ANALYSIS OF LONGITUDINAL CATEGORICAL AND COUNT DATA SUBJECT TO MEASUREMENT ERROR

YUNQI JI

Analysis of Longitudinal Categorical and Count Data Subject to Measurement Error

by

© Yunqi Ji

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy in Statistics

Department of Mathematics & Statistics

Memorial University of Newfoundland

January 2011

St. John's Newfoundland

Abstract

In biomedical, social, behavioral, and environmental studies, the data are frequently collected from surveys, registration systems, clinical trials, and other observational or experimental studies, which are often contaminated with measurement errors. This may be due to the imperfect instruments and procedures, limited experience and knowledge of examiners and examinees. Ignoring measurement errors in responses results in biased estimates of model parameters. Explicit models are required to describe the misclassifications on categorical responses and count errors on aggregation responses. To obtain more reliable inference, one needs to take the measurement errors into consideration when developing statistical methods to analyze mis-measured data.

In this thesis, we define a generalized thinning operation, based on which we propose a transition model for categorical longitudinal data. This new transition model can flexibly accommodate a variety of linear and nonlinear transition models. We also discuss a thinning-operation-based transition model and an ordinary linear transition model for dynamic count data.

Most importantly, we present some new measurement error models for categorical data and count data, which link the true responses with the observed, possibly mismeasured responses by explicit expressions. A meaningful application of the explicit misclassification model is to describe the unbalanced misclassifications in categorical data, which provides an alternative way to jointly model the data suffering from both misclassification and some missing values due to "unsure" answers. Moreover, the count error models which accommodate both the overcounted and undercounted

data can be used to describe some interesting count data of disease cases with different situations of the dynamic population sizes of an area. We apply these explicit measurement error models and transition models to analyze the longitudinal discrete data subject to measurement errors.

Methods based on the generalized estimating equations (GEE), generalized quasi-likelihood (GQL), the second order GQL (GQL2), and maximum likelihood (ML) are developed to obtain unbiased hence consistent estimates of the unknown parameters in longitudinal models for categorical and count responses. The explicit measurement error models lead to simple development of the GEE, GQL and GQL2 approaches. Intensive simulations are conducted to examine the performance of these approaches. These methods tend to provide satisfactory estimates of model parameters, estimated standard errors and confidence intervals. Surprisingly the generalized quasi-likelihood approach performs almost as good as the likelihood approach when the latter is applicable in some first-order transition models. In the linear transition model for dynamic count data, even the GQL approach provide almost as good estimates as the ML approach. These findings provide us an efficient alternative to analyze longitudinal data when complicated dependence structure is taken into account the modeling. The proposed methods are illustrated by an example of children asthma data from Harvard Six Cities Study.

Acknowledgements

I would like to take this opportunity to express my heartiest appreciation to my advisor, Dr. Zhaozhi Fan, for his insightful guidance, consistent support, and impressive encouragement throughout the four years of my PhD program. I would also like to thank all my PhD advisor committee members, Dr. Hong Wang, Dr. J Concepción Loredo-Osti, for their support during my research.

I sincerely thank Dr. Brajendra Sutradhar, Dr. Alwell Oyet, Dr. Yingwei Peng for their academic support during my program. I also want to acknowledge Dr. Gary Sneddon, for his kind support and help during the Graduate Program in Teaching.

I would like to address my sincere acknowledgements to the examination committee members, Dr. Yingwei Peng, Dr. Alwell Oyet, and Dr. Yanqing Yi. They gave me a lot of useful comments on my thesis.

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

Special thanks go to my family, especially my wife. Their endless love and support is always essential to my success.

Finally, it is my great pleasure to thank my friends and fellow student who encouraged and helped me during my Ph.D. program.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Longitudinal Studies

1.1.1 Overview

Longitudinal discrete data such as categorical and count data often appear in a wide range of areas: public health, medicine, economics, sociology, and so on. In economics, continuous data are often collected and are called panel data. In addition, there are other names for different types of data, for example, repeated measurements, clustered data, and time series. Basically, the longitudinal data are collected from a sample of subjects, each of the subjects are repeatedly measured over time. For example, the Harvard School of Public Health conducted a large longitudinal study, which began in the 1970's, in six cities to evaluate the effects of air pollution on respiratory health among adults and children [Ferris, et al. (1985), Ware, et al. (1984)]. As part of the study, a number of children were recruited to investigate their annual respiratory health status over a period of four years. Their health status was determined based on the information provided by their parents through some standard questionnaires and the results of pulmonary function test by means of a portable survey spirometry.

In longitudinal studies, besides the observations of interested responses, the data about some related covariates are also collected over time. These covariates may be time-dependent or time-independent. For time-dependent covariates, the values change with time. For example, in Harvard six cities study (H6CS), some covariates such as the cigarette smoking habits, weights, heights, diets and ages are collected and they often vary with time. The time-independent covariates may represent some baseline factors which do not change over time, for example, gender and race. As a cross-sectional study, only one outcome for each participant is collected. The focus is to evaluate the effects of covariates, whereas in a longitudinal study, the dynamics among the response variable over time is also of 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.

The defining characteristic of longitudinal data is that the multiple observations within subjects are not independent of each other. Therefore, in data analysis, one should take into account the correlation between observations from the same subject. To combine different correlation structures into the analysis, special statistical methods are required. This helps to draw more reliable statistical inference, especially for discrete data.

There are three models which are extensions of generalized linear models (GLMs) for longitudinal data: *marginal models*, *random effects models* and *transition models*. In *marginal models*, when the correlation structure of the responses is not of direct interest to the researchers, one mainly focuses on the effects of covariates. Hence the marginal expectation, $\mu_{ij} = E(Y_{ij})$, is modeled as a function of some explanatory variables. Therefore, the regression of the responses on covariates can be modeled separately from the correlation within subjects. For example, a logistic marginal model for longitudinal binary data can be given by

$$\begin{aligned} logit(\mu_{ij}) &= x'_{ij}\beta, \\ Var(Y_{ij}) &= \mu_{ij}(1-\mu_{ij}), \\ Corr(Y_{ij}, Y_{iu}) &= \rho \end{aligned}$$

where Y_{ij} is the response of subject i at the jth time point, and x_{ij} is the corresponding covariates, for $i = 1, 2, \ldots I$, and $j = 1, 2, \ldots, J$.

When we are interested in making statistical inference about the individuals but not population average, a *random effects model* will be helpful. Random effects models allow for the natural heterogeneity cross subjects by assuming that coefficients of some covariates follow a probability distribution. For example, a random effect model in the GLM framework can be given by

(1) $\mu^c_{ij} = g(x'_{ij}\beta + z'_{ij}U_i)$, where $\mu^c_{ij} = E(Y_{ij} = 1|U_i)$,

(2) U_i, $i = 1, 2, \ldots, I$, are mutually independent with a common multivariate distribution F,

(3) $g(\cdot)$ is the inverse of a specific link function in GLM. For binary data, it may be the logit or probit function, that is, $\text{logit}(x) = log(\frac{x}{1-x})$ or $\text{probit}(x) = \Phi^{-1}(x)$, and $Y_{ij}|U_i \sim b(1, \mu^c_{ij})$. Whereas, for count data, $g(x) = \log(x)$, and $Y_{ij}|U_i \sim \text{Poisson}(\mu^c_{ij})$. Given i, $Y_{ij}|U_i$, $j = 1, 2, \ldots, J$, are independent of each other.

If one is interested in both the effects of covariates and the dynamic dependence among observations within subjects, a transition model serves as a good alternative. Under a *transition model*, the present response is explicitly influenced by the historical observations prior to time j, which is denoted by $\mathcal{H}_{ij} = \{y_{iu}, u = 1, 2, \ldots, j-1\}$. Therefore, both the explanatory variables and the past outcomes are treated as predictor variables. Let $\mu_{ij}^c = E(Y_{ij}|\mathcal{H}_{ij})$ and $v_{ij}^c = \mathrm{Var}(Y_{ij}|\mathcal{H}_{ij})$ be the conditional expectation and variance of Y_{ij} given past outcomes and the covariates. Then, a general transition model can be given by

$$\mu_{ij}^c = g(x_{ij}'\beta, f(\mathcal{H}_{ij}'; \gamma)) \tag{1.1}$$

$$v_{ij}^c = v(\mu_{ij}^c \phi), \tag{1.2}$$

where $g(\cdot)$ is the inverse of a specific link function, and the transition from the previous states is represented by a series of known functions $f = (f_1, \ldots, f_s)'$, to the current response. Due to the intuitive dynamics among outcomes within subjects, in this thesis, we will focus on developing statistical analysis of transition models for the correlated categorical and count data.

1.1.2 Transition models for dynamic categorical data

There are different transition models proposed for dynamic categorical data.

1. Tong (1990, p. 113) discussed a linear transition model for dynamic binary data. This model is given by

$$y_{ij} = b_{ij} y_{i,j-1} + (1 - b_{ij})\epsilon_{ij}, \tag{1.3}$$

where $b_{ij} \sim b(1, \gamma_{ij})$ is the random dynamic dependence variable, $\epsilon_{ij} \sim b(1, \xi_{ij})$, and b_{ij} is independent of ϵ_{ij}. This can be easily generalized to accommodate the dynamic categorical data of dimension r by assuming that b_{ij} still follows $b(1, \gamma_{ij})$, but ϵ_{ij} follows a r-dimensional multinomial distribution, that is multinomial$_r(1, \xi_{ij})$. Similarly b_{ij} is assumed to be independent of ϵ_{ij}. It then follows that $Y_{ij} \sim b(1, \mu_{ij})$, where

$$\begin{aligned} \mu_{ij} &= \gamma\mu_{i,j-1} + (1-\gamma)\xi_{ij} \\ &= \gamma^{j-1}\mu_{i1} + (1-\gamma)\sum_{u=2}^{j}\gamma\xi_{iu}. \end{aligned} \tag{1.4}$$

As discussed by Sutradhar and Farrell (2007), the mean μ_{ij} in equation (1.4) is a function of not only the current covariates x_{ij} but also all historical covariate $\{x_{iu}, u < j\}$. Actually, this model can further be generalized to the kth order transition model for categorical data as follows

$$y_{ij} = \sum_{u=1}^{k} b_{ij(u)}y_{i,j-u} + (1 - \sum_{u=1}^{k} b_{ij(u)})\epsilon_{ij}, \tag{1.5}$$

where $b_{ij} \sim$ multinomial$(1, \gamma)$, and γ is a vector of probability. Sutradhar and Farrell (2007) pointed out that the correlation coefficients between binary observations under model (1.3) do not cover the full ranges from -1 to 1, which limits the use of this linear model in practice.

2. A similar linear transition model was proposed by Qaqish (2003). They used a family of multivariate binary distributions through a linear dynamic conditional expectation to construct the linear dynamic model. This conditional linear family of order k can be given by

$$P(Y_{ij} = 1|\mathcal{H}_{ij}^{k}) = E(Y_{ij}|\mathcal{H}_{ij}^{k}) = \mu_{ij} + \sum_{u=1}^{k} b_{ij(u)}(y_{i,j-u} - \mu_{i,j-u}), \tag{1.6}$$

where $\mathcal{H}_{ij}^k = (y_{i,j-1}, y_{i,j-2}, \ldots, y_{i,j-k})$ denotes the history at the previous k time points. The vector $b_{ij} = (b_{ij(1)}, b_{ij(2)}, \ldots, b_{ij(k)})'$ can be computed based on the specified correlation structure or using

$$b_{ij} = [Cov(\mathcal{H}_{ij}^k)]^{-1} Cov(\mathcal{H}_{ij}^k, Y_{ij}).$$

As mentioned by Mallick (2009), this model can use different working correlation structures, such as Gauss type AR(1), MA(1) or exchangeable correlation, by specifying it to C_i which is from $Cov(\mathcal{H}_{ij}^k) = V_i^{1/2} C_i V_i^{1/2}$, where $V_i = diag(\sigma_{i1}, \ldots, \sigma_{ik})$ with $\sigma_{iu} = Var(Y_{iu})$. However, it was noticed by Mallick (2009) and Farrell and Sutradhar (2006), the ranges for the correlations in C_i are bound to be restricted since $0 < E(Y_{ij}|\mathcal{H}_{ij}^k) < 1$ in (1.6).

3. A nonlinear transition model was discussed by some authors [Korn and Whittemore (1979); Zeger, Liang and Self (1985)]. They comprise a first-order Markov chain which is given by

$$logit(\mu_{ij}^c) = logit(P(Y_{ij} = 1|\mathcal{H}_{ij})) = x_{ij}'\beta + \gamma y_{i,j-1} \qquad (1.7)$$

Some econometricians [Amemiya (1985); Manski (1987)] called it a non-linear binary dynamic model. It has been shown by Farrell and Sutradhar (2006) that this model produces a reasonable correlation structure that allows the ranges of the correlations to be from -1 to 1. Diggle et al. (2002, p. 191) extended model (1.7) to order k, that is,

$$logit(\mu_{ij}^c) = logit(P(Y_{ij} = 1|\mathcal{H}_{ij})) = x_{ij}'\beta_k + \sum_{u=1}^{k} \gamma_u y_{i,j-u}, \qquad (1.8)$$

where β_k is the regression coefficient with the Markov chain of order k.

1.1.3 Transition models for dynamic count data

For dynamic count data, there are also some transition models proposed.

1. Wong (1986) discussed a transition model

$$\mu_{ij}^c = exp(x_{ij}'\beta)\{1 + exp(-\gamma_0 - \gamma_1 y_{i,j-1})\}, \text{ where } \gamma_0, \gamma_1 > 0. \tag{1.9}$$

   As a consequence of the constraints on γ_0 and γ_1, this model only allows for a negative correlation between the prior and current responses. In addition, the conditional expectation μ_{ij}^c must vary within a limited range from $exp(x_{ij}'\beta)$ to twice this value under the assumption about γ_0 and γ_1, which makes this model impractical in some cases.

2. Besag (1974) and Diggle, et al. (2002 p204) discussed a nonlinear dynamic model for longitudinal count data which is given by

$$\mu_{ij}^c = exp(x_{ij}'\beta + \gamma y_{i,j-1}). \tag{1.10}$$

   This model seems to be a analogy with the logistic transition model (1.7). However it has limited application in practice because the conditional expectation μ_{ij}^c increases as an exponential function of the previous observation $y_{i,j-1}$ when $\gamma > 0$. In the case that the conditional expectation is independent on covariates, the assumption $exp(x_{ij}'\beta) = \eta$ leads to a stationary process only when $\gamma < 0$. Hence the model can only characterize negative association without exponentially growing pattern over time.

3. Zeger and Qaqish (1988) introduced another log-linear transition model

$$\mu_{ij}^c = exp[x_{ij}'\beta + \gamma\{y_{i,j-1}^* - x_{i,j-1}'\beta\}]. \tag{1.11}$$

where $y^*_{i,j-1} = max(y_{i,j-1}, d)$ and $0 < d < 1$. When $\gamma = 0$, it reduces to an ordinary log-linear model. When $\gamma < 0$, there is a negative correlation between y_{ij} and $y_{i,j-1}$. When $\gamma > 0$ there is a positive correlation. This model describes a multiplicative pattern among the y_{ij}'s.

4. Blundell, Griffith and Windmeijer (2002) proposed a linear feedback model (LFM) to analyze the relationship between R&D and patents for a panel of US firms. Let $\xi_{ij} = exp(x'_{ij}\beta)$, and $E(Y_{ij}|\eta_i) = exp(x'_{ij}\beta + \eta_i) = \xi_{ij}v_i$, where η_i is the subject-specific random effect, and $v_i = exp(\eta_i)$, then the LFM is given by

$$E(Y_{ij}|y_{i,j-1}, v_i) = \gamma y_{i,j-1} + \xi_{ij}v_i \quad (1.12)$$

Since the $\xi_{ij}v_i$ is non-negative, the conditional mean of y_{ij} is bounded by $\gamma y_{i,j-1}$ from below

5. McKenzie (1988) discussed a stationary AR(1) model for count time series, and Sutradhar (2003) used it to model longitudinal count data. The model in longitudinal context is given by

$$y_{ij} = \gamma * y_{i,j-1} + \epsilon_{ij} \quad (1.13)$$

where $*$ is the binomial thinning operation. Under this model, $y_{ij-1} \sim Poisson(\mu_{i\cdot} = exp(x'_{i\cdot}\beta))$, $\epsilon_{ij} \sim Poisson((1-\gamma)\mu_{i\cdot})$, and ϵ_{ij} is independent of $y_{i,j-1}$. The constraints on the expectation of ϵ_{ij} leads to a stationary process for T_{ij}'s. Sutradhar, Jowaheer and Sneddon (2008) consider a non-stationary AR(1) model with the same form as (1.13) but different assumptions. Sutradhar and his colleagues assumed that $y_{i1} \sim Poisson(\mu_{i1})$ with $\mu_{ij} = exp(x'_{ij}\beta)$, for $j = 2, 3, \ldots, J$, and $\epsilon_{ij} \sim Poisson(\mu_{ij} - \gamma\mu_{i,j-1})$. Similarly, $y_{i,j-1}$ is assumed to be independent of ϵ_{ij}. Under the non-stationary model, one may then show that $E(Y_{ij}) = Var(Y_{ij}) = \mu_{ij} = exp(x'_{ij}\beta)$ for $j = 1, 2, \ldots, J$. This non-stationary model can be used for the dynamic count data with time-varying covariates.

1.1.4 Generalized estimating equations and generalized quasi-likelihood approaches

To estimate the unknown parameters in longitudinal models, different approaches are proposed. Among these approaches, the maximum likelihood (ML) method is considered to be the most efficient estimation procedure. Suppose that longitudinal data y_{ij}, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, follows a first-order transition model, the likelihood function can be written as

$$L(\theta|y) = \prod_{i=1}^{I} f(y_{i1}) \prod_{j=2}^{J} f(y_{ij}|y_{i,j-1}) \tag{1.14}$$

The estimates can be obtained by maximizing the likelihood function or the log-likelihood function $\ell(\theta) = \log\{L(\theta|y)\}$. For example, Sutradhar and Farrell (2007) developed the ML estimates of the parameters in the logistic transition model (1.7) for dynamic binary data. However, the ML approach strongly depends on the assumption that the joint distribution of y_i is known and exact. Its application is limited because of two reasons. The first one is that, in many cases, the joint distribution of y_i is very complicated which leads to considerable difficulty in developing the ML approach. For example, under the stationary AR(1) model (1.13), McKenzie (1988) presented the complex likelihood function. The other reason is that the assumptions about the joint distribution may be violated in practice, or it can even be completely unknown.

As an alternative of the ML approach, the generalized estimating equations method is proposed by Liang and Zeger (1986) for continuous or discrete longitudinal data. Their method is based on the estimating equations

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} W_i^{-1}(y_i - \mu_i) = 0, \tag{1.15}$$

where $\mu_i = E(Y_i)$, and W_i is the working covariance matrix which can be decomposed into $V_i^{1/2} C_i V_i^{1/2}$. In the decomposition of W_i, $V = \text{diag}(\sigma_{i1}, \ldots, \sigma_{iJ})$, and $C \equiv C(\theta, \alpha)$ is the working correlation matrix which may depend on the parameters θ in the mean structure μ_{ij} and a correlation parameter α. In practice, an estimate of α can also be obtained, either based on a moment estimator or a second set of estimating equations. There are several popular correlation structures in practice. For example, the independence structure in which observations are assumed to be independent, the exchangeable structure in which the correlation between observations within subjects is constant, the auto-regressive structure in which the correlation is a function of the time between observations, and the unstructured correlation in which there is not an assumed pattern of correlations. It can be seen that the consistency of the GEE estimates only depends on the mean structure μ_{ij}. Therefore, regardless of the choice of the working covariance structure, one can always obtain a consistent estimate of θ as long as the mean structure is correctly specified. The less dependence on the model assumptions makes the GEE approach one of the most popular methods in dealing with correlated data. More discussions can be found in Zeger and Liang (1986) [also see Zeger and Qaqish (1988); Hardin and Hilbe (2003); Diggle et al. (2002)].

In the GEE approach, one takes into account the correlation by choosing a working covariance structure W_i in the estimating equations. However, if the chosen W_i

is far from the true covariance matrix Σ_i, it will result in loss of efficiency. To improve the efficiency of estimation, a quasi-likelihood-based method was proposed by Wedderburn (1974) by using the true covariance matrix. Sutradhar (2003) and Sutradhar and Farrell (2007) further discussed this generalized quasi-likelihood (GQL) method, of which the estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{-1} (y_i - \mu_i) = 0, \tag{1.16}$$

where Σ_i is the true covariance matrix of Y_i. The use of Σ_i leads to higher efficiency of the GQL estimate than the GEE estimates which are derived by using a working covariance structure. The GQL approach only depends on the first and second order moments of response Y_i which are available for many specific models.

To further improve the efficiency of estimation of model parameters, Sutradhar and Farrell (2007) introduced a second order GQL approach by employing the first and the second order responses in the estimating procedure. We refer it to the GQL2 approach in this thesis. Let $F_i = (Y_i', S_i')'$, where $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$, and $S_i = (Y_{i1}^2, \ldots, Y_{iJ}^2, Y_{i1}Y_{i2}, \ldots, Y_{iu}Y_{iv}, \ldots, Y_{im-1}Y_{iJ})'$, then $\delta_i = E(F_i) = (\mu_i', \nu_i')'$ with $\nu_i = E(S_i)$. Further, let

$$\Omega_i = \begin{pmatrix} Cov(Y_i) & Cov(Y_i, S_i) \\ Cov(S_i, Y_i) & Cov(S_i) \end{pmatrix}$$

be the $m(m+3)/2 \times m(m+3)/2$ covariance matrix of F_i. The GQL2 estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \delta_i'}{\partial \theta} \Omega_i^{-1} (f_i - \delta_i) = 0, \tag{1.17}$$

where f_i is the observation of F_i. It can be seen that the GQL2 approach utilizes the moments up to order 4. However, the consistency of the estimates of model parameters depends on the correctly specified first and second order moments. Due to the use of more information from the data, the GQL2 is demonstrated to gain higher efficiency than both GQL and GEE approaches. In some cases, the GQL2 approach performs almost as well as the ML approach [Sutradhar and Farrell (2007)]. Therefore, in this situation, the GQL2 can be the optimal GQL (OGQL) approach.

1.2 Measurement Errors

Although most studies are well designed to obtain accurate information, measurement errors in data still occur due to many known and unknown factors such as imperfect instruments and procedures, limited knowledge and experience of examiners and examinees, and so on. Measurement errors may occur in continuous data (e.g., the weights of children), categorical data (e.g., infection status), and count data (e.g., reported number of cancer cases). For categorical data, the measurement error takes the form of misclassification (classification error) which refers to incorrectly assigning group-membership. For count data, it takes the form of miscount (count error) which is due to undernumerated or overnumerated aggregations.

