: The Encyclopedia Sponsored by Statistics and Probability. Available at `http://staprob.com/encyclopedia/GeneralizedQuasilikelihoodGQLInferences.html`

[40] Sutradhar B.C. and Patrick J. Farrell. *On Optimal Lag 1 Dependence Estimation for Dynamic Binary Models with Application to Asthma Data*, Sankhya: The Indian Journal of Statistics, Volume 69, Part 3: 448-467, 2007.

[41] Sutradhar B.C., R. Prabhakar Rao and V.N. Pandit. *Generalized Method of Moments Versus Generalized Quasilikelihood Inferences in Binary Panel Data Models*, Sankhya: The Indian Journal of Statistics, Volume 70-B, Part 1, 34-62, 2008.

[42] Sutradhar B.C. *An Overview on Regression Models for Discrete Longitudinal Response*, Statistical Science, 18, 377-393, 2003.

[43] Sutradhar B.C. and Mukerjee, R. *On Likelihood Inference in Binary Mixed Model with Application to COPD Data*, Comput. Statist. Data Anal, 48, 345-361, 2005.

[44] Sutradhar B.C. *On Exact Quasilikelihood Inference in Generalized Linear Mixed Model*, Sankya: The India Joural of Statitics, 66, 261-289, 2004.

[45] Tao Yi. *Generalized Quasi-likelihood Method for Misclassified Longitudinal Binary Data*, A practicum for Master degree of Memorial University of Newfoundland, June 30, 2010.

[46] Tong H. *Nonlinear Times Serise: A Dynamical System Approach*, Oxford University Press, Oxford, 1990.

[47] Wang Y.-G. and Carey V. *Working Correlation Structure Misspeccification, Estimation and Covariate Design: Implication for Generalized Estimating Equations Performance*, Biometrika, 90, 29-41, 2003.

[48] Ware J.H. *Linear Models for the Analysis of Several Measurements in Logitudinal Studies*, The American Statistican, 39, 95-101, 1985.

[49] Zeger S.L. and Liang K.-Y. *Longitudinal Data Analysis for Discrete and Continuous Outcomes*, Biometrics, 42, 121-130, 1986.

[50] Zorn C.J.W. *Generalized Estimating Equation Models for Correlated Data: A Review with Applications*, American Journal of Political Science, 45, 470-490, 2001.