As an example, in a large population-based study to examine the effect of passive smoking on children asthma, researchers often rely on some proforma questionnaires because they are relatively simple and economical to conduct when compared to the clinical examination of each child. However, it is impossible to design a perfectly reliable questionnaire due to the complexities and wide range of severity of the disease, triggers, and lack of medical knowledge among the public [Jenkins et al. (1996)]. Moreover, the accuracy of a diagnosis based only on reported symptoms may be very poor because of the great overlap of measurements between healthy children and those with previous wheezing. Therefore, the true health state of a child is not directly observable. Instead, what we can obtain is the diagnostic status based on some imperfect information from the questionnaires. The data may therefore be contaminated by classification errors. Another example is that the new cancer cases by state in USA collected from the National Cancer Institute's Surveillance, Epidemiology and End Results Program (SEER) are prone to errors due to the limited coverage of the SEER registries [Wang et al. (200, 2001)], as well as the inconsistent and incomplete case reporting [Furlow, (2007)], the misdiagnoses by medical facilities [Colby et al. (2002); Giercksky (1997); Motto, Watanabe and Sawabu (1996)], and the inaccuracies of data coding by hospitals [Fisher et al. (1992); Cooper et al. (1999)]. Therefore, the annually reported counts of cancer cases may be contaminated with measurement errors.

There are many literatures about measurement error models for different types of measurement errors in continuous data, for example, the classical measurement additive error model, the Berkson error model [Fuller (1987); Carroll et al. (2006); Buzas, Tosteson, and Stefanskiet (2003)], equation error model [Kipnis et al. (1999); Kipnis et al. (2003)], regression calibration model [Mallick and Gelfand (1996)]. As far as the misclassified categorical data are concerned, the classic way to describe the misclassification from the true (laten, inherent) response T to the observed (manifest, error-prone) response Y is only based on a series of misclassification probabilities π_{uv} as given in Table 1.1. In this table, a subject from the vth class may be categorized into the uth category with probability π_{uv}. We refer to the misclassification model based on Table 1.1 as the descriptive misclassification (DMC) model. To the best of

Table 1.1: Misclassification from true category T into observed category Y with equal numbers of categories.

	true category (Y)			
observed category(T)	1	2	$\cdots$	$r+1$
1	π_{11}	π_{12}	$\cdots$	$\pi_{1,r+1}$
2	π_{21}	π_{22}	$\cdots$	$\pi_{2,r+1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$r+1$	$\pi_{r+1,1}$	$\pi_{r+1,2}$	$\cdots$	$\pi_{r+1,r+1}$

our knowledge, there is not an explicit misclassification model proposed for categorical data which clearly describe the relationship between the true response T and the observed response Y that is similar to the continuous data case. In this thesis, we propose such explicit misclassification models for mis-measured categorical data. More interestingly, this model can accommodate unbalanced misclassification which can be used to deal with a special type of missing values.

Aside from the case of misclassified categorical data, there are two explicit models for mis-measured count data. The first model is the additive measurement error for Poisson count data proposed by Cameron and Trivedi (1998). It is given by

$$Y = T + e, \tag{1.18}$$

In this model, both the true count T and the additive error e are assumed to be nonnegative random variables, for example, $T \sim Poisson(\eta)$ and $e \sim Poisson(\xi)$. So the nonnegative measurement error leads to a larger mean and variance relative to T. Therefore, it is useful only for describing count inflation.

Whittemore and Gong (1991) presented the other count error model for the misclassified count based on an example about mortality rate of cervical cancer. Let n_1 and n_2 denote, respectively, the correct and incorrect disease classification, and they are assumed to be independent Poisson variables. However, it is also assumed that, given the sum $n_1 + n_2$, n_1 follows a binomial distribution, that is $n_1 \sim b(n_1 + n_2, \pi^+)$, where π^+ is the sensitivity. Suppose that T and Y are the true count and the reported count of disease cases, and L is the population size which is assumed to be known. Then count error model for Y is given by

$$\begin{aligned} Y &\sim P(\mu^c), \\ \mu^c &= \pi^+ \eta = \pi^+ \lambda L, \end{aligned}$$

where η is the mean of T. This model allows for statistical inference on the disease rate λ [Whittemore and Gong (1991); Bratcher and Stamey (2002), Stamey et al. (2005); Brandi, Young and Stamey (2009)]. However, the model is only suitable for the case of perfect specificity, i.e. $\pi^- = 1$, but imperfect sensitivity, that is, $\pi^+ < 1$ [Cameron and Trivedi (1998) p. 307-312]. As we know, there is not an explicit model which can accommodate the overcount, undercount, and miscount with both imperfect sensitivity and specificity. In this thesis, one of our objectives is to develop such a model.

Classic approaches for analysis of longitudinal data are often based on the assumption that there are no measurement errors in the observations. In practice, it is often not the case. Therefore, there exist literatures studying the adverse effects of measurement errors. Most of them are about the measurement errors in covariates. Fuller (1987) conducted an extensive discussion on linear measurement errors models, and Carroll et al. (2006) investigated measurement errors in nonlinear models. Also Stefanski and Carroll (1985) and Stefanski (1985) studied the effects of mis-measured covariates in generalized linear models, especially for logistic model for binary data [also see Schafer (1987); Speiegelman, Rosner and Logan (2000); Hossain and Gustafson (2009); Rabe-Hesketh, Pickles and Skrondal (2003)].

There are also some literature focusing on the mis-measured responses. For example, Gustafson (2007, 2003), Roy, Banerjee and Maiti (2005), Roy and Banerjiee (2009), Rosychuk (1999), Rosychuk and Thompson (2001), Rosychuk and Islam (2009) and Neuhaus (1999, 2002) discussed the adverse effects of misclassification on binary responses. All of these literatures claim that failure to account for measurement errors in covariates or responses caused biased and inconsistent parameter estimates. Neuhaus (1999 and 2002) further gave a formula of approximate bias for scaler regression coefficient for misclassified binary response. The bias hence leads to erroneous conclusions to various degrees in health-related studies. To correct the attenuation in the estimates due to ignoring measurement errors, some approaches are proposed, for example the Bayesian method [Gustafson (2003); McGlothlin, Stamey and Seaman, (2008); Rosychuk and Islam (2009)], SIMEX method [Küchenhoff, Mwalili and Lesaffre (2006)] and the expected estimating equations method by [Wang et al. (2008)]. In addition, Roy, Banerjee and Maiti (2005) and Roy and Banerjee (2009) discussed a model-based approach which is used to deal with the misclassified binary response with covariates subject to measurement errors.

In this thesis, we develop the corrected GEE, quasi-likelihood and maximum likelihood methods to handle the statistical inference on mis-measured longitudinal categorical and count data.

1.3 Objective of This Thesis

As mentioned before, measurement errors widely occur in covariates and responses from studies in epidemiology, medicine, economics, and sociology. Simply ignoring measurement errors in either covariates or responses leads to biased estimation of model parameters and loss of power in detecting interesting association among variables. However, there are less discussions about measurement errors in discrete responses in longitudinal context, especially for count data. One important reason is the difficulty in developing estimation approaches due to the unobserved true responses on which the interesting associations are defined. Another possible reason is the lack of explicit measurement error models for categorical data and count data.

There are two main objectives of this thesis. The first one is to develop explicit measurement error models for categorical data and count data which can clearly describe the relationship between the inherent responses and the observed responses. The other objective is to develop approaches to consistently estimate the unknown model parameters for longitudinal categorical data and count data.

The remainder of this thesis is organized as follows. In Chapter 2, we introduce a generalized thinning operation which can be used to describe the transition between integer-valued variables. Based on the generalized thinning operation, an explicit model is proposed to characterize measurement errors due to misclassification in categorical data or multinomial data. The explicit model can be used to describe the balanced and unbalanced misclassification. Different from the classic misclassification model, the new model helps in the simple development of estimation approaches to obtain effective estimates of unknown parameters. Two new count error models are proposed in Chapter 2 to describe the error-contaminated count data. These two count error models can accommodate both the overnumerated and undernumberated count data.

As mentioned before, the effects or associations of interest are often defined on the true responses. For example, effects of covariates and association with past outcomes may be included in a transition model for true responses. Therefore, we introduce some first order transition models for dynamic categorical data and count data in Chapter 3. The thinning-operation-based transition model for categorical data is very flexible and it can accommodate various linear and nonlinear transition models. For count data, we will propose a non-stationary AR(1) model and a linear transition model which have wide applications in practice.

For the misclassified longitudinal categorical data, we consider a nonlinear transition model for true categorical response and the explicit misclassification model for the relationship between the latent response and the observed response in Chapter 4. We develop three approaches, namely the GQL, OGQL and ML approaches, to obtain unbiased estimates of model parameters. In the framework of the OGQL approach, we use information from both the first and second order responses to gain higher efficiency on the parameter estimations. Under the ML approach, we use the expectation & maximization (EM) algorithm to get estimates of model parameters. In practice, the data may contain missing information due to some participants' "unsure" responses to a specific question. This can be modeled by the unbalanced misclassification model. In Section 4.4 of Chapter 4, we investigate the modeling of the imperfect categorical data caused by classification errors and a special type of missing information.

In Chapter 5, we apply the binomial count error model and the corrected additive error model to the mis-measured longitudinal count data. In Section 5.2, we use a combination of the binomial count error model, a binomial model for true count of disease cases, and the linear transition model for dynamic population sizes. The corrected additive error model is used, together with the linear transition model for the true count response, to analyze the miscounted data in areas with little information about population sizes. The GEE and GQL approaches are employed to consistently estimate the model parameters. Finally, we present the conclusions and future studies in Chapter 6.

Chapter 2

Classification Error and Count Error Models

2.1 Overview

Measurement errors in discrete variables usually take the form of misclassification. For example, a patient infected by asthma might be misclassified into the healthy group and an individual who is free of asthma may be misdiagnosed as an asthma case. When the variable of interest and its observation are both categorical, the classification error is often not independent of the inherent categorical variable. To describe the misclassification between the true (inherent, or latent) variable T and its surrogate Y, the observed (manifest) variable, the classical models are defined by a series of classification probabilities. The relationship between T and Y can be described by Table 2.1. It can be seen that a subject from the vth class may be categorized into all of classes with a probability vector $(\pi_{1v}, \pi_{2v}, \cdots, \pi_{s+1,v})'$, where

Table 2.1: Misclassification from true category T into observed category Y with unequal numbers of categories.

	true category (Y)			
observed category(T)	1	2	$\cdots$	$r+1$
1	π_{11}	π_{12}	$\cdots$	$\pi_{1,s+1}$
2	π_{21}	π_{22}	$\cdots$	$\pi_{2,s+1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$s+1$	$\pi_{s+1,1}$	$\pi_{s+1,2}$	$\cdots$	$\pi_{s+1,r+1}$

π_{uv} is the probability that a member of the vth category is classified into the uth category. In general, $r = s$, which implies the numbers of the observed categories and the true categories are the same. But sometimes it might be the case that $r > s$ or $r < s$. Some examples are given in Section 2.3 in this chapter.

We now define a matrix consisting of the classification probabilities in Table 2.1 as follows:

$$\widehat{\Pi} = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1,r+1} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2,r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{s+1,1} & \pi_{s+1,2} & \cdots & \pi_{s+1,r+1} \end{pmatrix}. \tag{2.1}$$

In the theory of stochastic processes, T represents the state of a process at a specific time point j and Y be the state of this process at a time point k after j. The matrix $\widehat{\Pi}$ can be used to model the dynamic transition of this process from time j to k. In this case, $r = s$, and $\widehat{\Pi}$ is the so-called transition matrix.

In the classification context, we refer to the matrix $\widehat{\Pi}$ as the full misclassification matrix (FMC-matrix) due to the fact that $\sum_{u=1}^{s+1} \pi_{uv} = 1$ for any $v = 1, \ldots, r+1$. This also implies that the misclassification from T to Y can be completely characterized by a simplified matrix Π obtained by deleting the last row from $\widehat{\Pi}$, since $\pi_{s+1,v} = 1 - \sum_{u=1}^{s} \pi_{uv}$. The matrix Π can be given by

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1,r+1} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2,r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{s1} & \pi_{s2} & \cdots & \pi_{s,r+1} \end{pmatrix}, \tag{2.2}$$

and it is named the misclassification matrix (MC-matrix). We rewrite MC matrix as $\Pi = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{r+1}]$, where $\boldsymbol{\pi}_i$ is a column vector of dimension s, for $i = 1, \ldots, r+1$. Let $\Pi_r = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_r]$ be a submatrix of Π with the last column $\boldsymbol{\pi}_{r+1}$ deleted from Π. Then Π_r describes the classification from the first r categories of T to Y, and $\boldsymbol{\pi}_{r+1}$ reflects the classification from the $(r+1)$th category to Y.

In existing literatures, the FMC-matrix is used to capture the relationship between the latent variable T and the manifest response Y. The classic model is just a descriptive way to characterize the misclassifying relationship between T and Y, therefore, we name it the descriptive misclassification (DMC) model. To the best of our knowledge, there is not yet an explicit expression clearly formulating the dynamic relationship between T and Y like the classic error model or Berkson error model for the mis-measured continuous data. To bridge the gap, we propose such an explicit misclassification (EMC) model which addresses the connection between T and Y. This new misclassification model is based on a generalized thinning operation defined in the next section.

2.2 Generalized Thinning Operation

In integer-valued time series, a probabilistic operation called binomial thinning operation is often used to describe the transition between variables at different time points. This operation has been proposed by Steutel and Harn (1979) and applied to model time series by McKenzie (1985, 1986, 1988). For a discrete random variable N defined on the non-negative integers and a scalar π such that $0 \leq \pi \leq 1$, the random variable $\pi * N|_{N=n}$ is defined as $\sum_{k=1}^{n} b_k(\pi)$, where $\{b_k(\pi)\}$ is a sequence of independent identically distributed (i.i.d.) binary random variables with $P[b_k(\pi) = 1] = \pi = 1 - P[b_k(\pi) = 0]$. If $n = 0$, $\pi * 0$ is then defined as $\sum_{k=1}^{0} b_k(\pi) = 0$. Generally, we write the binomial thinning as $\pi * N = \sum_{k=1}^{N} b_k(\pi)$ and $\pi * 0 = 0$. Binomial thinning operation works as the integer-valued analogue to multiplication by a scalar π. More detailed discussion can be found in [Steutel, Vervaat and Wolfe (1983)].

As a generalization of the binomial thinning operation, a multinomial thinning operation has been introduced by McKenzie (1991 and 2003) to develop the vector-valued time series models. It is also denoted by $*$. If $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_s)'$ is a s-dimensional vector of probabilities with $\sum_{j=1}^{s} \pi_j \leq 1$, and N is a nonnegative integer-valued random variable, $\pi * N$ conditional on $N = n$ is defined to be a random vector from Multinomial$(n, \boldsymbol{\pi})$. The ith element of $\boldsymbol{\pi} * N$ is the number of outcomes of class i in n independent and identical trials where the probability of such an outcome in a trial is π_i. Therefore the operation can be rewritten as $\boldsymbol{\pi} * N|_{N=n} = \sum_{j=1}^{n} U_j$ and $\boldsymbol{\pi} * N|_{N=0} = \sum_{j=1}^{0} U_j = \mathbf{0}$ (which implies $\boldsymbol{\pi} * 0 = \mathbf{0}$), where $\{U_j\} \overset{iid}{\sim} Multinomial(1, \boldsymbol{\pi})$, and $\mathbf{0}$ is the vector of zeros. Actually, for a non-zero n, we can equivalently define $\boldsymbol{\pi} * N|_{N=n} = U$, where $U \sim Multinomial(n, \boldsymbol{\pi})$. The inequality $\mathbf{1}'\boldsymbol{\pi} < 1$ means that there will be $N - \mathbf{1}'U$ subjects being classified into the $(s+1)$th category, while $\mathbf{1}'\boldsymbol{\pi} = 1$ implies that all of these n subjects are classified into the first s categories and no subjects are classified into the $(s+1)$th category. In the case of $n = 1$ and the inequality $\mathbf{1}'\boldsymbol{\pi} < 1$, a zero vector U implies that this subject is classified into the $(s+1)$th category but not one of the first s categories. If $\mathbf{1}'\boldsymbol{\pi} = 1$, one and only one element of U is equal to 1, which means that the subject should be classified into one and only one of the first s categories. In addition, McKenzie (2003) also defined the multinomial thinning operation in matrix form. Suppose that we have a non-negative integer-valued vector $N = (N_1, N_2, \ldots, N_r)'$ and a $s \times r$ matrix $\Pi = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_r]$, the thinning operation is defined as

$$U_{s\times 1} = \Pi * N = \sum_{i=1}^{r} \boldsymbol{\pi}_i * N_i.$$

In this thesis, we let $A_{m\times n}$ denote a matrix A of dimension $m \times n$.

To comprehensively use the multinomial thinning operation $*$ to model the relationship between multiple categorical variables, we further generalize its definition, especially in form of matrix, as

Def. 2.1 $\boldsymbol{\pi}_{s\times 1} * N_{1\times 1} \triangleq U_{s\times 1}$ where $U = \sum_{i=1}^{N} U_i$ and $U_i \overset{iid}{\sim} Multinomial(1, \boldsymbol{\pi})$. The notation $\triangleq$ means "is defined as".;

Def. 2.2 $\Pi_{s\times r} * N_{r\times k} \triangleq \sum_{i=1}^{r} (\boldsymbol{\pi}_i * N_{i1}, \boldsymbol{\pi}_i * N_{i2}, \ldots, \boldsymbol{\pi}_i * N_{ik})_{s\times k}$;

Def. 2.3 If $N_{r\times k\times m}$ is a three-dimension array, the matrix in its jth folder is

$$(N_{..j})_{r\times k} = \begin{pmatrix} N_{11j} & N_{12j} & \cdots & N_{1kj} \\ N_{21j} & N_{22j} & \cdots & \pi_{2kj} \\ \vdots & \vdots & \ddots & \vdots \\ N_{r1j} & N_{r2j} & \cdots & N_{rkj} \end{pmatrix},$$

where N_{uvj} is a scalar. Let $U_{s\times k\times m} = \Pi_{s\times r} * N_{r\times k\times m}$, the matrix in the jth folder of three-dimensional array U is

$$\begin{aligned}(U_{..j})_{s\times k} &= \Pi_{s\times r} * (N_{..j})_{r\times k} \\ &= [\Pi_{s\times r} * (N_{.1j})_{r\times 1}, \Pi_{s\times r} * (N_{.2j})_{r\times 1}, \ldots, \Pi_{s\times r} * (N_{.kj})_{r\times 1}]\end{aligned}$$

Some special cases are given as follows:

1. If $k = 1$, the formula in Def. 2.2 becomes $\Pi_{s\times r} * N_{r\times 1} \triangleq \left(\sum_{i=1}^{r} \pi_i * N_i\right)_{r\times 1}$.
2. If $r = 1$, the formula in Def. 2.2 becomes $\pi_{s\times 1} * N_{1\times k} \triangleq (\pi * N_1, \pi * N_1, \ldots, \pi * N_k)_{s\times k}$, where $N = (N_1, N_2, \ldots, N_k)$ and N_i's are scalars.

Actually, the two-dimensional matrix $N_{r\times k}$ can be viewed as a reduced form of the three-dimensional matrix $N_{r\times k\times m}$ in the case that there is only one folder, i.e. $m = 1$. Therefore, the Def. 2.3 can accommodates the Def. 2.2. Furthermore, the generalized thinning operation $*$ accommodates the ordinary multinomial thinning operation according to the definition above hence further to the binomial thinning operation.

From the definition, it can be seen that the generalized thinning operation is similar to the multiplication product for matrices in the operation rules. In addition, the multinomial thinning operations have good properties and some are given below.

1. $\boldsymbol{\pi}*(N_1+N_2) \stackrel{d}{=} \boldsymbol{\pi}*N_1+\boldsymbol{\pi}*N_2$, where N_i, $i = 1,2$ are non-negative integer-valued scalars, and the notation $\stackrel{d}{=}$ means "identical in distribution".

2. $\Pi_{s\times r} * (N_{r\times k} + M_{r\times k}) \stackrel{d}{=} \Pi_{s\times r} * N_{r\times k} + \Pi_{s\times r} * M_{r\times k}$

3. If $Z_{s\times k} = \Pi_{s\times r} * Y_{r\times k}$ and $Y_{r\times k} = \Gamma_{r\times m} * X_{m\times k}$, then $Z = \Pi*(\Gamma*X) \stackrel{d}{=} (\Pi\Gamma)*X$

It is straightforward to prove the first two properties from the definition. Here, we just give the justification of the third one. We first show that the matrix $\Lambda_{s\times m} = \Pi_{s\times r}\Gamma_{r\times m}$ can work as a MC-matrix in the operation. Let $\Lambda = [\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_s]$, $\Pi = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_r]$, and $\Gamma = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s]$, where $\boldsymbol{\lambda}_i$'s, $\boldsymbol{\pi}_j$'s and $\boldsymbol{\gamma}_l$'s are column vectors of probabilities. The sum of the probabilities in $\boldsymbol{\pi}_j$ and $\boldsymbol{\gamma}_i$, respectively, do not exceed 1, that is, $\mathbf{1}'\boldsymbol{\pi}_u \leq 1$ and $\mathbf{1}'\boldsymbol{\gamma}_j \leq 1$. Next, we will show that the sum of each column of Λ is at most 1.

We have

$$\Lambda = \Pi_{s\times r}\Gamma_{r\times m} \Leftrightarrow \boldsymbol{\lambda}_j = \sum_{u=1}^{r} \boldsymbol{\pi}_u \gamma_{uj} \Leftrightarrow \lambda_{ij} = \sum_{u=1}^{r} \pi_{iu}\gamma_{uj}.$$

The sum of elements in jth column of Λ

$$\mathbf{1}'\boldsymbol{\lambda}_j = \mathbf{1}'(\sum_{u=1}^{r} \pi_u\gamma_{uj}) = \sum_{u=1}^{r} \mathbf{1}'\boldsymbol{\pi}_u\gamma_{uj} \leq \sum_{u=1}^{r} \gamma_{uj} = \mathbf{1}'\boldsymbol{\gamma}_j \leq 1.$$

Therefore, Λ can work as a MC-matrix in the thinning operation.

Next we prove that the third property is true for vectors $Z_{s\times 1}$, $Y_{r\times 1}$ and $X_{m\times 1}$. We use the moment generating function (mgf) technique to prove that $Z \stackrel{d}{=} \Lambda * X$.

Given that $Y = y$, $Z = \Pi * Y = \sum_{u=1}^{r} W_u$, where $W_u = \boldsymbol{\pi}_u * y_u$ is a variate of $Multinomial(y_u, \boldsymbol{\pi}_u)$. It is clear that W_u's are independent of each other given $Y = y$ for $u = 1, 2, \ldots, r$, and their moment generating functions (mgf) are given by $M_{W_u}(t) = (1 - \mathbf{1}'\boldsymbol{\pi}_u + \boldsymbol{\pi}_u' e^t)^{y_u}$. Similarly, given $X = x$, $Y = \sum_{j=1}^{m} U_j$ and $U_j = \boldsymbol{\gamma}_j * x_j \sim Multinomial(x_j, \boldsymbol{\gamma}_j)$ for $j = 1, 2, \ldots, m$ are independent of each other with mgf $M_{S_j}(t^*) = (1 - \mathbf{1}'\boldsymbol{\gamma}_j + \boldsymbol{\gamma}_j' e^{t^*})^{x_j}$. Therefore, given $X = x$, the mgf of Z is given by

$$
\begin{aligned}
M_{Z|X=x}(t) &= E(e^{Z't}|X = x) \\
&= E[E(e^{Z't}|Y, X = x)] \\
&= E[E(e^{\sum_{i=1}^{r} W_i' t}|Y, X = x)] \\
&= E[\prod_{i=1}^{r}(1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t)^{Y_i}|X = x], \text{where } e^t = (e^{t_1}, \ldots, e^{t_s})' \\
&= E[exp\{\sum_{i=1}^{r} Y_i log(1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t)\}|X = x)], \text{let } t_i^* = log(1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t) \\
&= \prod_{j=1}^{m}(1 - \mathbf{1}'\boldsymbol{\gamma}_j + \boldsymbol{\gamma}_j' e^{t^*})^{x_j}, \text{where, } t^* = (t_1^*, \ldots, t_r^*)' \text{ then } e^{t^*} = (e^{t_1^*}, \ldots, e^{t_r^*})' \\
&= \prod_{j=1}^{m}\left[1 - \mathbf{1}'\boldsymbol{\gamma}_j + \sum_{u=1}^{r} \gamma_{uj}(1 - \mathbf{1}'\boldsymbol{\pi}_u + \boldsymbol{\pi}_u' e^t)\right]^{x_j} \\
&= \prod_{j=1}^{m}\left[1 - \mathbf{1}'\boldsymbol{\gamma}_j + \sum_{u=1}^{r} \gamma_{uj}(1 - \mathbf{1}'\boldsymbol{\pi}_u) + \sum_{u=1}^{r}\sum_{v=1}^{s} \gamma_{uj}\pi_{vu} e^{t_v}\right]^{x_j} \\
&= \prod_{j=1}^{m}\left[1 - \sum_{v=1}^{r}\sum_{u=1}^{r} \gamma_{uj}\pi_{vu} + \sum_{v=1}^{s}(\sum_{u=1}^{r} \gamma_{uj}\pi_{vu}) e^{t_v}\right]^{x_j} \\
&= \prod_{j=1}^{m}\left(1 - \mathbf{1}'\boldsymbol{\lambda}_j + \sum_{v=1}^{r} \lambda_{vj} e^{t_v}\right)^{x_j} \\
&= \prod_{j=1}^{m}\left(1 - \mathbf{1}'\boldsymbol{\lambda}_j + \boldsymbol{\lambda}_j' e^t\right)^{x_j}.
\end{aligned}
$$

This means $Z \overset{d}{=} \Lambda * X$.

As far as the matrices $Z_{s \times k}$, $Y_{r \times k}$, $X_{m \times k}$ are concerned, according to the Def. 2.2,

$$Z_{s \times k} = \Pi_{s \times r} * Y_{r \times k} = [\Pi * Y_1, \ldots, \Pi * Y_u, \ldots, \Pi * Y_k],$$

and

$$Y_{r \times k} = \Gamma_{r \times m} * X_{m \times k} = [\Gamma * X_1, \ldots, \Gamma * X_v, \ldots, \Gamma * X_k].$$

It naturally follows that Z's ith column vector $Z_i = \Pi * Y_i \overset{d}{=} (\Pi\Gamma) * X_i = \Lambda * X_i$, hence $Z \overset{d}{=} \Lambda * X$.

Note: In order to describe the full transitions between the three categorical variables X, Y and Z. We denote the complete classification indicator vectors of X, Y and Z by $\widehat{X} = (X', 1 - \mathbf{1}'X)'$, $\widehat{Y} = (Y', 1 - \mathbf{1}'Y)'$ and $\widehat{Z} = (Z', 1 - \mathbf{1}'Z)'$, respectively. We further denote that the corresponding FMC-matrices are $\widehat{\Pi}$ and $\widehat{\Gamma}$, respectively. The sum of elements in each column of $\widehat{\Pi}$ and $\widehat{\Gamma}$ is equal to 1. Therefore the transition from $\widehat{X}$ to $\widehat{Y}$ is expressed as $\widehat{Y} = \widehat{\Gamma} * \widehat{X}$, which is equivalent to

$$Y = \Gamma * \widehat{X} = \Gamma_m * X + \gamma_{m+1} * (1 - \mathbf{1}'X),$$

Similarly, the transition from $\widehat{Y}$ to $\widehat{Z}$ is given by $\widehat{Z} = \widehat{\Pi} * \widehat{Y}$, and it is equivalent to

$$Z = \Pi * \widehat{Y} = \Pi_r * Y + \pi_{r+1} * (1 - \mathbf{1}'Y)$$

Hence, the transition from X to Z can be fully described as $\widehat{Z} = \widehat{\Pi}*(\widehat{\Gamma}*\widehat{X}) \overset{d}{=} (\widehat{\Pi}\widehat{\Gamma})*\widehat{X}$.

In the Def. 2.1, let $Y = \boldsymbol{\pi} * N$, it is easy to obtain the conditional expectation and variance of Y given N are given by

$$E(Y|N) = \boldsymbol{\pi} N, \tag{2.3}$$

$$Var(Y|N) = V_{\pi} N, \tag{2.4}$$

where $V_{\boldsymbol{\pi}}$ is defined as a diagonal matrix derived from a vector $\boldsymbol{\pi}$, that is, $V_{\boldsymbol{\pi}} \triangleq \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$. Similarly, in the Def. 2.2, let $Y_{s\times 1} = \Pi_{s\times r} * N_{r\times 1}$, the expectation and variance of Y given N can be given by

$$E(Y|N) = \Pi N, \tag{2.5}$$

$$Var(Y|N) = \sum_{i=1}^{s} N_i V_{\pi_i}. \tag{2.6}$$

Similar to the Kronecker product for matrices in algebra, we also define the Kronecker thinning operation $\circledast$ which may be useful in the future development based on the generalized thinning operation. The Kronecker thinning operation is defined as

1.

$$\boldsymbol{\pi} \circledast N \triangleq \begin{pmatrix} \pi * n_{11} & \pi * n_{12} & \cdots & \pi * n_{1k} \\ \pi * n_{21} & \pi * n_{22} & \cdots & \pi * n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \pi * n_{m1} & \pi * n_{m2} & \cdots & \pi * n_{mk} \end{pmatrix}_{sr \times k},$$

where $\boldsymbol{\pi}$ is a vector of dimension s, and N is a matrix of dimension $m \times k$. Specially, when $m = 1$, $\boldsymbol{\pi}_{s\times 1} \circledast N_{1\times k} = \boldsymbol{\pi}_{s\times 1} * N_{1\times k}$

2.

$$\Pi \circledast n \triangleq (\boldsymbol{\pi}_1 * n, \boldsymbol{\pi}_2 * n, \ldots, \boldsymbol{\pi}_r * n)_{s\times r},$$

where $\Pi = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_r]$ is a $s \times r$ matrix, and n is a non-negative integer.

3.

$$\Pi \circledast N \triangleq \begin{pmatrix} \Pi \circledast n_{11} & \Pi \circledast n_{12} & \cdots & \Pi \circledast n_{1k} \\ \Pi \circledast n_{21} & \Pi \circledast n_{22} & \cdots & \Pi \circledast n_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \Pi \circledast n_{m1} & \Pi \circledast n_{m2} & \cdots & \Pi \circledast n_{mk} \end{pmatrix}_{sm \times rk},$$

where Π is a $s \times r$ matrix, and N is a $m \times k$ matrix.

Example:

Suppose that we have longitudinal categorical data Y_{ij} of dimension s and its inherent variable T_{ij} of dimension r for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. We denote that $\widehat{T}_{ij} = (T'_{ij}, 1 - \mathbf{1}'T_{ij})'$ and $\widehat{Y}_{ij} = (Y'_{ij}, 1 - \mathbf{1}'Y_{ij})'$. There are three different ways to describe the misclassification between the observed and the inherent responses.

Case I: If our focus is on the transition between Y_{ij} and T_{ij}, it can be written as

$$\begin{aligned} Y_{ij} &= \Pi * \widehat{T}_{ij} \\ &= \Pi_r * T_{ij} + \boldsymbol{\pi}_{r+1} * (1 - \mathbf{1}'T_{ij}) \\ &= \sum_{u=1}^{r} \boldsymbol{\pi}_u * T_{ij(u)} + \boldsymbol{\pi}_{r+1} * (1 - \mathbf{1}'T_{ij}), \end{aligned}$$

or in an alternative way,

$$\widehat{Y}_{ij} = \widehat{\Pi} * \widehat{T}_{ij}.$$

Case II: If we are interested in the transition between $(Y_i)_{s\times J}$ and $(T_i)_{r\times J}$, we can write it as

$$Y_i = \Pi * \widehat{T}_i = (\Pi * \widehat{T}_{i1}, \Pi * \widehat{T}_{i2}, \ldots, \Pi_i * \widehat{T}_{iJ})_{s\times J},$$

or

$$\widehat{Y}_i = \widehat{\Pi} * \widehat{T}_i,$$

where $Y_i = [Y_{i1}, Y_{i2}, \ldots, Y_{iJ}]$ and $\widehat{Y}_i = [\widehat{Y}_{i1}, \widehat{Y}_{i2}, \ldots, \widehat{Y}_{iJ}]$, similarly, $T_i = [T_{i1}, T_{i2}, \ldots, T_{iJ}]$ and $\widehat{T}_i = [\widehat{T}_{i1}, \widehat{T}_{i2}, \ldots, \widehat{T}_{iJ}]$.

Case III: If we are interested in the transition between $Y_{s\times J\times I}$ and $T_{r\times J\times I}$, then we have

$$Y = \Pi * \widehat{T},$$

or

$$\widehat{Y} = \widehat{\Pi} * \widehat{T},$$

where the matrix in the ith folder of T, $\widehat{T}$, Y and $\widehat{Y}$ are, respectively, T_i, $\widehat{T}_i$, Y_i, and $\widehat{Y}_i$.

2.3 Classification Error Models

In this section, we introduce a classification error model with an explicit expression based on the generalized thinning operation. Let $Y_{s\times 1}$ and $T_{r\times 1}$ represent the observed and the inherent multinomial variables, where $T \sim Multinomial(N, p)$. Then $\boldsymbol{\eta} \triangleq E(T) = Np$. Let $\widehat{T} = (T', N - \mathbf{1}'T)$ and $\widehat{Y} = (Y', N - \mathbf{1}'Y)$ denote the full vectors of classification variables T and Y. The misclassification model for multinomial data can be expressed as

$$Y = \Pi * \widehat{T} = \Pi_r * T + \boldsymbol{\pi}_{r+1} * (N - \mathbf{1}'T). \tag{2.7}$$

In this thesis, we refer to model (2.7) as explicit multinomial misclassification (EMMC) model for multinomial data contaminated with classification errors. For the misclassified categorical data with $N = 1$, model (2.7) is named the explicit misclassification (EMC) model. For binomial variable T and Y, that is $r = s = 1$, we refer to model (2.7) as the explicit binomial misclassification (EBMC) model.

Based on the thinning operation, we can also build the marginal EMMC models for each element of Y. For the jth element Y_j, it can be written as

$$Y_j = \sum_{u=1}^{r} \pi_{ju} * T_u + \pi_{j,r+1} * (N - \mathbf{1}'T),$$

where $\Pi_{j.} = (\pi_{j1}, \pi_{j2}, \ldots, \pi_{j,r+1})'$ is the vector composed by elements in the jth row of Π. The marginal model for the reported count $\widehat{Y}_{s+1}$ of the subjects classified into the $(s+1)$th observed category of Y is given as

$$\widehat{Y}_{s+1} = \sum_{u=1}^{r} \pi_{s+1,u} * T_u + (1 - \mathbf{1}'\boldsymbol{\pi}_{r+1}) * (N - \mathbf{1}'T).$$

Notice that in the joint misclassification model (2.7), $\widehat{Y}_{s+1} = N - \mathbf{1}'Y$.

It should be pointed out that the s marginal misclassification models are not enough to describe the transition between the two categorical variables. This is because, given T, the Y_j's from the marginal models are independent and there are no constraints on the correlation between the Y_j's. For example, in the case of $N = 1$, both Y_{j_1} and Y_{j_2} may be 1 at the same time, which is impossible for a categorical variable. However, if we are only interested in a specific category, for example, the jth category, the marginal model for Y_j can be useful to describe how many subjects are classified into this category. The jth marginal model can completely describe the classification of all of the categories into the jth category.

Generally, the mean and variance of a misclassified categorical or multinational variable may be used in developing estimating equations to estimate model parameters. So, we give the mean and variance of Y as follows:

$$\begin{aligned}
\mu &= E(Y) \\
&= E[\Pi_r * T + \boldsymbol{\pi}_{r+1} * (N - \mathbf{1}'T)] \\
&= N\boldsymbol{\pi}_{r+1} + (\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')\boldsymbol{\eta} \\
&= N[\boldsymbol{\pi}_{r+1} + (\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')p], \qquad (2.8)
\end{aligned}$$

where $\boldsymbol{\eta} = E(T) = Np$, and

$$
\begin{aligned}
Var(Y) &= E[Var(Y|T)] + Var[E(Y|T)] \\
&= E[\sum_{i=1}^{r} V_{\boldsymbol{\pi}_i} T_i] + Var((\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')T) \\
&= \sum_{i=1}^{r} V_{\boldsymbol{\pi}_i} \boldsymbol{\eta}_i + V_{\boldsymbol{\pi}_{r+1}}(N - \mathbf{1}'\boldsymbol{\eta}) + (\Pi_{r-1} - \boldsymbol{\pi}_r\mathbf{1}')Var(T)(\Pi_{r-1} - \boldsymbol{\pi}_r\mathbf{1}')' \\
&= \sum_{i=1}^{r} NV_{\boldsymbol{\pi}_i} p_i + NV_{\boldsymbol{\pi}_{r+1}}(1 - \mathbf{1}'p) \\
&\quad + N(\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')V_p(\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')'. \qquad (2.9)
\end{aligned}
$$

Let $q = \boldsymbol{\pi}_{r+1} + (\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')p$, it can be shown from the following mgf of Y that the error-prone variable Y also follows a multinomial distribution with a probability vector q, that is $Y \sim Multinomial(N, q)$. The mgf of Y can be calculated as

$$
\begin{aligned}
M_Y(t) &= E(exp(Y't)) = E[E(exp(Y't)|T)] \\
&= E(exp(\sum_{i=1}^{r+1} B_i t)), \text{where } M_i = \boldsymbol{\pi}_i * T_i,\ i = 1, \ldots, r \text{ and } M_{r+1} = \boldsymbol{\pi}_{r+1} * (N - \mathbf{1}'T) \\
&= E[\prod_{i=1}^{r} (\frac{1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t}{1 - \mathbf{1}'\boldsymbol{\pi}_{r+1} + \boldsymbol{\pi}_{r+1}' e^t})^{T_i} (1 - \mathbf{1}'\boldsymbol{\pi}_{r+1} + \boldsymbol{\pi}_{r+1}' e^t)^N] \\
&= E[exp\{\sum_{i=1}^{r} T_i log(\frac{1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t}{1 - \mathbf{1}'\boldsymbol{\pi}_{r+1} + \boldsymbol{\pi}_{r+1}' e^t})\} (1 - \mathbf{1}'\boldsymbol{\pi}_{r+1} + \boldsymbol{\pi}_{r+1}' e^t)^N] \\
&= [1 - \mathbf{1}'p + \sum_{i=1}^{r} (p_i \frac{1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t}{1 - \mathbf{1}'\boldsymbol{\pi}_{r=1} + \boldsymbol{\pi}_{r+1}' e^t})]^n (1 - \mathbf{1}'\boldsymbol{\pi}_{r+1} + \boldsymbol{\pi}_{r+1}' e^t)^N \\
&= [(1 - \mathbf{1}'p)(1 - \mathbf{1}'\boldsymbol{\pi}_{r+1} + \boldsymbol{\pi}_{r+1}' e^t) + \sum_{i=1}^{r} (p_i(1 - \mathbf{1}'\boldsymbol{\pi}_i + \boldsymbol{\pi}_i' e^t)]^N \\
&= [(1 - \mathbf{1}'\boldsymbol{\pi}_{\mathbf{r+1}})(1 - \mathbf{1}'p) + \sum_{i=1}^{r} p_i(1 - \mathbf{1}'\boldsymbol{\pi}_i) + \sum_{i=1}^{r} p_i \boldsymbol{\pi}_i' e^t + (1 - \mathbf{1}'p)\boldsymbol{\pi}_{r+1}' e^t]^N \\
&= [1 - \mathbf{1}'q + q'e^t]^N,
\end{aligned}
$$

where q is the vector of multinomial probabilities of dimension s.

Table 2.2: Balanced misclassification of air quality in an environmental study.

	True Level (T)		
Classified Level (Y)	H (1)	M (2)	L(3)
H (1)	π_{11}	π_{12}	π_{13}
M(2)	π_{21}	π_{22}	π_{23}
L(3)	π_{31}	π_{32}	π_{33}

Therefore, it follows that

$$E(Y) = Nq, \tag{2.10}$$

$$Var(Y) = NV_q. \tag{2.11}$$

Comparing the expression (2.9) with expression (2.11), V_q should be equal to $\sum_{i=1}^{r} V_{\pi_i} p_i + V_{\pi_{r+1}}(1 - \mathbf{1}'p) + (\Pi_r - \pi_{r+1}\mathbf{1}')V_p(\Pi_r - \pi_{r+1}\mathbf{1}')'$.

Generally, the true variable T and the observed variable Y in the EMMC model (2.7) often have equal total numbers of categories, i.e. $r+1 = s+1$, leading to $r = s$. We refer to this case as the balanced misclassification (BMC). For example, in the environmental studies given in Table 2.2, air quality in an area can be categorized into three levels high (H), medium (M) and low (L). There may be classification errors involved in the categorical data due to imperfect measurement instruments and procedures.

In some cases, $r < s$, which implies that there are more observed categories than the true categories. We name this type of classification as the unbalanced misclassification (UBMC). For example, in diagnosis of a kind of epidemic disease given in Table

Table 2.3: Unbalanced misclassification with $r < s$ in an example of asthma study.

	Disease (T)		
Diagnosis of test (Y)	Positive (1)	Negative (2)	Suspected(3)
Infected (1)	π_{11}	π_{12}	π_{13}
Healthy (2)	π_{21}	π_{22}	π_{23}

2.3, an individual may be diagnosed as "Positive" which means "infected", or "Negative" which implies "healthy", or even diagnosed as "suspected". In some situations, the arising of the extra category on observed variable Y may be due to incomplete information such as an "unsure" answer from a questionnaire. For example, in the study of the children asthma, the parents may give an "unsure" response about their children' asthma statuses. Another example in economic studies which focus on the level of annual incomes of individuals, some people may refuse to answer the question about the level of income in the past year. However in the questionnaires, they may provide some other information like the occupation, age, working experience, education level etc. Based on the supplementary information, we can make an estimate of the probability that his/her income belongs to either of the high, medium and low level.

In addition, model (2.7) can be used to describe the relationship between latent genotypes and manifest phenotypes. For example, it is known that there are four blood types for humans, that is, A, B, AB and O, and three types of related genes, i.e. A, B and O. All of possible genotype of an individual are AA, AB, BB, AO, BO and OO. In genetics, it is well known that both the genes A and B are dominant over

Table 2.4: Blood types and their genotypes, an example of unbalanced misclassification with $s < r$

	Genotype (T)					
Blood type (Y)	AA	AB	BB	AO	BO	OO
A	1	0	0	1	0	0
B	0	0	1	0	1	0
AB	0	1	0	0	0	0
O	0	0	0	0	0	1

the gene O. Therefore, the relationship between the genotypes and blood types can be described by Table 2.4. Actually, in practice, the blood type of an individual may be misclassified due to the mistakes by unexperience examiners, which means some π_{uv}'s for $u \neq v$ may be not zeros.

From the discussions above, it can be seen that the new EMMC model (2.7) can be widely used to characterize different classification patterns for categorical or multinomial data.

2.4 Count Error Models

In epidemiologic studies, data like the total of patients infected by a kind of epidemic disease in different areas are frequently used to evaluate the environmental effects on population health. These data are often collected by some survivance systems. For example, the new cancer cases by state in USA can be obtained from

SEER. But the data are prone to errors due to different reasons mentioned in Section 1.2 of Chapter 1.

Statistical analysis which takes measurement errors in count data into consideration is of great interest. However, as mentioned in Chapter 1, there is not yet a count error model which can accommodate both the overnumerated and undernumerated count data for imperfect sensitivity and specificity. Therefore, in this section, we develop the measurement error models for count data.

2.4.1 Multinomial count error model

In the previous section, we discussed the classification error model for categorical and multinomial data. For the misclassified multinomial data, we assume that the population size N in a district is fixed and known, which is reasonable in a censal year. However, the population size may be unknown in the intercensal years. Further, it can be unfixed due to emigration, immigration, death, birth and so on. It is assumed that $N \sim (\phi, \tau^2)$, where ϕ and τ^2 are the mean and variance of N, respectively. Then the reported count of disease cases Y and the true count T in an open area with an unknown and random population size N can be modeled by

$$Y = \Pi * \widehat{T} = \Pi_r * T + \boldsymbol{\pi}_{r+1} * (N - \mathbf{1}'T) \tag{2.12}$$

We name this model as the multinomial count error model for a random N.

From expressions (2.11) and (2.12) in section 2.3, we derived the conditional expectation and variance of Y given the population size N to be:

$$E(Y|N) = N[\boldsymbol{\pi}_r + (\Pi_r - \boldsymbol{\pi}_{r+1}\mathbf{1}')p],$$

and

$$
\begin{aligned}
Var(Y|N) &= \sum_{i=1}^{r} NV_{\pi_i} p_i + NV_{\pi_{r+1}}(1 - \mathbf{1}'p) \\
&\quad + N(\Pi_r - \pi_{r+1}\mathbf{1}')Var(T|N)(\Pi_r - \pi_{r+1}\mathbf{1}')',
\end{aligned}
$$

where $Var(T|N) = NV_p$. So the unconditional expectation and variance of Y based on the assumptions on N can be given as

$$
\mu = E[E(Y|N)] = \phi[\pi_r + (\Pi_r - \pi_{r+1}\mathbf{1}')p], \tag{2.13}
$$

and

$$
\begin{aligned}
Var(Y) &= E[Var(Y|N)] + Var[E(Y|N)] \\
&= E\{\sum_{i=1}^{r} NV_{\pi_i} p_i + NV_{\pi_{r+1}}(1 - \mathbf{1}'p) \\
&\quad + (\Pi_r - \pi_{r+1}\mathbf{1}')Var(T|N)(\Pi_r - \pi_{r+1}\mathbf{1}')'\} \\
&\quad + Var[N[\pi_r + (\Pi_r - \pi_{r+1}\mathbf{1}')p]] \\
&= \phi[\sum_{i=1}^{r} V_{\pi_i} p_i + V_{\pi_{r+1}}(1 - \mathbf{1}'p) \\
&\quad + (\Pi_r - \pi_{r+1}\mathbf{1}')V_p(\mathbf{\Pi}_r - \pi_{r+1}\mathbf{1}')'] \\
&\quad + [\pi_r + (\Pi_r - \pi_{r+1}\mathbf{1}')p]Var(N)[\pi_{r+1} + (\Pi_r - \pi_{r+1}\mathbf{1}')p]'.
\end{aligned}
$$

Alternatively, if we let $q = \pi_{r+1} + (\Pi_r - \pi_{r+1}\mathbf{1}')p$, we can get that, given $N = n$, $Y \sim multinomial(n, q)$ from the previous section. Therefore, the unconditional variance of Y has another form, that is

$$
\begin{aligned}
Var(Y) &= E[Var(Y|N)] + Var[E(Y|N)] \\
&= E[NV_q + Var(qN) \\
&= \phi V_q + qVar(N)q'.
\end{aligned} \tag{2.14}
$$

Now we consider a special case that the population size $N \sim Poisson(\phi)$. Suppose that T_j represents the count of subjects belonging to the jth category among the population in an area, where $j = 1, 2, \ldots, r+1$. So given $N = n$, $T \sim multinomial(n, p)$ and $T_j \sim b(n, p_j)$, where $\sum_{j=1}^{r+1} p_j = 1$, hence $\sum_{j=1}^{r+1} T_j = N$.

The joint moment generating function of T can be developed as follows:

$$
\begin{aligned}
M_T(t) &= E(e^{T't}) \\
&= E(e^{\sum_{j=1}^{r+1} T_j t_j}) \\
&= E[E(e^{\sum_{j=1}^{r+1} T_j t_j} | N)] \\
&= E[(\sum_{j=1}^{r+1} p_j e^{t_j})^N] \\
&= E[exp\{N log(\sum_{j=1}^{r+1} p_j e^{t_j})\}] \\
&= exp(\phi[\sum_{j=1}^{r+1} p_j e^{t_j} - 1]) \\
&= e^{\phi \sum_{j=1}^{r+1} p_j (e^{t_j} - 1)} \\
&= \prod_{j=1}^{r+1} M_{T_j}(t_j).
\end{aligned}
$$

This mgf implies that, unconditionally, T_j $j = 1, 2, \ldots, r+1$ are independent Poisson variables and $T_j \sim Poisson(\eta_j = \phi p_j)$. Similarly, we can also conclude that Y_l, the count of subjects classified into the lth category, for $l = 1, 2, \ldots, s+1$, are also independent Poisson variables and $Y_l \sim Poisson(\mu_l = \phi q_l)$. This means that under the assumption that the population size N follows a Poisson distribution, the counts of subjects falling into different categories are independent of each other.

In an open area, there are two situations leading to a dynamic population size. The first is migration including immigration and emigration, and the other is the natural growth or decrease of the population due to birth and death. In both situations, the population size N can be assumed to be random. If one is interested in modeling the count of people who are attacked by an epidemic disease, for example, asthma in an open district, then the population can be partitioned into two subpopulations, healthy group and infected group. This leads to binomial count data with a random population size N. We will present some results for this kind of data in the following paragraphs.

As shown by Figure 2.1, T denotes the true count of infected subjects in an area, and Y be the observed count of reported disease cases from some registration systems. Suppose that, given the population size $N = n$, $T \sim b(n,p)$, where, p is the true disease rate in this area. The relationship between T and Y is given by

$$Y = \pi^+ * T + (1 - \pi^-) * (N - T), \tag{2.15}$$

where π^+ is the sensitivity and π^- is the specificity in Table 2.5. We can conclude that given $N = n$, $Y|N = n \sim b(n,q)$, where $q = 1 - \pi^- + (\pi^- + \pi^+ - 1)p$ is the reported disease rate in this area. The marginal distribution of Y is a Poisson distribution, that is, $Y \sim P(\mu = q\phi)$. Actually, $T \stackrel{d}{=} p * N$, similarly $Y \stackrel{d}{=} q * N$.

Table 2.5: Example of misclassified disease cases

	Disease cases			
Reported cases	Healthy $(1-p)$	Count $(N-T)$	Infected (p)	Count (T)
Negative	π^-	$N-T-(1-\pi^-)*(N-T)$	$1-\pi^+$	$T-\pi^+*T$
Positive	$1-\pi^-$	$(1-\pi^-)*(N-T)$	π^+	π^+*T

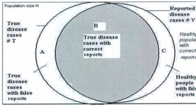


Figure 2.1: The True and Reported Disease Cases in An Area

2.4.2 Corrected additive count error models

From the last paragraph in Section 2.4.1, the size of infected subpopulation is independent of the size of healthy subpopulation under the assumption that the population size N follows a Poisson distribution. The sizes of the two subpopulations follow two different Poisson distributions. We let T denote the true size of infected population and T^0 represent the size of healthy population in an area. Similarly, let Y and Y^0 denote the reported size of infected population and healthy population, respectively. So the size of the total population in this area $N = T + T^0 = Y + Y^0$. Under the assumption that $N \sim Poisson(\phi)$, T is independent of T^0, and Y is independent of Y^0. Furthermore, $T \sim Poisson(p\phi)$ and $T^0 \sim Poisson((1-p)\phi)$, similarly, $Y \sim Poisson(q\phi)$ and $Y^0 \sim Poisson((1-q)\phi)$, where $q = 1 - \pi^- + (\pi^+ + \pi^- - 1)p$. We rewrite the binomial count error model (2.15) in an alternative form because we do not have any knowledge of ϕ in the Poisson distribution of N. The new expression is given by

$$Y = \pi^+ * T + (1 - \pi^-) * T^0. \tag{2.16}$$

In the expression (2.16), the count of infected subjects being correctly classified $\pi^+ * T$ is independent of the count of healthy subjects being misclassified into infected category $(1 - \pi^-) * T^0$. If we let $e = (1 - \pi^-) * T^0$, the new count error model can be rewritten as

$$Y = \pi^+ * T + e, \tag{2.17}$$

where π^+ is the sensitivity. We call this model as the corrected additive model when it is compared with the additive model (1.18) which was discussed by Cameron and Trivedi (1998).

Therefore, we can apply the model (2.17) to model the miscounted disease cases in an area with unknown population size. We suppose that the unknown population size follows a Poisson distribution. This is a popular and reasonable assumption in practice, but we may not have any knowledge about its expectation. In this situation, it can be rationalely assumed that the number of individuals misdiagnosed as infected cases among healthy subpopulation in an open area is independent of the number of correctly reported disease cases $\pi^+ * T$. We further assume that $T \sim Poisson(\eta)$ where $\eta = exp(x'\beta)$ where x represents covariates associated with the true count of disease cases, for example, the environmental exposures. The additive error $e \sim Poisson(\psi)$ with $\psi = exp(z'\alpha)$ where z are covariates related to the miscount from healthy people into the infected population, for example, the health care level of the medical facilities in this area. There may be some common covariates shared by x and z.

The expectation of Y is given by the following expression:

$$\mu = E(Y) = E[E(Y|T)] = \pi^+\eta + \psi, \tag{2.18}$$

and the variance of Y is formulated by

$$\begin{aligned} Var(Y) &= Var[E(Y|T)] + E[Var(Y|T)] \\ &= Var(\pi^+T + \psi) + E[\pi(1-\pi)T + \psi] \\ &= (\pi^+)^2 Var(T) + \pi^+(1-\pi^+)\eta + \psi \\ &= \mu + (\pi^+)^2[Var(T) - \eta]. \end{aligned}$$

It is easy to see that in the corrected additive count error model (2.17), the expectation of Y can be greater than the expectation of T, that is, $\mu > \eta$ when $\psi > (1 - \pi^+)\eta$.

On the other hand, the expectation of Y can be smaller than the expectation of T, that is, $\mu < \eta$ when $\psi < (1 - \pi^+)\eta$. Therefore, the corrected additive error model (2.17) can accommodate both the overcounted and undercounted data, with imperfect sensitivity ($\pi^+ < 1$) and imperfect specificity ($\psi > 0$). In addition, if $\psi = 0$ leading to $e = 0$, this implies that no healthy people were misclassified into the infected group. Therefore, this model can accommodate the case of perfect specificity and, definitely, the perfect sensitivity ($\pi^+ = 1$).

It should be pointed out that, in the corrected additive count error model (2.17), the assumption that population size follows Poisson distribution may be violated when we focus on a specific subpopulation. For example, if we know the total population size N in a district, but we do not know the size of the subpopulation at least 50 years old N_1. Then N_1 can be assumed to follow a binomial distribution with probability ρ, that is, $N_1 \sim b(N, \rho)$. In this case, the size of the healthy group T^0 and the size of the group infected by lung cancer T in this subpopulation are not independent variables. In this situation, the miscount data can be modeled by model (2.12) if we know the probability ρ.

However, even we do not know the distribution of the population size of an area in an intercensal year in practice, it is a popular assumption that the sizes of the healthy and the infected groups are independent of each other. The count of the misreported disease cases among healthy subpopulation $e = \pi^- * T^0$ is often assumed to be an approximate Poisson variable taking into account the approximation of binomial variable to a Poisson distribution when the size of the healthy subpopulation is very large. Therefore, the corrected additive error model (2.17) is still applicable when the distribution of the population size is unknown.

Chapter 3

Longitudinal Transition Models for Categorical Data and Count Data

3.1 Transition Models for Categorical Data

3.1.1 A transition model for dynamic categorical data

In this section, we develop a transition model for dynamic categorical data based on the generalized thinning operation. This transition model has similar structure as the explicit misclassification model (2.7) with $N = 1$ in Section 2.3 of Chapter 2.

In a longitudinal study, we let t_{i0} denote the baseline observation of the categorical variable $T_{ij} = (T_{ij(1)}, \dots, T_{ij(u)}, \dots, T_{ij(r)})'$ for the ith subject in a longitudinal study, where $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$. We define a matrix $\tilde{\eta}_{ij} = \left(\tilde{\eta}_{ij(u,v)}\right)_{r \times r}$ with element $\tilde{\eta}_{ij(u,v)} = P(T_{ij(u)} = 1 | t_{i,j-1(v)} = 1)$, that is the probability of the transition from the vth state at the $(j-1)$th time point to the uth state at the jth time point, for $u, v = 1, 2, \dots, r$. Define the vector $\bar{\eta}_{ij} = \left(\bar{\eta}_{ij(u)}\right)_{r \times 1}$ with elements

Table 3.1: Transition probabilities from $T_{i,j-1}$ to T_{ij}

	$T_{i,j-1}$				
T_{ij}	1	2	$\cdots$	r	$r+1$
1	$\bar{\eta}_{ij(1,1)}$	$\bar{\eta}_{ij(1,2)}$	$\cdots$	$\bar{\eta}_{ij(1,r)}$	$1-\mathbf{1}'\bar{\eta}_{ij(1,\cdot)}$
2	$\bar{\eta}_{ij(2,1)}$	$\bar{\eta}_{ij(2,2)}$	$\cdots$	$\bar{\eta}_{ij(r,2)}$	$1-\mathbf{1}'\bar{\eta}_{ij(2,\cdot)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$\bar{\eta}_{ij(r,1)}$	$\bar{\eta}_{ij(r,2)}$	$\cdots$	$\bar{\eta}_{ij(r,r)}$	$1-\mathbf{1}'\bar{\eta}_{ij(r,\cdot)}$
$r+1$	$\bar{\eta}_{ij(1)}$	$\bar{\eta}_{ij(2)}$	$\cdots$	$\bar{\eta}_{ij(r)}$	$1-\mathbf{1}'\bar{\eta}_{ij}$

$\bar{\eta}_{ij(u)} = P(T_{ij(u)} = 1|\mathbf{1}'t_{i,j-1} = 0)$, that is, the probability of the transition from the state $r+1$ at time $j-1$ to the uth state at time j, where $u = 1, 2, \ldots, r$. In addition, the probability of transiting from state u at time $j-1$ to state $r+1$ at the next time point is equal to $1-\mathbf{1}'\bar{\eta}_{ij(\cdot,u)}$. Similarly, the probability that a subject with state $r+1$ at time $j-1$ keeps his/her state at the next time point j is $1-\mathbf{1}'\bar{\eta}_{ij}$.

Table 3.1 shows the transition probabilities of the ith subject's state from the $(j-1)$th time point to the jth time point. Similar to the misclassification problem for categorical data described in Section 2.3 of the previous Chapter, the full transition matrix is given by

$$\widehat{\Lambda}_{ij} = \begin{bmatrix} \bar{\eta}_{ij} & \bar{\eta}_{ij} \\ \mathbf{1}' - \mathbf{1}'\bar{\eta}_{ij} & 1 - \mathbf{1}'\bar{\eta}_{ij} \end{bmatrix} = \begin{pmatrix} \bar{\eta}_{ij(1,1)} & \cdots & \bar{\eta}_{ij(1,r)} & \bar{\eta}_{ij(1)} \\ \bar{\eta}_{ij(2,1)} & \cdots & \bar{\eta}_{ij(2,r)} & \bar{\eta}_{ij(2)} \\ \vdots & \ddots & \vdots & \vdots \\ \bar{\eta}_{ij(r,1)} & \cdots & \bar{\eta}_{ij(r,r)} & \bar{\eta}_{ij(r)} \\ 1 - \mathbf{1}'\bar{\eta}_{ij(\cdot,1)} & \cdots & 1 - \mathbf{1}\bar{\eta}_{ij(\cdot,r)} & 1 - \mathbf{1}'\bar{\eta}_{ij} \end{pmatrix}, \tag{3.1}$$

and a simplified transition matrix can be defined as

$$\Lambda_{ij} = [\bar{\eta}_{ij}, \bar{\eta}_{ij}] = \begin{pmatrix} \bar{\eta}_{ij(1,1)} & \cdots & \bar{\eta}_{ij(1,r)} & \bar{\eta}_{ij(1)} \\ \bar{\eta}_{ij(2,1)} & \cdots & \bar{\eta}_{ij(2,r)} & \bar{\eta}_{ij(2)} \\ \vdots & \ddots & \vdots & \vdots \\ \bar{\eta}_{ij(r,1)} & \cdots & \bar{\eta}_{ij(r,r)} & \bar{\eta}_{ij(r)} \end{pmatrix}. \tag{3.2}$$

By defining $\widetilde{T}_{ij} = (T_{ij}, 1 - \mathbf{1}'T_{ij})$, the new transition model based on the generalized thinning operation can be defined as

$$\begin{aligned} T_{ij} &= \Lambda * \widehat{T}_{i,j-1} \\ &= \bar{\eta}_{ij} * T_{i,j-1} + \bar{\eta}_{ij} * (1 - \mathbf{1}'T_{i,j-1}). \end{aligned} \tag{3.3}$$

The first part on the right side of the model (3.3), $\bar{\eta}_{ij} * T_{i,j-1}$, denotes the transition from the first r states, and the second part $\bar{\eta}_{ij} * (1 - \mathbf{1}'T_{i,j-1})$ represents the transition from the last state $r+1$ at time point $j-1$.

We next present some useful results of calculations about expectations, variances, and covariances. Firstly, based on model (3.3), it is easy to see that the conditional expectation of T_{ij} given the previous $T_{i,j-1}$ is

$$\begin{aligned} E(T_{ij}|T_{i,j-1}) &= \bar{\eta}_{ij}T_{i,j-1} + \bar{\eta}_{ij}(1 - \mathbf{1}'T_{i,j-1}) \\ &= (\bar{\eta}_{ij} - \bar{\eta}_{ij}\mathbf{1}')T_{i,j-1} + \bar{\eta}_{ij}. \end{aligned} \tag{3.4}$$

We further have the expectation of T_{ij} given T_{ik} for $k < j$

$$E(T_{ij}|T_{ik}) = \prod_{v=k+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')T_{ik} + \sum_{l=k+1}^{j} \prod_{v=l+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\bar{\eta}_{il}. \tag{3.5}$$

It should be noticed that, in this section, $\prod_{v=k}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')$ for $k \leq j$ is defined as $(\bar{\eta}_{ij}-\bar{\eta}_{ij}\mathbf{1}')(\bar{\eta}_{i,j-1}-\bar{\eta}_{i,j-1}\mathbf{1}')\cdots(\bar{\eta}_{ik}-\bar{\eta}_{ik}\mathbf{1}')$ but not $(\bar{\eta}_{ik}-\bar{\eta}_{ik}\mathbf{1}')(\bar{\eta}_{i,v+1}-\bar{\eta}_{i,k+1}\mathbf{1}')\cdots(\bar{\eta}_{ij}-\bar{\eta}_{ij}\mathbf{1}')$.

The unconditional expectations can be given as

$$\begin{aligned} \eta_{ij} &\triangleq E(T_{ij}) \\ &= (\bar{\eta}_{ij} - \bar{\eta}_{ij}\mathbf{1}')\eta_{ij-1} + \bar{\eta}_{ij} &(3.6) \\ &= \prod_{v=k+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\eta_{ik} + \sum_{l=k+1}^{j} \prod_{v=l+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\bar{\eta}_{il}. &(3.7) \end{aligned}$$

For any $k < j$, the expectation of the pairwise product $T_{ij}T'_{ik}$ is

$$\begin{aligned} E(T_{ij}T'_{ik}) &= E[E(T_{ij}|T_{ik})T_{ik}] \\ &= \prod_{v=k+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')E(T_{ik}T'_{ik}) + \sum_{l=k+1}^{j} \prod_{v=l+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\bar{\eta}_{il}\eta'_{ik}. \end{aligned} \tag{3.8}$$

Hence, the covariance between T_{ij} and T_{ik} is

$$\begin{aligned} Cov(T_{ij}, T_{ik}) &= E(T_{ij}T'_{ik}) - E(T_{ij})E(T'_{ik}) \\ &= \prod_{v=k+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')E(T_{ik}T'_{ik}) + \sum_{l=k+1}^{j} \prod_{v=l+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\bar{\eta}_{il}\eta'_{ik} \\ &\quad - \prod_{v=k+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\eta_{ik}\eta'_{ik} + \sum_{l=k+1}^{j} \prod_{v=l+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')\bar{\eta}_{il}\eta'_{ik} \\ &= \prod_{v=k+1}^{j} (\bar{\eta}_{iv} - \bar{\eta}_{iv}\mathbf{1}')Var(T_{ik}). \end{aligned} \tag{3.9}$$

For categorical variable $T_{ij} \sim multinomial(1, \eta_{ij})$, it is obvious that

$$Var(T_{ij}) = V_{\eta_{ij}} = \mathrm{diag}(\eta_{ij}) - \eta_{ij}\eta_{ij}'.$$

If we rewrite $T_i = (T_{i1}', T_{i2}', \ldots, T_{iJ}')_{rJ\times 1}'$ and let $\eta_i = E(T_i)$, the variance-covariance matrix of T_i can be written as

$$\Sigma_i = \begin{pmatrix} \Sigma_{i11} & \Sigma_{i12} & \cdots & \Sigma_{i1J} \\ \Sigma_{i21} & \Sigma_{i22} & \cdots & \Sigma_{i2J} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{iJ1} & \Sigma_{iJ2} & \cdots & \Sigma_{iJJ} \end{pmatrix}_{rJ\times rJ}, \tag{3.10}$$

where $\Sigma_{jk} = \Sigma_{kj}' = Cov(T_{ij}, T_{jk})$ for $j \neq k$, and $\Sigma_{jj} = Var(T_{ij}) = V_{\eta_{ij}}$.

Model (3.3) can accommodate various transition models based on different assumptions on $\widehat{\Lambda}_{ij}$ or Λ_{ij}. In addition, Λ_{ij} can be a constant matrix over time for any subject, or it can be a matrix function of some covariates, even time-varying covariates. For example, in the full transition matrix Λ_{ij}, we can suppose that

$$\tilde{\eta}_{ij(u,v)} = \frac{exp(x_{ij}'\beta_u + \gamma_{uv})}{1 + \sum_{l=1}^{r} exp(x_{ij}'\beta_l + \gamma_{lv})}, \tag{3.11}$$

and

$$\bar{\eta}_{ij(u)} = \frac{exp(x_{ij}'\beta_u)}{1 + \sum_{l=1}^{r} exp(x_{ij}'\beta_l)}. \tag{3.12}$$

In expression (3.11), $x_{ij} = (x_{ij(1)}, \ldots, x_{ij(p)})'$ is a vector consisting of p explanatory variables. The parameter matrix

$$\mathscr{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1r} \\ \vdots & \cdots & \vdots \\ \beta_{p1} & \cdots & \beta_{pr} \end{pmatrix}$$

with $\beta_u = (\beta_{1u}, \ldots, \beta_{pu})'$ denoting the effects of covariates, and

$$\Upsilon = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1r} \\ \vdots & \cdots & \vdots \\ \gamma_{r1} & \cdots & \gamma_{rr} \end{pmatrix}$$

denotes the dynamic dependence.

Indeed, the transition model (3.3) can accommodate the following model (3.13-3.14) in nature. The latter one is just an alternative form of the special case of model (3.3) based on the assumptions (3.11) and (3.12). The model is given by

$$\eta_{ij}^c = E(T_{ij}|T_{i,j-1} = t_{i,j-1}, x_{ij}) = \begin{pmatrix} \eta_{ij(1)}^c \\ \vdots \\ \eta_{ij(u)}^c \\ \vdots \\ \eta_{ij(r)}^c \end{pmatrix}, \tag{3.13}$$

where η_{ij}^c is the conditional expectation given the prior state of the process and current values of covariates. The element of η_{ij}^c is defined by

$$\eta_{ij(u)}^c = \frac{exp(x_{ij}'\beta_u + t_{i,j-1}'\gamma_{u\cdot})}{1 + \sum\limits_{l=1}^{r} exp(x_{ij}'\beta_l + t_{i,j-1}'\gamma_{l\cdot})} \tag{3.14}$$

for $j = 1, 2, \ldots, J$, where $\gamma_{u\cdot} = (\gamma_{u1}, \ldots, \gamma_{ur})'$ consists of the uth row of matrix Υ. This model leads to more complicated derivation of moments compared with model (3.3).

To address the issue of estimation, we next present the first estimating method based on model (3.3) with assumption (3.11) and (3.12). The interested parameters include all the elements of $\mathscr{B}$ and Υ.

Let $\theta = (Vec(\mathscr{B})', Vec(\Upsilon)')'$, where Vec is the operation of vectorizing a matrix. The GEE approach [Liang and Zeger, (1986)] has the estimating equations given by

$$\sum_{i=1}^{I} \frac{\partial \eta_i'}{\partial \theta} W_i^{-1}(t_i - \eta_i) = 0, \tag{3.15}$$

where W_i is the "working" covariance matrix of T_i. If the true covariance matrix Σ_i in (3.10) is used, the GEE becomes the GQL method [Sutradhar (2003)]. The derivatives involved in $\partial \eta_i / \partial \theta$ are given as

$$\begin{aligned}
\frac{\partial \eta_{ij(u)}}{\partial \beta_{mu}} &= \sum_{k=1}^{r} \left(\frac{\partial \tilde{\eta}_{ij(u,k)}}{\partial \beta_{mu}} - \frac{\partial \bar{\eta}_{ij(u)}}{\partial \beta_{mu}}\right)\eta_{i,j-1(k)} + \sum_{k=1}^{r} (\tilde{\eta}_{ij(u,k)} - \bar{\eta}_{ij(u)}) \frac{\partial \eta_{i,j-1(k)}}{\partial \beta_{mu}} + \frac{\partial \bar{\eta}_{ij(u)}}{\partial \beta_{mu}} \\
&= \sum_{k=1}^{r} [\tilde{\eta}_{ij(u,k)}(1 - \tilde{\eta}_{ij(u,k)}) - \bar{\eta}_{ij(u)}(1 - \bar{\eta}_{ij(u)})] x_{ij(m)} \eta_{i,j-1(k)} \\
&\quad + \sum_{k=1}^{r} [\tilde{\eta}_{ij(u,k)}(1 - \tilde{\eta}_{ij(u,k)})] \frac{\partial \eta_{i,j-1(k)}}{\partial \beta_{mu}} + \bar{\eta}_{ij(u)}(1 - \bar{\eta}_{ij(u)}) x_{ij(m)}, \qquad (3.16) \\
\frac{\partial \eta_{ij(u)}}{\partial \beta_{mv}} &= \sum_{k=1}^{r} [\bar{\eta}_{ij(u)} \bar{\eta}_{ij(v)} - \tilde{\eta}_{ij(u,k)} \tilde{\eta}_{ij(v,k)}] x_{ij(m)} \eta_{i,j-1(k)} \\
&\quad + \sum_{k=1}^{r} [\tilde{\eta}_{ij(u,k)}(1 - \tilde{\eta}_{ij(u,k)})] \frac{\partial \eta_{i,j-1(k)}}{\partial \beta_{mv}}, \qquad (3.17) \\
\frac{\partial \eta_{ij(u)}}{\partial \gamma_{uk}} &= \sum_{v=1}^{r} \left(\frac{\partial \tilde{\eta}_{ij(u,v)}}{\partial \gamma_{uk}} - \frac{\partial \bar{\eta}_{ij(u)}}{\partial \gamma_{uk}}\right)\eta_{i,j-1(v)} + \sum_{v=1}^{r} (\tilde{\eta}_{ij(u,v)} - \bar{\eta}_{ij(u)}) \frac{\partial \eta_{i,j-1(v)}}{\partial \gamma_{uk}} + \frac{\partial \bar{\eta}_{ij(u)}}{\partial \gamma_{uk}} \\
&= \tilde{\eta}_{ij(u,k)}(1 - \tilde{\eta}_{ij(u,k)}) \eta_{i,j-1(k)} + \sum_{v=1}^{r} [\tilde{\eta}_{ij(u,v)} - \bar{\eta}_{ij(u)}] \frac{\partial \eta_{i,j-1(v)}}{\partial \gamma_{uk}}, \qquad (3.18) \\
\frac{\partial \eta_{ij(u)}}{\partial \gamma_{mk}} &= -\tilde{\eta}_{ij(u,k)} \tilde{\eta}_{ij(u,k)} \eta_{i,j-1(k)} + \sum_{v=1}^{r} [\tilde{\eta}_{ij(u,u)} - \bar{\eta}_{ij(u)}] \frac{\partial \eta_{i,j-1(v)}}{\partial \gamma_{mk}}, \qquad (3.19)
\end{aligned}$$

where

$$
\begin{aligned}
\frac{\partial \tilde{\eta}_{ij(u,k)}}{\partial \beta_{mu}} &= \tilde{\eta}_{ij(u,k)}(1-\tilde{\eta}_{ij(u,k)})x_{ij(m)}, \\
\frac{\partial \tilde{\eta}_{it(u,k)}}{\partial \beta_{mv}} &= -\tilde{\eta}_{ij(v,k)}\tilde{\eta}_{ij(u,k)}x_{ij(m)}, \\
\frac{\partial \tilde{\eta}_{ij(u,k)}}{\partial \gamma_{uk}} &= \tilde{\eta}_{ij(u,k)}(1-\tilde{\eta}_{ij(u,k)}), \\
\frac{\partial \tilde{\eta}_{ij(u,k)}}{\partial \gamma_{vm}} &= 0 \quad m \neq k \text{ for any } v \\
\frac{\partial \tilde{\eta}_{ij(u,k)}}{\partial \gamma_{mk}} &= -\tilde{\eta}_{ij(u,k)}\tilde{\eta}_{ij(m,k)}, \\
\frac{\partial \bar{\eta}_{ij(u)}}{\partial \beta_{ml}} &= \bar{\eta}_{ij(u)}(1-\bar{\eta}_{ij(u)})x_{ij(m)}, \\
\frac{\partial \bar{\eta}_{ij(u)}}{\partial \beta_{mv}} &= -\bar{\eta}_{ij(u)}\bar{\eta}_{ij(v)}x_{ij(m)}, \\
\frac{\partial \bar{\eta}_{ij(u)}}{\partial \Upsilon} &= \mathbf{0} \quad \text{(matrix)}.
\end{aligned}
$$

The most efficient estimates of the model parameters can be obtained by applying the maximum likelihood approach. Based on model (3.3), the likelihood function given observations $T = t$ is

$$L(\theta) = \prod_{i=1}^{I}\prod_{j=1}^{J} g_{ij|j-1}, \tag{3.20}$$

where

$$
\begin{aligned}
g_{ij|j-1} = {} & \left[\prod_{v=1}^{r}\prod_{u=1}^{r} \tilde{\eta}_{ij(u,v)}^{t_{ij(u)}t_{i,j-1(v)}}\right]\left[\prod_{v=1}^{r}(1-\mathbf{1}'\tilde{\eta}_{ij(\cdot,v)})^{(1-\mathbf{1}'t_{ij})t_{i,j-1(v)}}\right] \\
& \times \left[\prod_{u=1}^{r} \bar{\eta}_{ij(u)}^{t_{ij(u)}(1-\mathbf{1}'t_{i,j-1})}\right]\left[(1-\mathbf{1}'\bar{\eta}_{ij})^{(1-\mathbf{1}'t_{ij})(1-\mathbf{1}'t_{i,j-1})}\right].
\end{aligned}
$$

Then the log-likelihood function is

$$
\begin{aligned}
\ell(\theta) = {} & \sum_{i=1}^{I}\sum_{j=1}^{J}\{\sum_{v=1}^{r}[\sum_{u=1}^{r} t_{ij(u)}t_{i,j-1v}log(\tilde{\eta}_{ij(u,v)}) + (1-\mathbf{1}'t_{ij})t_{i,j-1(v)}log(1-\mathbf{1}'\tilde{\eta}_{ij(\cdot,v)})] \\
& +(1-\mathbf{1}'t_{ij})(1-\mathbf{1}'t_{i,j-1})log(1-\mathbf{1}'\bar{\eta}_{ij}) + \sum_{u=1}^{r} t_{ij(u)}(1-\mathbf{1}'t_{i,j-1})log(\bar{\eta}_{ij(u)})\}.
\end{aligned}
$$

To maximize the function $\ell(\theta)$, we solve the following equations:

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta_u} &= \sum_{i=1}^{I}\sum_{j=1}^{J}[\sum_{v=1}^{r}(t_{ij(u)} - \tilde{\eta}_{ij(u,v)})t_{i,j-1(v)} + (t_{ij(u)} - \tilde{\eta}_{ij(u)})(1 - \mathbf{1}' t_{i,j-1})]x'_{ij} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{m}[\sum_{v=1}^{r}(t_{ij(u)} - E(T_{ij(u)}|T_{i,j-1} = t_{i,j-1})]x'_{ij} = 0, \quad (3.21) \\
\frac{\partial \ell(\theta)}{\partial \gamma_{uv}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(t_{ij(u)} - \tilde{\eta}_{ij(u,v)})t_{i,j-1(v)} = 0. \quad (3.22)
\end{aligned}$$

However, the expressions (3.13-3.14) lead to a simpler development of ML approach than model (3.3). In the likelihood function based on (3.13) and (3.14),

$$\begin{aligned}
g_{i,j|j-1} &= (1 - \mathbf{1}'\eta^c_{ij})^{1-\mathbf{1}' t_{ij}} \prod_{u=1}^{r}[\eta^c_{ij(u)}]^{t_{ij(u)}} \\
&= \frac{\prod_{u=1}^{r} exp[\sum_{u=1}^{r}(x_{ij}\beta_u + t'_{ij-1}\gamma_u)t_{iju}]}{1 + \sum_{u=1}^{r} exp(x_{ij}\beta_u + t'_{i,j-1}\gamma_u)},
\end{aligned}$$

and the log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\sum_{u=1}^{r}(x_{i1}\beta_j + t'_{i,j-1}\gamma_u)t_{iju} - log\{1 + \sum_{u=1}^{r} exp(x_{ij}\beta_u + t'_{i,j-1}\gamma_u)\}\right]. \quad (3.23)$$

This leads to a series of simpler score equations

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta_u} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(t_{ij(u)} - \eta^c_{ij(u)})x'_{ij} = 0, \quad (3.24) \\
\frac{\partial \ell(\theta)}{\partial \gamma_u} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(t_{ij(u)} - \eta^c_{ij(u)})t_{i,j-1} = 0, \quad (3.25)
\end{aligned}$$

and they are equivalent to equations (3.21) and (3.22).

It can be seen that, in the first order transition model, the pairwise products $t_{i1}t'_{i2}, \ldots, t_{ij}t'_{iu}, \ldots, t_{i,J-1}t'_{iJ}$ along with the first order responses $t_{i1}, \ldots, t_{ij}, \ldots, t_{iJ}$ provide sufficient information for the estimation of $\mathscr{B}$ and Υ. This finding implies that estimators based on the second-order GQL (GQL2) may be as efficient as the ML estimators. This suggests the promising application of GQL2 in the case that the full likelihood function is difficult to develop.

3.1.2 The transition model for dynamic binary data

Analogous to model (3.3), the dynamic binary data model can be written as

$$T_{ij} = \tilde{\eta}_{ij} * T_{i,j-1} + \bar{\eta}_{ij} * (1 - T_{i,j-1}), \tag{3.26}$$

with the baseline observations t_{i0} in a longitudinal study, where $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots, J$.

Under assumptions that $\tilde{\eta}_{ij} = \gamma_{ij} + (1 - \gamma_{ij})\xi_{ij}$ and $\bar{\eta}_{ij} = (1 - \gamma_{ij})\xi_{ij}$, for $j = 1, 2, \ldots, J$, model (3.26) becomes the thinning-operation-based linear transition model of the following linear binary dynamic model given by

$$T_{ij} = b_{ij} T_{i,j-1} + (1 - b_{ij})\epsilon_{ij} \tag{3.27}$$

[Tong (1990)]. In model (3.27), it is also assumed that $b_{ij} \sim b(1, \gamma_{ij})$, and b_{ij} is independent of ϵ_{ij}.

Let $\tilde{\eta}_{ij} = \frac{exp(x'_{ij}\beta+\gamma)}{1+exp(x'_{ij}\beta+\gamma)}$ and $\bar{\eta}_{ij} = \frac{exp(x'_{ij}\beta)}{1+exp(x'_{ij}\beta)}$, model (3.26) will be the thinning-operation-version of the non-linear binary dynamic model which is given by

$$\begin{aligned} \eta^c_{ij} &= P(T_{ij} = 1 | T_{i,j-1} = t_{i,j-1}) \\ &= \frac{exp(x'_{ij}\beta + t_{i,j-1}\gamma)}{1 + exp(x'_{ij}\beta + t_{i,j-1}, \gamma)}, \text{ for j=1,2,..., J,} \end{aligned} \tag{3.28}$$

[Amemiya (1985); Manski (1987)]. In fact, $\eta^c_{ij} = \tilde{\eta}_{ij} t_{i,j-1} + \bar{\eta}_{ij}(1 - t_{i,j-1})$, is the conditional expectation $E(T_{ij} | T_{i,j-1} = t_{i,j-1})$ derived from model (3.26). This is a special case of model (3.14) for longitudinal binary data.

The mean and variance of T_{ij} based on model (3.26) are given by

$$\eta_{ij} \triangleq E(T_{ij}) \;=\; \eta_{i,j-1}(\tilde{\eta}_{ij} - \bar{\eta}_{ij}) + \bar{\eta}_{ij}, \tag{3.29}$$

$$\sigma_{ij}^2 \triangleq Var(T_{ij}) \;=\; \eta_{ij}(1 - \eta_{ij}), \tag{3.30}$$

and the covariance between T_{ij} and T_{iu} is

$$\sigma_{iju} \triangleq Cov(T_{ij}, T_{iu}) \;=\; Var(T_{iu}) \prod_{k=u+1}^{j} (\tilde{\eta}_{ik} - \bar{\eta}_{ik}), \text{ for } u < j, \tag{3.31}$$

where $\eta_{i0} = t_{i0}$.

Recently, the non-linear dynamic model (3.28) was applied by Sutradhar and Farrell (2007) to analyze a longitudinal children asthma data set. They developed three approaches to estimate the model parameters β and γ, namely, the generalized quasi-likelihood (GQL), the second order GQL (GQL2) and maximum likelihood (ML) approaches. In the GQL2 approach, they combined the first and second order statistics of the responses into the estimating procedures and obtained highly efficient estimates which are comparable to ML estimates. They also showed that the GQL2 approach is the optimal GQL (OGQL) method in the lag 1 dynamic dependence model (3.27). The detailed estimating procedures under the three approaches are given below.

1. GQL approach

The parameters $\theta = (\beta', \gamma)'$ are estimated by solving the estimating equations [Sutradhar (2003)]

$$\sum_{i=1}^{I} \frac{\partial \eta_i'}{\partial \theta} \Sigma_i^{-1}(t_i - \eta_i) = 0. \tag{3.32}$$

Once we have $\hat{\theta}_{GQL}$, its covariance matrix can be estimated by

$$\hat{V}(\hat{\theta}_{GQL}) = \left(\sum_{i=1}^{I} \frac{\partial \eta_i'}{\partial \theta} \Sigma_i^{-1} \frac{\partial \eta_i}{\partial \theta} \right)^{-1} |_{\theta = \hat{\theta}_{GQL}}. \tag{3.33}$$

2. ML approach

The likelihood function of the observations $t = \{t_{ij}, i = 1, 2, \ldots, I, \text{ and } j = 1, 2, \ldots, J.\}$ is given by

$$L(\theta) = \prod_{i=1}^{I} \prod_{j=1}^{J} g_{i,j|j-1}, \tag{3.34}$$

where

$$g_{i,j|j-1} = \bar{\eta}_{ij}^{t_{ij}t_{i,j-1}} (1 - \bar{\eta}_{it})^{(1-\mathbf{1}'t_{ij})t_{i,j-1}} \bar{\eta}_{ij}^{t_{ij}(1-t_{i,j-1})} (1 - \bar{\eta}_{ij})^{(1-t_{ij})(1-t_{i,j-1})}.$$

Then the log-likelihood function is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{I} \sum_{j=1}^{J} \{t_{ij}t_{i,j-1} log(\bar{\eta}_{ij}) + (1 - t_{ij})t_{i,j-1} log(1 - \bar{\eta}_{it}) \\ &\quad + t_{ij}(1 - t_{i,j-1}) log(\bar{\eta}_{ij}) + (1 - t_{ij})(1 - t_{i,j-1}) log(1 - \bar{\eta}_{ij})\}. \end{aligned}$$

To maximize the function $\ell(\theta)$, we solve the following equations:

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \beta_u} &= \sum_{i=1}^{I} \sum_{j=1}^{J} [(t_{ij} - \bar{\eta}_{ij})t_{i,j-1} + (t_{ij} - \bar{\eta}_{ij})(1 - t_{i,j-1})] x'_{ij(u)} \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} [(t_{ij} - \eta_{ij}^{c}] x'_{ij(u)} = 0, \quad \text{for } u = 1, 2, \ldots, p, \qquad (3.35) \\ \frac{\partial \ell(\theta)}{\partial \gamma} &= \sum_{i=1}^{I} \sum_{j=1}^{J} (t_{ij} - \bar{\eta}_{ij}) t'_{i,j-1} = 0. \qquad (3.36) \end{aligned}$$

If one wants to derive the ML method based on model (3.28), the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{I} \prod_{j=1}^{J} g_{i,j|j-1},$$

where $g_{i,j|j-1} = \left(\eta_{ij}^{c}\right)^{t_{ij}} \left(1 - \eta_{ij}^{c}\right)^{1-t_{ij}}$, and the log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{I} \sum_{j=1}^{J} t_{ij}(x'_{ij}\beta + \gamma t_{i,j-1}) - \sum_{i=1}^{I} \sum_{j=1}^{J} log[1 + exp(x'_{ij}\beta + \gamma t_{i,j-1})].$$

This yields equivalent score equations to (3.35) and (3.36).

The covariance matrix of $\hat{\theta}_{ML}$ can be estimated by

$$\hat{V}(\hat{\theta}_{ML}) = \left(I_T^{-1}\right)_{\theta=\hat{\theta}_{ML}}, \tag{3.37}$$

where

$$I_T = \left[-E\{\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\}\right]$$

is the Fisher Information matrix.

3. GQL2 approach

To utilize more information of data, Sutradhar and Farrell (2007) suggested that it would be more efficient to combine the second order statistics of the responses into the estimating procedure. The estimating equations [Sutradhar and Farrell (2007)] are given by

$$\sum_{i=1}^{I} \frac{\partial \varphi_i'}{\partial \theta} \Omega_i^{-1}(h_i - \phi_i) = 0, \tag{3.38}$$

where h_i is the observation of $H_i = (T_i', R_i')'$, and $\varphi_i = E(H_i)$. Here $T_i = (T_{i1}, \ldots, T_{iJ})'$, and $R_i = (T_{i1}T_{i2}, \ldots, T_{iu}T_{iv}, \ldots, T_{i,J-1}T_{iJ})'$. Ω_i is the covariance matrix of H_i, that is,

$$\Omega_i = Cov(H_i) = \begin{pmatrix} Cov(T_i) & Cov(T_i, R_i) \\ Cov(R_i, T_i) & Cov(R_i) \end{pmatrix}_{J(J+1)/2 \times J(J+1)/2}. \tag{3.39}$$

The GQL2 approach involves some moments up to order four, say

$$\begin{aligned} E(T_{ij}T_{iu}) &= \sigma_{iuj} + \eta_{iu}\eta_{ij}, \text{for } u \neq j, \\ E(T_{ij}T_{iu}T_{iv}) &= \sum_{S_1^*}(\prod_{j=1}^{J} g_{i,j|j-1})_{t_{iu}=1,t_{iv}=1,t_{ij}=1}, \text{for different } j, u, v, \\ E(T_{ij}T_{iu}T_{iv}T_{il}) &= \sum_{S_2^*}(\prod_{j=1}^{J} g_{i,j|j-1})_{t_{il}=1,t_{iu}=1,t_{iv}=1,t_{ij}=1}, \text{for different } j, u, v, l, \end{aligned}$$

where $g_{i,j|j-1} = (\eta_{i,j}^c)^{t_{ij}}(1-\eta_{ij}^c)^{1-t_{ij}}$ for $j = 1, 2, \ldots, J$, and $\sigma_{iuj} = Cov(T_{ij}, T_{iu})$ $\sum_{S_1^*}$ indicates the summation over all $t_{ik} = 0$ for $k \neq u, v, j$, and similarly $\sum_{S_2^*}$ indicates the summation over all $t_{ik} = 0$ for $k \neq l, u, v, j$.

The covariance matrix of $\hat{\theta}_{GQL2}$ is estimated by

$$\hat{V}(\hat{\theta}_{GQL2}) = \left(\sum_{i=1}^{I} \frac{\partial \phi_i'}{\partial \theta} \Omega_i^{-1} \frac{\partial \phi_i}{\partial \theta}\right)^{-1} |_{\theta=\hat{\theta}_{GQL2}}. \tag{3.40}$$

As far as the matrix Ω_i (3.39) is concerned, some authors used a "working" covariance matrix, for example, normality based covariance [Zhao and Prentice, (1990)], independence covariance [Sutradhar, (2003)]. But these working covariance matrix may results in loss of efficiency.

To conduct statistical inference, for example, constructing confidence interval or testing hypothesis, one needs the asymptotic distributions of these estimators $\hat{\theta}_{GQL}$, $\hat{\theta}_{GQL2}$, and $\hat{\theta}_{ML}$. Under some mild regularity conditions [Newey and McFadden (1993)], it follows from equations (3.32) and (3.38), (3.35-3.36) that as $I \to \infty$,

$$\sqrt{I}(\hat{\theta}_{GQL} - \theta) \sim N\left(0, \left\{\sum_{i=1}^{I} \frac{\partial \eta_i'}{\partial \theta} \Sigma_i^{-1} \frac{\partial \eta_i}{\partial \theta}\right\}^{-1}\right), \tag{3.41}$$

$$\sqrt{I}(\hat{\theta}_{GQL2} - \theta) \sim N\left(0, \left\{\sum_{i=1}^{I} \frac{\partial \varphi_i'}{\partial \theta} \Omega_i^{-1} \frac{\partial \varphi_i}{\partial \theta}\right\}^{-1}\right), \tag{3.42}$$

and

$$\sqrt{I}(\hat{\theta}_{ML} - \theta) \sim N\left(0, I_T^{-1}\right). \tag{3.43}$$

3.2 Longitudinal Models for Count Data

As mentioned in Chapter 1, a lot of authors discussed nonlinear transition models [Besag (1974); Wong (1986); Zeger and Qaqish (1988); Diggle et al. (2002)]. In practice, some count data may have linear relationship among the responses T_{ij}'s, for example, the dynamic population sizes in a district. Among the population of size $T_{i,j-1}$ in the prior year, some people may die or move out in the $(j-1)$th year, and the rest are still living in this area. At the same time, there may be some newborns or immigrants in this area.

Another example is the prevalence count of an epidemic disease which is defined as the total number of the disease cases among the population at a given year. Let $T_{i,j-1}$ denote the prevalence count in the $(j-1)$th year and T_{ij} denote the prevalence count in the jth year. The overlap of $T_{i,j-1}$ and T_{ij} consists of those patients who survived over the $(j-1)$th year and are still suffering from the disease in the jth year. In fact, the new cancer cases or new mortality cases due to cancers are not exponentially or multiplicatively growing over time. The expected count in the next year may be approximately linearly correlated with the average of the count of disease cases in the previous year. Therefore, the linear transition model is very promising in epidemiologic studies.

In the following subsections, we introduce two new models for dynamic count data which characterize the linear dependence among follow-up observations.

3.2.1 Non-stationary AR(1) model

Let T_{ij} denote the count response in district i in the jth year, $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. McKenzie (1988) and Sutradhar (2003) discussed a stationary AR(1) model for count data. The model is given by

$$T_{ij} = \gamma * T_{i,j-1} + \epsilon_{ij}, \tag{3.44}$$

where $*$ is the binomial thinning operation, and t_{i0}'s denote the baseline observations. It was assumed that $T_{i,j-1} \sim Poisson(\eta_{i\cdot} = exp(x_{i\cdot}'\beta))$, and $\epsilon_{ij} \sim Poisson((1-\gamma)\eta_{i\cdot})$. This model can only describe the dynamic count data with time-constant covariates, which limits its application in practice. Sutradhar et al. (2008) modified the model (3.44) into a non-stationary AR(1) model under different assumptions. They assumed that $y_{i1} \sim Poisson(\mu_{i1})$ with $\mu_{ij} = exp(x_{ij}'\beta)$, $\epsilon_{ij} \sim Poisson(\mu_{ij} - \gamma\mu_{i,j-1})$, and $y_{i,j-1}$ is independent of ϵ_{ij}, $j = 2, \ldots, J$. Under the non-stationary model, the expectations are time-varying, that is, $\mu_{ij} = exp(x_{ij}'\beta)$. The new model can be used to describe the longitudinal count data with time-dependent covariates. However, the restriction of the expectation of the additive error ϵ_{ij} may not be suitable in some practical cases.

In this thesis, we generalize model (3.44) to a new non-stationary AR(1) (NS-AR(1)) model which allows for time-varying covariates and unrestricted expectation of the additive error. The new model is given by

$$T_{ij} = \gamma * T_{i,j-1} + D_{ij}. \tag{3.45}$$

This model has a similar expression as (3.44) but different assumptions. It is assumed that $D_{ij} \sim Poisson(\xi_{ij})$, where ξ_{ij} may be a function of $t_{i,j-1}$ and some explanatory variables. However, given $T_{i,j-1} = t_{ij-1}$, D_{ij} is independent of $\gamma * t_{i,j-1}$. It is obvious that, given the prior observation $T_{i,j-1} = t_{i,j-1}$, the conditional expectation of T_{ij} can be formulated by

$$\eta_{ij}^c = E(T_{ij}|T_{i,j-1} = t_{i,j-1}) = \xi_{ij} + \gamma t_{i,j-1}. \tag{3.46}$$

In practice, the new model (3.45) can be exploited to model different types of data sets with various background, then the assumptions about ξ_{ij} may vary accordingly. Two examples are presented in the following paragraphs.

1. For the dynamic population, T_{ij} denotes the population size at the jth year in district i. In model (3.45), the first term $\gamma * T_{i,j-1}$ consists of people who are living in district i from the $(j-1)$th year to the jth year, whereas D_{ij} represents new residents due to birth and immigration. $T_{i,j-1} - \gamma * T_{i,j-1}$ is the number of people who died or emigrated during the $(j-1)$th year. In this situation, the number of people included in D_{ij} due to newborns may be related to the previous population size $T_{i,j-1}$. Therefore ξ_{ij} can be a function of the previous population size t_{ij-1} and covariates x_{ij}, for example, $\xi_{ij} = exp(x'_{ij}\beta + \rho t_{i,j-1})$.

2. For the data about prevalence of a disease, T_{ij} denotes the prevalence count of the disease cases at the jth year in district i. The first part in model (3.45) $\gamma * T_{i,j-1}$ is composed of cases who survived over the $(j-1)$th year and are still suffering the disease in the jth year. D_{ij} includes new disease cases including people who became infected in the jth year and patients immigrating from other districts. $T_{i,j-1} - \gamma * T_{i,j-1}$ includes patients who died or move out of district i in the jth year before the survey. D_{ij} can be reasonably assumed to be independent of the prior observation $t_{i,j-1}$.

Alternatively, the term $\gamma * T_{i,j-1}$ in the two examples can also be interpreted as the dynamic dependence of T_{ij} on $T_{i,j-1}$. Therefore, in the first example, if one assumes that $\gamma * T_{i,j-1}$ can completely characterize the dynamic dependence, the D_{ij} can be supposed to be independent of $T_{i,j-1}$, which leads to simple development of the inference of the model. Therefore, in this thesis, we assume that these D_{ij}'s are mutually independent and they are independent of the previous observations $T_{i,j-1}$'s. As far as ξ_{ij} is concerned, it can be assumed that $\xi_{ij} = exp(x'_{ij}\beta)$.

For the NS-AR(1) model (3.45), it can be concluded from Section 2.4 in Chapter 3 that a Poisson variable $T_{i,j-1}$ leads to a Poisson-distributed $\gamma * T_{i,j-1}$, then it further leads to a Poisson variable T_{ij}. There are three different cases of model (3.44).

Case 1. The zero baseline observation $t_{i0} = 0$ leads to a unconditionally Poisson variable T_{i1}, hence leads to a Poisson sequence $\{T_{ij} \sim Poisson(\eta_{ij})\}$. This can be used to model new cases of a disease in different periods from the outbreak of the disease in small areas. However, in a large district, it is difficult to go back to the outbreak of some disease, for example, cancer.

Case 2. Suppose that we have a non-zero baseline t_{i0}. It leads to a Binomial variable $\gamma * T_{i0}$. Therefore we get a non-Poisson variable T_{i1} and hence a sequence of non-Poisson variables $\{T_{ij}\}$.

Case 3. If the baseline observations t_{i0}'s are not available, we can assume that T_{i1} follows a Poisson distribution with expectation ξ_{i1} with a regression intercept β_0 describing the baseline expectation. This assumption results in a sequence of Poisson variables $\{T_{ij}\}$.

In the second case with non-zero baseline observations, the likelihood function of the data becomes very complicated, which will be shown in the following subsection.

3.2.2 Linear transition model

In some cases, T_{ij} and $T_{i,j-1}$ may not follow the NS-AR(1) model (3.45), but the conditional mean structure (3.46) describing the relationship between these two may still hold. For example, to model the incidence count of an epidemic disease in district

i, the count of the new disease cases in the jth year T_{ij} may not follow the NS-AR(1) model (3.45). Even the dynamic population data may not exactly follow the NS-AR(1) model due to the mixture of migration with death and birth. Taking this into consideration, the linear transition (LT) model based on the conditional mean structure (3.46) becomes an appropriate alternative of the NS-AR(1) model (3.45).

The LT model is given by

$$T_{ij}|T_{i,j-1} = t_{i,j-1} \sim Poisson(\eta^c_{i,j} = \xi_{ij} + \gamma t_{i,j-1}). \tag{3.47}$$

This model can be viewed as a special case of the linear feedback model (LFM) by Blundell, Griffith and Windmeijer (2002) in which there is no subject-specific random effect. However, our model has different interpretations from the LFM. For the incidence count of a disease, under model (3.47), t_{ij} denotes the incidence count of the disease at the jth time in district i. The term $\gamma t_{i,j-1}$ reflects the dynamic dependence of T_{ij} on $T_{i,j-1}$, and ξ_{ij} can be assumed to be a function of covariates such as environmental factors which contribute to the expectation of the occurrences of the disease.

When the LT model is applied to data sets in the two examples in the previous subsection, the terms $\gamma t_{i,j-1}$, ξ_{ij} and $(1-\gamma)t_{i,j-1}$ have similar interpretations to those under the NS-AR(1) model presented in Section 3.2.1. However, the term $\gamma t_{i,j-1}$ can also completely accommodate the dynamic dependence of T_{ij} on the prior observation $t_{i,j-1}$.

It is apparent that the conditional distribution of T_{ij} given the prior observation $t_{i,j-1}$ is a Poisson distribution. However, except T_{i1} with baseline $t_{i0} = 0$, the marginal distribution of T_{ij} is not Poisson, for $j = 1, 2, \ldots, J$.

3.2.3 Moments of the NS-AR(1) and LT models

In this subsection, we provide the calculations of some moments of the responses based on the NS-AR(1) model (3.45) and the LT model (3.47). These moments are used in the construction of estimating equations for the GQL and GQL2 approaches. Some higher order moments, for example, the third and fourth order moments, are also required for the GQL2 estimating approach, we provide the relevant calculations at the end of this section. One may notice that some moments under both NS-AR(1) model and LT model share the same expressions.

Under the NS-AR(1) model and LT model, we have the same expression of the expectation of T_{ij} which is given by

$$\eta_{ij} = \xi_{ij} + \gamma\eta_{i,j-1} \Rightarrow \eta_{ij} = \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \gamma^{j-u}\eta_{iu}, \text{for } u < j, \tag{3.48}$$

where $\eta_{i0} = t_{i0}$. We also have the same expression of the expectation of the pairwise product $T_{ij}T_{iu}$ as follows:

$$\begin{aligned} E(T_{ij}T_{iu}) &= \eta_{iu}\xi_{ij} + \gamma E(T_{iu}T_{i,j-1}) = \eta_{iu}\sum_{k=0}^{j-u-1}\gamma^k\xi_{i,j-k} + \gamma^{j-u}E(T_{iu}^2) \\ &\Rightarrow \; Cov(T_{iu}, T_{ij}) = \gamma^{j-u}\sigma_{iu}^2 \\ &\Rightarrow \; Corr(T_{iu}, T_{ij}) = min\{1, \gamma^{j-u}\frac{\sigma_{ij}}{\sigma_{iu}}\}. \end{aligned}$$

The expectations of T_{ij}^2 under NS-AR(1) and LT models are, respectively, given by

$$\text{NS-AR(1) model:} \qquad E(T_{ij}^2) = \eta_{ij} + \eta_{ij}^2 + \gamma^2[E(T_{i,j-1}^2) - \eta_{i,j-1} - \eta_{i,j-1}^2]$$

$$\Rightarrow \quad \sigma_{ij}^2 = \eta_{ij} + \gamma^2(\sigma_{i,j-1}^2 - \eta_{i,j-1}),$$

$$\text{LT model:} \qquad E(T_{ij}^2) = \eta_{ij} + \eta_{ij}^2 + \gamma^2[E(T_{i,j-1}^2) - \eta_{i,j-1}^2]$$

$$\Rightarrow \quad \sigma_{ij}^2 = \eta_{ij} + \gamma^2\sigma_{i,j-1}^2.$$

Under both NS-AR(1) model and LT model, some third and forth order moments with the same expressions are given by

$$\begin{aligned}
E(T_{iu}^2 T_{ij}) &= E(T_{iu}^2)\xi_{ij} + \gamma E(T_{iu}^2 T_{i,j-1}), \quad u < j \\
&= E(T_{iu}^2) \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \rho^{j-u} E(T_{iu}^3), \\
E(T_{iu}^3 T_{ij}) &= E(T_{iu}^3)\xi_{ij} + \gamma E(T_{iu}^3 T_{i,j-1}), \quad u < j \\
&= E(T_{iu}^3) \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \rho^{j-u} E(T_{iu}^4), \\
E(T_{iv} T_{iu} T_{ij}) &= E(T_{iv} T_{iu})\xi_{ij} + \gamma E(T_{iv} T_{iu} T_{i,j-1}), \quad v < u < j \\
&= E(T_{iv} T_{iu}) \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \rho^{j-u} E(T_{iv} T_{iu}^2), \\
E(T_{iv}^2 T_{iu} T_{ij}) &= E(T_{iv}^2 T_{iu})\xi_{ij} + \gamma E(T_{iv}^2 T_{iu} T_{i,j-1}), \quad v < u < j \\
&= E(T_{iv}^2 T_{iu}) \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \rho^{j-u} E(T_{iv}^2 T_{iu}^2), \\
E(T_{iv} T_{iu}^2 T_{ij}) &= E(T_{iv} T_{iu}^2)\xi_{ij} + \gamma E(T_{iv} T_{iu}^2 T_{i,j-1}), \quad v < u < j \\
&= E(T_{iv} T_{iu}^2) \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \rho^{j-u} E(T_{iv} T_{iu}^3), \\
E(T_{il} T_{iv} T_{iu} T_{ij}) &= E(T_{il} T_{iv} T_{iu})\xi_{ij} + \gamma E(T_{il} T_{iv} T_{iu} T_{i,j-1}), \quad l < v < u < j \\
&= E(T_{il} T_{iv} T_{iu}) \sum_{k=u+1}^{j} \xi_{ik}\gamma^{j-k} + \rho^{j-u} E(T_{il} T_{iv} T_{iu}^2).
\end{aligned}$$

Under the NS-AR(1) model, the third and fourth order moments can also be calculated as follows:

$$
\begin{aligned}
E(T_{ij}^3) &= \eta_{ij} + 3\eta_{ij}^2 + \eta_{it}^3 + \gamma^3[E(T_{i,j-1}^3) - \eta_{ij-1} - 3\eta_{ij-1}^2 - \eta_{ij}^3] \\
&\quad +3\gamma^2(1+\xi_{ij}-\gamma)[E(T_{i,j-1}^2) - \eta_{i,j-1} - \eta_{i,j-1}^2], \\
E(T_{ij}^4) &= \eta_{ij} + 7\eta_{ij}^2 + 6\eta_{ij}^3 + \eta_{ij}^4 + \gamma^4[E(T_{i,j-1}^4) - \eta_{i,j-1} - 7\eta_{i,j-1}^2 - 6\eta_{i,j-1}^3 - \eta_{i,j-1}^4] \\
&\quad +\gamma^3(6+4\xi_{ij}-6\gamma)[E(T_{i,j-1}^3) - \eta_{i,j-1} - 3\eta_{i,j-1}^2 - \eta_{i,j-1}^3] \\
&\quad +\gamma^2[11\gamma^2 - 3\gamma(6+4\xi_{ij}) + 7 + 18\xi_{ij} + 6\xi_{ij}^2][E(T_{i,j-1}^2) - \eta_{i,j-1} - \eta_{i,j-1}^2], \\
E(T_{iu}T_{ij}^2) &= \underbrace{(\xi_{ij}+\xi_{ij}^2)\eta_{iu} + \gamma(1-\gamma+2\xi_{ij})E(T_{iu}T_{i,j-1})}_{A11_{uj}} + \gamma^2 E(T_{iu}T_{i,j-1}^2), \quad u<j \\
&= \sum_{k=u+1}^{j} A11_{uk}\gamma^{2(j-k)} + \gamma^{2(j-u)}E(T_{iu}^3) \\
E(T_{iu}^2T_{ij}^2) &= \underbrace{(\xi_{ij}+\xi_{ij}^2)E(T_{iu}^2) + \gamma(1-\gamma+2\xi_{ij})E(T_{iu}^2T_{i,j-1})}_{A12_{uj}} + \gamma^2 E(Y_{iu}^2Y_{i,j-1}^2), \quad u<j \\
&= \sum_{k=u+1}^{j} A12_{uk}\gamma^{2(j-k)} + \gamma^{2(j-u)}E(T_{iu}^4), \\
E(T_{iu}T_{ij}^3) &= \underbrace{\eta_{iu}(\xi_{ij}+3\xi_{ij}^2+\xi_{ij}^3) + \gamma(1-3\gamma+2\gamma^2+6\xi_{ij}+3\xi_{ij}^2-3\xi_{ij}\gamma)E(T_{iu}T_{i,j-1})} \\
&\quad \underbrace{+3\gamma^2(1-\gamma+\xi_{ij})E(T_{iu}T_{i,j-1}^2)} + \gamma^3 E(T_{iu}T_{i,j-1}^3), \quad u<j \\
&= \sum_{k=u+1}^{j} A13_{uk}\gamma^{3(j-k)} + \gamma^{3(j-u)}E(T_{iu}^4), \text{the underbraced part} = A13_{uj}, \\
E(T_{iv}T_{iu}T_{ij}^2) &= \underbrace{(\xi_{ij}+\xi_{ij}^2)E(T_{iv}T_{iu}) + \gamma(1-\gamma+2\xi_{ij})E(T_{iv}T_{iu}T_{i,j-1})}_{A14_{uj}} \\
&\quad +\gamma^2 E(T_{iv}T_{iu}T_{i,j-1}^2), \quad v<u<j \\
&= \sum_{k=u+1}^{j} A14_{uk}\gamma^{2(j-k)} + \gamma^{2(j-u)}E(T_{iv}T_{iu}^3),
\end{aligned}
$$

whereas under the LT model these moments can are

$$
\begin{aligned}
E(T_{ij}^3) &= E[\eta_{ij}^c + 3(\eta_{ij}^c)^3 + (\eta_{ij}^c)^3] \\
&= \eta_{ij} + 3\eta_{ij}^2 + \eta_{iu}^3 + 3\gamma^2(1+\xi_{ij})[E(T_{i,j-1}^2] - \eta_{ij-1}^2] + \gamma^3[E(T_{i,j-1}^3) - \eta_{i,j-1}^3], \\
E(T_{ij}^4) &= E[\eta_{ij}^c + 7(\eta_{ij}^c)^2 + 6(\eta_{ij}^c)^3 + (\eta_{ij}^c)^4] \\
&= \eta_{ij} + 7\eta_{ij}^2 + 6\eta_{iu}^3 + \eta_{ij}^4 + \gamma^2(7 + 18\xi + 6\xi_{ij}^2)[E(T_{i,j-1}^2) - \eta_{i,j-1}^2] \\
&\quad + \gamma^3(1+4\xi_{ij})[E(T_{i,j-1}^3] - \eta_{i,j-1}^3] + \gamma^4[E(T_{i,j-1}^4] - \eta_{i,j-1}^4], \\
E(T_{iu}T_{ij}^2) &= \underbrace{(\xi_{ij} + \xi_{ij}^2)\eta_{iu} + \gamma(1+2\xi_{ij})E(T_{iu}T_{i,j-1})}_{A21_{uj}} + \gamma^2 E(T_{iu}T_{i,j-1}^2), \quad u < j \\
&= \sum_{k=u+1}^{j} A21_{uk}\gamma^{2(j-k)} + \gamma^{2(j-u)}E(T_{iu}^3), \\
E(T_{iu}^2 T_{ij}^2) &= \underbrace{(\xi_{ij} + \xi_{ij}^2)E(T_{iu}^2) + \gamma(1+2\xi_{ij})E(T_{iu}^2 T_{i,j-1})}_{A22_{uj}} + \gamma^2 E(T_{iu}^2 T_{i,j-1}^2), \quad u < j \\
&= \sum_{k=u+1}^{j} A22_{uk}\gamma^{2(j-k)} + \gamma^{2(j-u)}E(T_{iu}^4), \\
E(T_{iu}T_{ij}^3) &= \underbrace{(\xi_{ij} + 3\xi_{ij}^2 + \xi_{it}^3)\eta_{iu} + \gamma(1 + 6\xi_{it} + 3\xi_{ij}^2)E(T_{iu}T_{i,j-1})} \\
&\quad + \underbrace{3\gamma^2(1+\xi_{ij})E(T_{iu}T_{i,j-1}^2)} + \gamma^3 E(T_{iu}T_{i,j-1}^3), \quad u < j \\
&= \sum_{k=u+1}^{j} A23_{uk}\gamma^{3(j-k)} + \gamma^{3(j-u)}E(T_{iu}^4), \text{the underbraced part} = A23_{uj}, \\
E(T_{iv}T_{iu}T_{ij}^2) &= \underbrace{(\xi_{ij} + \xi_{ij}^2)E(T_{iv}T_{iu}) + \gamma(1+2\xi_{ij})E(T_{iv}T_{iu}T_{i,j-1})}_{A24_{uj}} \\
&\quad + \gamma^2 E(T_{iv}T_{iu}T_{i,j-1}^2), \quad v < u < j \\
&= \sum_{k=u+1}^{j} A24_{uk}\gamma^{2(j-k)} + \gamma^{2(j-u)}E(T_{iv}T_{iu}^3).
\end{aligned}
$$

3.2.4 Estimation of the model parameters

The parameters of interest in these two models are $\theta = (\beta', \gamma')'$, where β represents the effects of covariates, and γ is the dynamic dependence parameter reflecting the correlation between the current outcome and the prior outcome.

3.2.4.1 Generalized quasi-likelihood method

The generalized quasi-likelihood method (GQL) [Sutradhar, (2003)] is to obtain the estimates by solving the estimating equations:

$$\sum_{i=1}^{I} \frac{\partial \eta_i'}{\partial \theta} \Sigma_i^{-1}(t_i - \eta_i) = 0, \tag{3.49}$$

where $\partial \eta_i / \partial \theta$ is the first derivative matrix of η_i with respect to θ and is of dimension $(p+1) \times J$. p is the dimension of x_{ij} in ξ_{ij}. Among $\partial \eta_i / \partial \theta$,

$$\frac{\partial \eta_{ij}}{\partial \beta_k} = exp(x_{ij}'\beta)x_{ij(k)} + \gamma \frac{\partial \eta_{i,j-1}}{\partial \beta_k} \tag{3.50}$$

for $k = 1, \ldots, p$, $j = 1, \ldots, J$, and

$$\frac{\partial \eta_{ij}}{\partial \gamma} = \eta_{i,j-1} + \gamma \frac{\partial \eta_{i,j-1}}{\partial \gamma} \tag{3.51}$$

for $j = 1, \ldots, J$. $\Sigma_i = \Lambda_i W_i \Lambda_i$ is the variance-covariance matrix of T_i, where $\Lambda_i = diag(\sigma_{i1}, \ldots, \sigma_{iJ})$ and W_i is the true correlation structure of T_i. If W_i is replaced by a general "working" correlation structure W_i, the GQL approach becomes the GEE approach [Liang and Zeger (1986)]. Once we have the estimate $\hat{\theta}_{GQL}$, its corresponding covariance matrix can be estimated by

$$\hat{V}(\hat{\theta}_{GQL}) = \left(\sum_{i=1}^{I} \frac{\partial \eta_i'}{\partial \theta} \Sigma_i^{-1} \frac{\partial \eta_i}{\partial \theta} \right)^{-1} |_{\theta = \hat{\theta}_{GQL}}. \tag{3.52}$$

3.2.4.2 GQL2 approach

To obtain more efficient estimators of the model parameters, Sutradhar and Farrell (2007) developed the GQL2 approach which utilizes both the first and second order statistics of responses under the model (3.28) for dynamic binary data. In this section, we develop the GQL2 approach by exploiting the first and second order dynamic count responses. The GQL2 estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \varphi_i'}{\partial \theta} \Omega_i^{-1} (h_i - \varphi_i) = 0 \tag{3.53}$$

[Sutradhar and Farrell, (2007)], where $H_i = (T_i', R_i')'$ with $T_i = (T_{i1}, \ldots, T_{iJ})'$ and $R_i = (T_{i1}^2, \ldots, T_{iJ}^2, T_{i1}T_{i2}, \ldots, T_{iu}T_{iv}, \ldots, T_{i,J-1}T_{iJ})'$. h_i is the observation of H_i, and $\varphi_i = E(H_i)$ is the expectation of H_i. The variance-covariance matrix of H_i is given by

$$\Omega_i = \begin{pmatrix} Cov(T_i) & Cov(T_i, R_i) \\ Cov(R_i, T_i) & Cov(R_i) \end{pmatrix}_{J(J+3)/2 \times J(J+3)/2}.$$

Some quantities required in this approach such as φ_i, $Cov(T_i)$, $Cov(R_i, T_i)$ and $Cov(R_i)$ can be easily calculated based on the moments calculated in Section 3.2.3.

Some useful derivatives required in (3.53) are given below.

Under the NS-AR(1) model,

$$\begin{aligned}
\frac{\partial E(T_{ij}^2)}{\partial \beta} &= \frac{\partial \eta_{ij}}{\partial \beta}(1 + 2\eta_{ij}) + \gamma^2 [\frac{\partial E(T_{i,j-1}^2)}{\partial \beta} - \frac{\partial \eta_{i,j-1}}{\partial \beta}(1 + 2\eta_{i,j-1})], \\
\frac{\partial E(T_{ij}^2)}{\partial \gamma} &= \frac{\partial \eta_{ij}}{\partial \gamma}(1 + 2\eta_{ij}) + \gamma^2 [\frac{\partial E(T_{i,j-1}^2)}{\partial \gamma} - \frac{\partial \eta_{i,j-1}}{\partial \gamma}(1 + 2\eta_{i,j-1})] \\
&\quad + 2\gamma [E(T_{i,j-1}^2) - \eta_{i,j-1} - \eta_{i,j-1}^2].
\end{aligned}$$

Under the LT model,

$$\begin{aligned}
\frac{\partial E(T_{ij}^2)}{\partial \beta} &= \frac{\partial \eta_{ij}}{\partial \beta}(1 + 2\eta_{ij}) + \gamma^2 (\frac{\partial E(T_{i,j-1}^2)}{\partial \beta} - 2\eta_{i,j-1}\frac{\partial \eta_{i,j-1}}{\partial \beta}), \\
\frac{\partial E(T_{ij}^2)}{\partial \gamma} &= \frac{\partial \eta_{ij}}{\partial \gamma}(1 + 2\eta_{ij}) + \gamma^2 (\frac{\partial E(T_{i,j-1}^2)}{\partial \gamma} - 2\eta_{i,j-1}\frac{\partial \eta_{i,j-1}}{\partial \gamma}) + 2\gamma [E(T_{i,j-1}^2) - \eta_{i,j-1}^2].
\end{aligned}$$

Under both of the NS-AR(1) and LT models,

$$\begin{aligned}
\frac{\partial E(T_{iu}T_{ij})}{\partial\beta} &= \frac{\partial\eta_{iu}}{\partial\beta}\xi_{ij} + \eta_{iu}\xi_{ij}x_{ij} + \gamma\frac{\partial E(T_{iu}T_{i,j-1})}{\partial\beta} \\
&= \frac{\partial\eta_{iu}}{\partial\beta}\sum_{j=u+1}^{t}\xi_{ij}\gamma^{t-j} + \eta_{iu}\sum_{j=u+1}^{t}\xi_{ij}x_{ij}\gamma^{t-j} + \gamma^{t-u}\frac{\partial E(T_{iu}^2)}{\partial\beta}, \\
\frac{\partial E(T_{iu}T_{ij})}{\partial\gamma} &= \frac{\partial\eta_{iu}}{\partial\gamma}\xi_{ij} + E(T_{iu}T_{i,j-1}) + \gamma\frac{\partial E(T_{iu}T_{ij-1})}{\partial\gamma} \\
&= \frac{\partial\eta_{iu}}{\partial\gamma}\sum_{j=u+1}^{t}\xi_{ij}\gamma^{t-j} + \eta_{iu}\sum_{j=u+1}^{t}\gamma^{t-j-1}\xi_{ij}\gamma^{t-j} \\
&\quad + \gamma^{t-u}\frac{\partial E(T_{iu}^2)}{\partial\gamma} + (t-u)\gamma^{t-u-1}E(T_{iu}^2).
\end{aligned}$$

Once the estimate $\hat{\theta}_{GQL2}$ is obtained, the corresponding covariance matrix can be estimated by

$$\hat{V}(\hat{\theta}_{GQL2}) = \left(\sum_{i=1}^{n}\frac{\partial\varphi_i'}{\partial\theta}\Omega_i^{-1}\frac{\partial\varphi_i}{\partial\theta}\right)^{-1}\Big|_{\theta=\hat{\theta}_{GQL2}}. \tag{3.54}$$

As mentioned in the Section 3.1.2, it is shown by Sutradhar and Farrell (2007) that the GQL2 method provides as efficient estimators as the ML approaches under the first order transition model (3.28) for dynamic binary data, which is the reason that GQL2 is named OGQL. Similarly, the GQL2 method in the first order transition model (3.47) is also expected to produce almost as efficient as the MLE's for model parameters, which can be demonstrated in our numeric studies. Therefore, we do not conduct simulations under the GQL2 method under the LT model. We expect that the three methods yield estimates with very similar efficiency.

3.2.4.3 Maximized likelihood method

In this subsection, we develop parameter estimation based on the likelihood approach.

Under NS-AR(1) model, the conditional probability that $T_{ij} = t_{ij}$ given the prior observation $t_{i,j-1}$ is

$$\begin{aligned}
&P(T_{ij} = t_{ij}|T_{i,j-1} = t_{i,j-1}) \\
&= \sum_{k_{ij}=0}^{min\{t_{ij},t_{i,j-1}\}} P(T_{ij} = t_{ij}|T_{i,j-1} = t_{i,j-1}, K_{ij} = k_{ij})P(T_{i,j-1} = t_{i,j-1}|K_{ij} = k_{ij}) \\
&= \sum_{k_{ij}=0}^{min\{t_{ij},t_{i,j-1}\}} P(N_{ij} = t_{ij} - k_{ij})\frac{t_{i,j-1}!}{k_{ij}!(t_{i,j-1} - k_{ij})!}\gamma^{k_{ij}}(1-\gamma)^{t_{i,j-1}-k_{ij}} \\
&= \sum_{k_{ij}=0}^{min\{t_{ij},t_{i,j-1}\}} \frac{\xi_{ij}^{t_{ij}-k_{ij}}e^{-\xi_{ij}}}{(t_{ij} - k_{ij})!}\frac{t_{i,j-1}!}{k_{ij}!(t_{i,j-1} - k_{ij})!}\gamma^{k_{ij}}(1-\gamma)^{t_{i,j-1}-k_{ij}} \qquad (3.55)
\end{aligned}$$

It is very difficult to calculate the joint likelihood of T_{ij} with the observation t_{ij} because of the complicated conditional probability (3.55). Therefore, the ML estimators of model parameters are difficult to obtained under the NS-AR(1) model. In the numeric studies, we only conduct simulations to check the performance of GQL and GQL2 methods for the NS-AR(1) model.

As far as the LT model is concerned, the joint likelihood given observations $T = t$ is written as

$$\begin{aligned}
L(\theta|T = t) &= \prod_{i=1}^{I} f(t_i) \\
&= \prod_{i=1}^{I}\prod_{t=1}^{J} f(t_{ij}|t_{i,j-1}) \\
&= \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{(\eta_{ij}^c)^{t_{ij}}exp(-\eta_{ij}^c)}{t_{ij}!}, \qquad (3.56)
\end{aligned}$$

where $T = \{T_{ij}, i = 1, 2, \ldots, I, j = 1, 2, \ldots, J\}$, its observations are $t = \{t_{ij}, i = 1, 2, \ldots, I, j = 1, 2, \ldots, J\}$ with baseline observations $t_{i0}, i = 1, 2, \ldots, I$, and $\eta_{i1} = exp(x_{i1}\beta) + \gamma t_{i0}$. The log-likelihood can be expressed as

$$\ell\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}log[exp(x_{ij}\beta) + \gamma t_{i,j-1}] - \sum_{i=1}^{I}\sum_{j=1}^{J}[exp(x_{ij}\beta) + \gamma t_{i,j-1}]. \qquad (3.57)$$

The ML estimates $\hat{\theta}_{ML}$ are obtained by solving the equations:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(t_{ij}/\eta_{ij}^{c}-1)exp(x_{ij}'\beta)x_{ij} = 0, \tag{3.58}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(t_{ij}/\eta_{ij}^{c}-1)t_{i,j-1} = 0. \tag{3.59}$$

To estimate the covariance matrix of $\hat{\theta}_{ML}$, one need to calculate the Fisher information matrix which can be given by

$$\begin{aligned} I &= -E\left\{\frac{\partial^2 \ell_i(\theta)}{\partial\theta^2}\right\} \\ &= \begin{pmatrix} -E\left\{\frac{\partial^2 \ell_i(\theta)}{\partial\beta^2}\right\} & -E\left\{\frac{\partial^2 logL(\theta)}{\partial\beta\partial\gamma}\right\} \\ -E\left\{\frac{\partial^2 \ell_i(\theta)}{\partial\beta'\partial\gamma}\right\} & -E\left\{\frac{\partial^2 logL(\theta)}{\partial\gamma^2}\right\} \end{pmatrix} \\ &= E\left[\sum_{j=1}^{J}\begin{pmatrix} (\frac{\xi_{ij}t_{ij}}{(\eta_{ij}^{c})^2}-\frac{y_{ij}}{\eta_{ij}^{c}}+1)\xi_{ij}x_{ij}x_{ij}' & \frac{\xi_{ij}}{(\eta_{ij}^{c})^2}t_{ij}t_{i,j-1}x_{ij} \\ \frac{\xi_{ij}}{(\eta_{ij}^{c})^2}t_{ij}t_{i,j-1}x_{ij}' & \frac{t_{ij}}{(\eta_{ij}^{c})^2}t_{i,j-1}^2 \end{pmatrix}\right] \\ &= E\left[\sum_{j=1}^{J}\begin{pmatrix} \frac{\xi_{ij}^2}{\eta_{ij}^{c}}x_{ij}x_{ij}' & \frac{\xi_{ij}}{\eta_{ij}^{c}}t_{i,j-1}x_{ij} \\ \frac{\xi_{ij}}{\eta_{ij}^{c}}t_{i,j-1}x_{ij}' & \frac{t_{i,j-1}^2}{\eta_{ij}^{c}} \end{pmatrix}\right]. \end{aligned} \tag{3.60}$$

Then, covariance matrix of $\hat{\theta}_{ML}$ can be estimated by

$$\hat{V}(\hat{\theta}_{ML}) = I\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\begin{pmatrix} \frac{\xi_{ij}^2}{\eta_{ij}^{c}}x_{ij}x_{ij}' & \frac{\xi_{ij}}{\eta_{ij}^{c}}t_{i,j-1}x_{ij} \\ \frac{\xi_{ij}}{\eta_{ij}^{c}}t_{i,j-1}x_{ij}' & \frac{t_{i,j-1}^2}{\eta_{ij}^{c}} \end{pmatrix}\right]_{\theta=\hat{\theta}_{ML}}^{-1} \tag{3.61}$$

.

It can be seen from (3.58) and (3.59), the ML estimating equations under the LT model (3.47) only involve the first and second order response, which is very similar to (3.35-3.36) under dynamic binary model (3.28). This means that the GQL2 approaches employing the first and the second responses tends to yield almost identical estimates to the ML estimates. Therefore, the GQL2 under the LT model may be the optimal GQL (OGQL) approach.

3.2.5 Simulation studies

In this section, we conduct simulations to examine some interesting issues in statistical inference under different model settings and designs of covariates. Firstly, we will check the performance of the proposed estimation approaches. Under the NS-AR(1) model (3.45), 1000 simulations are carried out for both the GQL and GQL2 approaches. However, we do not conduct simulations for the ML approach under the NS-AR(1) model because of the complexity of the likelihood function. Similarly, under the LT model (3.47), 1000 simulations are implemented for the GQL and ML method. The reason that we do not conduct simulations for the GQL2 approach is that the GQL approach produces almost same efficient estimates as the ML approach. So the GQL2 estimates are expected to be almost same as the GQL and ML estimates.

Next, it is noticed that correct specification of the baseline observations is also of importance producing reliable estimates of model parameters in transition models including the NS-AR(1) model (3.45) and the LT model (3.47). In practice, there may be different choices of the baseline observations, for example, the observations in the year prior to the studying period, or the averages of observations in several years prior to the studying period if available. Therefore, it is possible that false baseline observations are used in estimation procedures. 500 simulations about misspecification of baseline observations are conducted in Section 3.2.5.3.

It can be seen that the NS-AR(1) model and the LT model are very similar in many aspects including model structure, conditional and unconditional expectation structures as well as some higher moments shown in Section 3.2.3. These similarities may lead to challenges in distinguishing them. Therefore, it is meaningful to check the possible misspecification of the two models. For this purpose, 500 simulations are carried out in Section 3.2.5.4.

3.2.5.1 Designs

In the simulation studies in Section 3.2.5.2 and Section 3.2.5.3, we let $\xi = exp(\beta_0 + \beta_1 x_{ij(1)} + \beta_2 x_{ij(2)})$ for both the NS-AR(1) model (3.45) and LT model (3.47). However, we set up different designs of the two covariates $x_{ij(1)}$ and $x_{ij(2)}$ under the two models as below.

Design I: In the NS-AR(1) model, we have chosen the sample size $I = 60$, and number of repeated observations $J = 4$. Three values are chosen for γ: $0.2, 0.5$, and 0.8. The baseline observations t_{i0} are sampled from a Poisson distribution, that is $t_{i0} \sim Poisson(50)$. Among the two time dependent covariates follows, the first one denotes a categorical variable, for example the economic level, air quality level.

$$x_{ij(1)} = \begin{cases} 1, & \text{j=1,2 and i=1,}\ldots\text{, I/2;} \\ 0, & \text{j=3,4 and i=1,}\ldots\text{, I/2;} \\ 0, & \text{j=1,2 and i=n/2+1,}\ldots\text{, I;} \\ 1, & \text{j=3,4 and i=n/2+1,}\ldots\text{, I,} \end{cases}$$

and the second covariate is a subject-constant but time-varying variable which represents a consituous variable.

$$x_{ij(2)} \sim N(j/10 - 0.1, 1), \quad j = 1, 2, 3, 4 \text{ for all } i = 1, 2, \ldots, 60.$$

Design II: In the LT model, we choose the same sample size $I = 60$, but different number of repeated measurements $J = 6$. For different values are assigned to γ: $0.05, 0.2, 0.5$, and 0.9. The baseline observations t_{i0} are also sampled from the Poisson distribution $Poisson(50)$. Similar to two covariates in the NS-AR(1) model, the two time-varying covariates are given by

$$x_{ij(1)} = \begin{cases} -1, & \text{j=1,2 and i=1,..., I/3;} \\ 0, & \text{j=3,4 and i=1,..., I/3;} \\ 1, & \text{j=5,6 and i=1,..., I/3;} \\ 0, & \text{j=1,2,3 and i=I/3+1,..., 2I/3;} \\ 1, & \text{j=4,5,6 and i=I/3+1,..., 2I/3;} \\ 1, & \text{j=1,2,3 and i=2I/3+1,..., I;} \\ 0, & \text{j=4,5,6 and i=2I/3+1,..., I,} \end{cases}$$

and

$$x_{ij(2)} \sim \begin{cases} N((j-1)/10, 1), & \text{for } j = 1, \ldots, 4, \text{and} i = 1, \ldots, I; \\ N(3/10, 1), & \text{j=5; and i=1,...,I} \\ N(2/10, 1), & \text{j=6; and i=1,...,I.} \end{cases}$$

In Section 3.2.5.4, we use Design II for both the NS-AR(1) model and the LT model to compare the misspecification between them.

3.2.5.2 Estimation of model parameters

In this section, 1000 simulation are conducted to check the performance of the proposed approaches in estimating interested parameters in the NS-AR(1) model (3.45) and LT model (3.47). The results under the NS-AR(1) model and LT model are given in Table 3.2 and Table 3.3, respectively.

In each case of γ, we applied the GQL and GQL2 approaches to estimate the model parameters $\theta = (\beta', \gamma)'$, where $\beta = (\beta_0, \beta_1, \beta_2)'$ are the regression coefficients, γ is the dynamic dependence parameter. From Table 3.2, it can be seen that both the GQL and GQL2 approaches yield approximately unbiased estimates of model parameters. For example, when $\gamma = 0.2$, we have the GQL estimates of $\hat{\beta}_{GQL} = (1.0009, -1.0010, 0.9988)'$, and the GQL estimate of γ is 0.1999. In the same case, $\hat{\beta}_{GQL2} = (1.0093, -1.0000, 0.9985)'$ and $\hat{\gamma}_{GQL2} = 0.1997$. They are very close to the true values $\beta = (1, -1, 1)$ and $\gamma = 0.2$. The estimated standard errors (ESE) derived from GQL estimator (3.52) and GQL2 estimator (3.54) are almost identical to their corresponding simulated standard errors (SSE). Therefore, the coverage probabilities of 95% CI's are mostly close to the true nominal level 0.95 under the two approaches. Another point is that the GQL2 approach tends to have slightly higher efficiency than the GQL approach when the ESE's and SSE's are concerned. For example, in the case that $\gamma = 0.8$, the (SSE, ESE) of $\hat{\beta}_1$ are (0.1899, 0.1874) under the GQL approach, and (0.1891, 0.1870) under the GQL2 approach. However, in most cases the GQL performs almost as well as the GQL2 as far as the SM, SSE, ESE and CPr are concerned.

Under the LT model, for different values of the dynamic dependence parameter, the GQL and ML approaches are used to estimate model parameters $\theta = (\beta', \gamma)'$ with $\beta = (\beta_0, \beta_1, \beta_2)'$. The simulation results are given in Table 4.3. It is shown that the GQL and ML approaches yield approximately unbiased estimates of θ. For example, when $\gamma = 0.5$, we have the GQL estimates of $\hat{\beta}_{GQL} = (1.0004, -1.0007, 0.9989)'$ with SSE's (0.1045, 0.0843,0.0431) and ESE's (0.1070, 0.0856, 0.0430), and the GQL estimate of γ is 0.4994 with a SSE 0.0100 and an ESE 0.0098. In this case, the ML

Table 3.2: Simulation results for the NS-AR(1) model with the true values of parameters $\beta = (1, -1, 1)$

	$\gamma = 0.2$		$\gamma = 0.5$		$\gamma = 0.8$	
Quantity	GQL	GQL2	GQL	GQL2	GQL	GQL2
SM(β_0)	1.0009	1.0093	0.9949	0.9951	0.9976	0.9983
SSE	0.0701	0.0692	0.1032	0.1031	0.0964	0.0961
ESE	0.0698	0.0691	0.1019	0.1018	0.0966	0.0965
CPr	0.942	0.946	0.943	0.941	0.940	0.944
SM(β_1)	-1.0010	-1.0000	-1.0163	-1.0160	-1.0128	-1.0127
SSE	0.0572	0.0564	0.1825	0.1823	0.1899	0.1891
ESE	0.0545	0.0541	0.1796	0.1793	0.1874	0.1870
CPr	0.941	0.944	0.952	0.952	0.950	0.952
SM(β_2)	0.9988	0.9985	0.9986	0.9984	0.9991	0.9987
SSE	0.0347	0.0346	0.0567	0.0566	0.0409	0.0408
ESE	0.0346	0.0345	0.0563	0.0562	0.0418	0.0417
CPr	0.944	0.951	0.951	0.949	0.951	0.953
SM(γ)	0.1999	0.1997	0.4999	0.4999	0.7999	0.7999
SSE	0.0080	0.0079	0.0083	0.0083	0.0053	0.0053
ESE	0.0079	0.0079	0.0083	0.0083	0.0054	0.0054
CPr	0.949	0.949	0.950	0.954	0.956	0.956

estimates are $\hat{\beta}_{ML} = (1.0014, -1.0002, 0.9989)'$ with SSE's (0.1020, 0.0819, 0.0425) and ESE's (0.1045, 0.0836, 0.0425) and $\hat{\gamma}_{ML} = 0.4993$ with a SSE 0.0099 and an ESE 0.0098. These estimates of θ are very close to the true values $\theta = (1, -1, 1, 0.5)$, and these ESE's are also very close to the corresponding SSE's.

It also can be seen through Table 3.3 that ML approach tends to have slightly higher efficiency than the GQL approach, that is, the SSE's (ESE's) under the ML approach are a little smaller than the corresponding SSE's (ESE's) under the GQL approach. However, in most cases, the GQL approach produces estimates that are very close to the ML estimates as far as the SM, SSE, ESE and CPr are concerned. This implies that the first order responses $\{T_{ij}\}$ include almost all information about the model parameters θ. This is also why we did not conduct the simulation about the GQL2 approach under the LT model.

In summary, the proposed approaches can effectively estimate the model parameters in the NS-AR(1) model and the LT model. The GQL approach yields highly efficient estimates of parameters. This implies meaningful application of this method because of its simplicity compared with the GQL2 method under the NS-AR(1) model and the ML method under the LT model.

3.2.5.3 Misspecified baseline observations

To check the influence of possible misspecification of baseline observations on parameter estimation, in this subsection, we conduct 500 simulations to check the performance of the proposed approaches for NS-AR(1) model (3.45) and LT model (3.47) taking the mis-specified baseline observations into consideration. In simulation, we choose $\beta = (\beta_0, \beta_1, \beta_2)' = (1, -1, 1)'$ and $\gamma = 0.65$. The true baseline observations

Table 3.3: Simulation results for the LT model with the true values of parameters $\beta = (1, -1, 1)$

	$\gamma = 0.05$		$\gamma = 0.2$		$\gamma = 0.5$		$\gamma = 0.9$	
Quantity	GQL	ML	GQL	ML	GQL	ML	GQL	ML
SM(β_0)	1.0025	1.0029	1.0007	1.0024	1.0004	1.0014	0.9909	0.9934
SSE	0.0599	0.0600	0.0701	0.0692	0.1054	0.1020	0.1956	0.1928
ESE	0.0597	0.0595	0.0698	0.0691	0.1070	0.1045	0.1876	0.1847
CPr	0.948	0.950	0.942	0.946	0.950	0.949	0.938	0.935
SM(β_1)	-0.9990	-0.9988	-1.0010	-1.0000	-1.0007	-1.0002	-1.0047	-1.0032
SSE	0.0539	0.0539	0.0572	0.0564	0.0843	0.0819	0.1585	0.1564
ESE	0.0522	0.0521	0.0545	0.0541	0.0856	0.0836	0.1507	0.1484
CPr	0.937	0.937	0.941	0.944	0.962	0.955	0.937	0.934
SM(β_2)	0.9972	0.9971	0.9988	0.9985	0.9989	0.9989	1.0013	1.0008
SSE	0.0278	0.0278	0.0347	0.0346	0.0431	0.0425	0.0653	0.0647
ESE	0.0277	0.0277	0.0346	0.0345	0.0430	0.0425	0.0643	0.0639
CPr	0.954	0.954	0.944	0.951	0.950	0.949	0.939	0.946
SM(γ)	0.0499	0.0498	0.1999	0.1997	0.4994	0.4993	0.8994	0.8994
SSE	0.0054	0.0053	0.0080	0.0079	0.0100	0.0099	0.0083	0.0082
ESE	0.0053	0.0052	0.0079	0.0079	0.0098	0.0098	0.0083	0.0082
CPr	0.944	0.939	0.949	0.949	0.942	0.950	0.951	0.946

t_{i0} are generated from $Poisson(50)$ under both the NS-AR(1) model and the LT model, whereas, in conducting statistical inference for the two models, we assume that all the baseline observations are mis-specified to be 50. The simulation results in presence of the mis-specified baseline observations under the NS-AR(1) model and LT model are given in Table 3.4.

As discussed in Section 3.2.1, the baseline observations have much influence on the NS-AR(1) model. Actually, misspecified baseline observations are expected to have significant influence on the statistical inference based on both the NS-AR(1) model and the LT model. This is because the misspecified baseline observations lead to wrong expectations η_{i1}, hence lead to incorrect η_{ij} for $j = 2, \ldots, J$. Further, all moments of the response are incorrect due to the misspecification of baseline observations. Therefore, statistical inferences based on the estimating equation-based approaches and the ML approach which highly depends on the accuracy of data are not reliable any more. This can be demonstrated from Table 3.4. For example, under the NS-AR(1) model, the GQL estimate of $\beta_0 = 1$ is 0.9300, and the corresponding GQL2 estimate is 0.9411. The CPr's under the two approaches are 0.900 and 0.912, which are significantly lower than the nominal level 0.95. Similarly, under the LT model, $\hat{\beta}_{0(GQL)} = 0.9471$ and $\hat{\beta}_{0(GQL2)} = 0.9524$ both of which have significant biases from the true value 1. The CPr's of β_0 under the GQL approach is 0.936 which is much smaller than 0.95. Further note that the biases of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\gamma}$ under both approaches are significant relative to the biases of these estimates employing correct baseline observations, and the CPr's in the case of mis-specified observations are significantly biased from 0.95.

Based on the simulation results in Table 3.4. we conclude that the baseline ob-

Table 3.4: Mis-specifying the baseline observation $t_{i0} = 50$ when $t_{i0} \sim Pois(50)$ with $\beta = (1, -1, 1)$ and $\gamma = 0.65$.

		Correct				Misspecified			
		β_0	β_1	β_2	γ	β_0	β_1	β_2	γ
Method		NS-AR(1) model							
GQL	SM	0.9954	-1.0129	1.0000	0.6494	0.9300	-1.1531	1.0279	0.6573
	SSE	0.0966	0.1833	0.0422	0.0069	0.0966	0.2055	0.0410	0.0069
	ESE	0.0971	0.1876	0.0433	0.0069	0.0979	0.2185	0.0428	0.0069
	CPr	0.940	0.960	0.946	0.950	0.900	0.974	0.900	0.814
GQL2	SM	0.9959	-1.0132	0.9997	0.6495	0.9411	-1.0829	1.0243	0.6557
	SSE	0.0967	0.1830	0.0422	0.0069	0.0977	0.1956	0.0419	0.0069
	ESE	0.0969	0.1866	0.0433	0.0069	0.0974	0.2048	0.0427	0.0070
	CPr	0.942	0.962	0.946	0.946	0.912	0.978	0.910	0.862
		LT model							
GQL	SM	0.9978	-1.0024	0.9974	0.6494	0.9471	-1.0293	1.0168	0.6536
	SSE	0.1352	0.1070	0.0525	0.0094	0.1391	0.1094	0.0539	0.0094
	ESE	0.1369	0.1077	0.0516	0.0094	0.1376	0.1093	0.0508	0.0095
	CPr	0.956	0.958	0.960	0.954	0.936	0.948	0.906	0.938
ML	SM	1.0012	-1.0000	0.9968	0.6493	0.9524	-1.0256	1.0157	0.6535
	SSE	0.1303	0.1040	0.0516	0.0094	0.1344	0.1062	0.0529	0.0094
	ESE	0.1340	0.1054	0.0510	0.0094	0.1352	0.1072	0.0502	0.0094
	CPr	0.962	0.964	0.958	0.952	0.952	0.960	0.916	0.934

servations do influence the statistical inferences about models. Therefore, one should be careful to choose baseline values in practice.

3.2.5.4 Misspecification of models

In this section, 500 simulation are conducted to check the influence of the misspecification of models on the estimates of model effects. We consider two cases of the misspecification of models.

Case 1. Under the true NS-AR(1) model (3.45), we mis-specify the model as the LT model (3.47), of which the simulation results are given in the Table 3.5.

Case 2. Under the true LT model (3.47), we mis-specify the model as the NS-AR(1) model (3.45), of which the results are given in Table 3.6.

It can be seen from the two tables that, the estimates under all approaches in both Case I and Case II are still approximately unbiased. For example, when $\gamma = 0.5$ for Case I in Table 3.5, the GQL estimates $(\hat{\beta}'_{GQL}, \hat{\gamma}_{GQL}) = (0.9984, -1.0002, 0.9996, 0.4999)$, and the ML estimates $(\hat{\beta}'_{ML}, \hat{\gamma}_{ML}) = (0.9996, -0.9996, 0.9993, 0.4998)$. For Case II in Table 3.6, the GQL estimates $(\hat{\beta}'_{GQL}, \hat{\gamma}_{GQL}) = (1.0009, -1.0239, 0.9988, 0.4997)$, and the ML estimates $(\hat{\beta}'_{ML}, \hat{\gamma}_{ML}) = (1.0066, -0.9826, 0.9976, 0.4983)$. All of these estimates have ignorable biases from the true values $(\beta', \gamma) = (1, -1, 1, 0.5)$. This is because that the expectations η_{ij} under both the NS-AR(1) model and the LT model share the same formula according to (3.48) in Section 3.2.3, which leads to the unbiased GQL estimating equations (3.49) even under the misspecified models. Similarly, the same conditional expectations η^c_{ij} according to (3.46) and (3.47) leads to the unbiased ML estimating equations (3.58-3.59) under the misspecified models.

As far as the GQL2 approach under the misspecified NS-AR(1) model in Table 3.6 is concerned, the GQL2 estimates may have slightly greater biases than the GQL estimates have, especially for large values of γ, for instance, when $\gamma = 0.9$, the GQL estimate of β_0 is 0.9913, which has less bias than the corresponding GQL2 estimate $\hat{\beta}_{0(GQL2)} = 0.9845$. This may be due to the different expectations of T_{ij}^2 under the NS-AR(1) and LT models showed in Section 3.2.3, which leads to a little bias of the GQL2 estimating equations (3.53) under the misspecified model. However, it is obvious that the biases of the GQL2 estimates are not significant compared with the true values of parameters. This is because that the first order responses $\{T_{ij}\}$ and the pairwise products of the response $\{T_{ij}T_{iu}\}$, which have the same expectations under the two models, have already included almost sufficient information about the model parameters, especially for small values of γ. Therefore, the GQL2 estimating equations (3.53) have tiny bias which do not have significant influence on the estimates of model parameters.

However, the estimated standard errors and coverage probabilities of 95% CI's do not have satisfactory performance in the case that the dynamic dependence parameter γ takes large values in both Case I and Case II. In Case I, the GQL estimators and the ML estimators of standard errors derived from (3.52) and (3.61), respectively, tend to have worse overestimated standard errors as the value of γ increases, which leads to conservative confidence intervals. As a result, the CPr's are much higher than the nominal level 0.95. For example, when $\gamma = 0.9$ in Table 3.5, the ESE of $\hat{\beta}_{0(GQL)}$ is 0.1925, whereas the corresponding SSE takes a much smaller value 0.1001. Similarly the value of ESE of $\hat{\beta}_{0(ML)}$ is 0.1920 which is also much greater than the corresponding value of SSE which is 0.0999. The CPr's of β_0 under the two approaches are, respectively, 0.999 and 1.000, which are all much greater than 0.95.

As far as Case II is concerned, the GQL estimators and the GQL2 estimators of standard errors derived from (3.52) and (3.54), respectively, tend to have more severely underestimated standard errors as γ increases, which leads to lower CPr's than the nominal level 0.95. For instance, when $\gamma = 0.9$ in Table 3.6, the ESE of $\hat{\beta}_{0(GQL)}$ is 0.0891, and it is much smaller than the value of the corresponding SSE 0.1966. Similarly the value of ESE of $\hat{\beta}_{0(GQL2)}$ is 0.0890, while the corresponding SSE is a much larger value 0.1964. The CPr's for β_0 under the two approaches have the same value 0.624 which is much smaller than the nominal level 0.95. The occurrence of this phenomenon may be because much error information is used in estimating standard errors and in constructing confidence intervals under all approaches in the two cases. It may be because the other moments, besides those expectations of the first order response and pairwise products of the response, also play important roles in estimating standard errors of $\hat{\theta}$ as the degree of dynamic dependence increases.

In summary, the misspecification of models described in Case I and Case II do not affect the unbiasedness of estimates of model parameters. However, the misspecification may lead to severely biased estimation of standard errors of $\hat{\theta}$, then lead to poor confidence intervals.

Table 3.5: Misspecified LT model under true NS-AR(1) model, where $\beta = (1, -1, 1)$.

	$\gamma = 0.05$		$\gamma = 0.2$		$\gamma = 0.5$		$\gamma = 0.9$	
Quantity	GQL	ML	GQL	ML	GQL	ML	GQL	ML
SM(β_0)	0.9970	0.9972	0.9977	0.9987	0.9984	0.9996	0.9953	0.9958
SSE	0.0590	0.0588	0.0702	0.0699	0.0900	0.0894	0.1001	0.0999
ESE	0.0599	0.0597	0.0720	0.0712	0.1065	0.1050	0.1925	0.1920
CPr	0.953	0.953	0.958	0.956	0.980	0.976	0.999	1.000
SM(β_1)	-0.9998	-0.9997	-1.0005	-0.9999	-1.0002	-0.9996	-1.0036	-1.0034
SSE	0.0523	0.0521	0.0575	0.0571	0.0723	0.0720	0.0853	0.0852
ESE	0.0523	0.0522	0.0599	0.0594	0.0851	0.0839	0.1510	0.1506
CPr	0.947	0.949	0.951	0.956	0.975	0.978	0.999	0.999
SM(β_2)	0.9999	0.9999	0.9988	0.9986	0.9996	0.9993	0.9999	0.9999
SSE	0.0270	0.0270	0.0318	0.0317	0.0372	0.0370	0.0362	0.0362
ESE	0.0277	0.0277	0.0320	0.0318	0.0430	0.0428	0.0665	0.0665
CPr	0.958	0.958	0.955	0.953	0.972	0.976	0.998	0.998
SM(γ)	0.0500	0.0500	0.2003	0.2002	0.4999	0.4998	0.9000	0.9000
SSE	0.0053	0.0053	0.0077	0.0076	0.0070	0.0070	0.0032	0.0032
ESE	0.0053	0.0053	0.0082	0.0082	0.0097	0.0097	0.0082	0.0082
CPr	0.943	0.939	0.960	0.965	0.994	0.994	1.000	1.000

Table 3.6: Misspecified NS-AR(1) model under true LT model, where .

	$\gamma = 0.05$		$\gamma = 0.2$		$\gamma = 0.5$		$\gamma = 0.9$	
Quantity	GQL	GQL2	GQL	GQL2	GQL	GQL2	GQL	GQL2
SM(β_0)	0.9980	0.9986	0.9947	0.9950	1.0009	1.0066	0.9913	0.9845
SSE	0.0693	0.0690	0.0794	0.0791	0.1169	0.1170	0.1966	0.1964
ESE	0.0660	0.0657	0.0736	0.0730	0.0911	0.0908	0.0891	0.0890
CPr	0.946	0.944	0.936	0.934	0.866	0.864	0.624	0.624
SM(β_1)	-0.9966	-0.9970	-1.0022	-1.0016	-1.0239	-0.9826	-1.0945	-1.0099
SSE	0.1016	0.1009	0.1296	0.1294	0.2598	0.2490	0.7070	0.6753
ESE	0.1030	0.1011	0.1284	0.1279	0.1884	0.1811	0.2080	0.1731
CPr	0.954	0.952	0.956	0.952	0.876	0.848	0.646	0.646
SM(β_2)	0.9992	0.9993	1.0025	1.0021	0.9988	0.9970	0.9983	1.0022
SSE	0.0325	0.0323	0.0364	0.0361	0.0527	0.0529	0.0769	0.0765
ESE	0.0318	0.0315	0.0357	0.0352	0.0425	0.0423	0.0385	0.0384
CPr	0.942	0.940	0.940	0.939	0.880	0.866	0.674	0.682
SM(γ)	0.0499	0.0496	0.1995	0.1997	0.4997	0.4983	0.8992	0.8993
SSE	0.0053	0.0053	0.0082	0.0081	0.0110	0.0110	0.0101	0.0101
ESE	0.0050	0.0050	0.0077	0.0076	0.0081	0.0081	0.0039	0.0039
CPr	0.934	0.937	0.934	0.932	0.852	0.842	0.548	0.546

Chapter 4

Modeling Misclassified Longitudinal Categorical Data

4.1 Overview

Suppose that T_{ij} is the true but unobservable categorical response and Y_{ij} is the observed response for subject i at the jth time point, $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. It is assumed that the true categorical response T_{ij} follows a multinomial distribution, that is, $T_{ij} \sim Multinomial(1, \eta_{ij})$. We also assume that the dynamic pattern of T_{ij} follows the nonlinear transition model (3.3, 3.11, 3.12) or (3.13-3.14) developed in Section 3.1 of Chapter 3. Suppose that the categorical variable Y_{ij} is a vector of dimension s and it's inherent variable T_{ij} is of dimension r. In most cases, $s = r$. However there are certain situations where s and r are different, see examples in Chapter 2. Let $\widehat{T}_{ij} = (T'_{ij}, 1 - \mathbf{1}'T_{ij})'$ and $\widehat{Y}_{ij} = (Y'_{ij}, 1 - \mathbf{1}'Y_{ij})'$ be the full latent and observed categorical variables, respectively. We assume that the FMC matrix $\widehat{\Pi}$

is constant over time and subjects.

$$\widehat{\Pi} = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1,r+1} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2,r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{s+1,1} & \pi_{s+1,2} & \cdots & \pi_{s+1,r+1} \end{pmatrix},$$

where $\pi_{uv} = P(Y_{ij(u)} = 1|T_{ij(v)} = 1)$, $\pi_{u,r+1} = P(Y_{ij(u)} = 1|\mathbf{1}'T_{ij} = 0)$. $\pi_{s+1,v} = P(\mathbf{1}'Y_{ij} = 0|T_{ij(v)} = 1) = 1 - \sum_{k=1}^{s} \pi_{kv}$, and $\pi_{s+1,r+1} = P(\mathbf{1}'Y_{ij} = 0|\mathbf{1}'T_{ij} = 0) = 1 - \sum_{k=1}^{s} \pi_{k,r+1}$ for $u = 1, 2, \ldots, s$ and $v = 1, 2, \ldots, r$.

The MC matrix Π is given by

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1,r+1} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2,r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{s,1} & \pi_{s,2} & \cdots & \pi_{s,r+1} \end{pmatrix} \triangleq [\pi_1, \pi_2, \ldots, \pi_{r+1}],$$

where π_i's are vectors of dimension s. If $\mathbf{1}'\pi_i = 1$ for all $i = 1, 2, \ldots, r+1$, Π will become the FMC-matrix describes the misclassification from $r+1$ inherent categories to s observed categories. Let $\Pi_r = [\pi_1, \pi_2, \ldots, \pi_r]$ be the submatrix of Π deleting the last column π_{r+1}.

The misclassification model for the longitudinal data Y_{ij} and T_{ij} is given by

$$\begin{aligned} Y_{ij} &= \Pi * \widetilde{T}_{ij} \\ &= \Pi_r * T_{ij} + \pi_{r+1} * (1 - \mathbf{1}'T_{ij}) \qquad (4.1) \\ &= \sum_{u=1}^{r} \pi_s * T_{ij(u)} + \pi_{r+1} * (1 - \mathbf{1}'T_{ij}). \end{aligned}$$

According to the definition of the generalized thinning operation, the expectation of the observed categorical variable Y_{ij} in this model is given by the following formulas

$$\begin{aligned}\mu_{ij} &= E(Y_{ij}) \\ &= E[\Pi_r * T_{ij} + \pi_{r+1} * (1 - \mathbf{1}'T_{ij})] \\ &= (\Pi_r - \pi_{r+1}\mathbf{1}')'\eta_{ij} + \pi_{r+1}. \end{aligned} \tag{4.2}$$

The covariance matrix of Y_{ij} is

$$\begin{aligned} Var(Y_{ij}) &= Var[E(Y_{ij}|T_{ij})] + E[Var(Y_{ij}|T_{ij})] \\ &= Var(\Pi\widetilde{T}_{ij}) + E[\sum_{u=1}^{r} T_{ij(u)}V_{\pi_u} + E[(1 - \mathbf{1}'T_{ij})V_{\pi_{r+1}}] \\ &= \Pi Var(\widetilde{T}_{ij})\Pi' + \sum_{u=1}^{r} \eta_{ij(u)}V_{\pi_u} + (1 - \mathbf{1}'\eta_{ij})V_{\pi_{r+1}} \\ &= [\Pi_r, \pi_{r+1}] \begin{pmatrix} Var(T_{ij}) & -Var(T_{ij}) \\ -\mathbf{1}'Var(T_{ij}) & \mathbf{1}'Var(T_{ij})\mathbf{1} \end{pmatrix} [\Pi_r, \pi_{r+1}]' \\ &\quad + \sum_{u=1}^{r} \eta_{ij(u)}V_{\pi_u} + (1 - \mathbf{1}'\eta_{ij})V_{\pi_{r+1}} \\ &= \sum_{u=1}^{r} \eta_{ij(u)}V_{\pi_u} + (1 - \mathbf{1}'\eta_{ij})V_{\pi_{r+1}} \\ &\quad + (\Pi_r - \pi_{r+1}\mathbf{1}')Var(T_{ij})(\Pi_r - \pi_{r+1}\mathbf{1}')'. \end{aligned}$$

As shown in Section 2.3 of Chapter 2, $Y_{ij} \sim Multinomial(1, \mu_{ij})$. Hence the covariance of Y_{ij} can be written in an alternative form

$$Var(Y_{ij}) = V_{\mu_{ij}} = diag(\mu_{ij}) - \mu_{ij}\mu_{ij}'. \tag{4.3}$$

The covariance between Y_{ij} and Y_{iu} is given by

$$\begin{aligned}
Cov(Y_{ij}, Y_{iu}) &= E(Y_{ij}Y'_{iu}) - \mu_{ij}\mu'_{iu} \\
&= E\{[\pi_{r+1} - (\Pi_r - \pi_{r+1}\mathbf{1}')T_{ij}][\pi_{r+1} - (\Pi_r - \pi_{r+1}\mathbf{1}')T_{iu}]'\} \\
&\quad -[\pi_{r+1} - (\Pi_r - \pi_{r+1}\mathbf{1}')\eta_{ij}][\pi_{r+1} - (\Pi_r - \pi_{r+1}\mathbf{1}')\eta_{iu}]' \\
&= (\Pi_r - \pi_{r+1}\mathbf{1}')Cov(T_{ij}, T_{iu})(\Pi_r - \pi_{r+1}\mathbf{1}')'.
\end{aligned}$$

These moments will be used to develop the GEE, GQL methods to estimate model parameters. Actually, it is much easier to calculate theses moments based on the explicit model (4.1) than that based on the classic descriptive misclassification model.

We now discuss the maximum likelihood (ML) approach which produces efficient estimators of the interested parameters. Suppose that we have observations of the manifest variable Y and the latent T, that is, $y = \{y_{ij}, i = 1, \ldots, I \text{ and } j = 1, \ldots, J\}$ and $t = \{t_{ij}, i = 1, \ldots, I \text{ and } j = 0, \ldots, J\}$, where t_{i0}'s are baseline observations and assumed to be known. The complete likelihood function is given by

$$\begin{aligned}
L(\theta|y,t) &= \prod_{i=1}^{I} f(y_i|t_i)f(t_i) \\
&= \prod_{i=1}^{I}\prod_{j=1}^{J} f(y_{ij}|t_{ij}) \prod_{i=1}^{I}\prod_{j=1}^{J} g_{i,j|j-1},
\end{aligned} \tag{4.4}$$

where $g_{i,j|j-1}$ is given in section 1 of Chapter34. Under some regularity conditions, such as all elements of FMC-matrix $\widehat{\Pi}$ are within the interval $(0, 1)$, which implies that $0 < \pi_{uv} < 1$, for $u = 1, \ldots, s+1$ and $v = 1, \ldots, r+1$, the conditional likelihood function of observations y_{ij} given t_{ij} is given by

$$
\begin{aligned}
f(y_{ij}|t_{ij}) &= \left[(1-\mathbf{1}'\pi_{r+1})^{(1-\mathbf{1}'y_{ij})}\prod_{u=1}^{s}(1-\mathbf{1}'\pi_{r+1})^{y_{ij(u)}}\right]^{1-\mathbf{1}'t_{ij}} \\
&\quad\times\prod_{v=1}^{r}\left[(1-\mathbf{1}'\pi_v)^{(1-\mathbf{1}'y_{ij})}\prod_{u=1}^{s}\pi_{uv}^{y_{ij(u)}}\right]^{t_{ij(v)}} \\
&= \left[\prod_{v=1}^{r}\prod_{u=1}^{s}\pi_{uv}^{y_{ij(u)}t_{ij(v)}}\right]\left[\prod_{v=1}^{r}(1-\mathbf{1}'\pi_v)^{(1-\mathbf{1}'y_{ij(u)})t_{ij(v)}}\right] \\
&\quad\times\left[\prod_{u=1}^{s}\pi_{ur+1}^{y_{ij(u)}(1-\mathbf{1}'t_{ij})}\right]\left[(1-\mathbf{1}'\pi_{r+1})^{(1-\mathbf{1}'y_{ij})(1-\mathbf{1}'t_{ij})}\right]. \qquad (4.5)
\end{aligned}
$$

From expressions (4.4) and (4.5), it can be seen that it is difficult to calculate the marginal likelihood of the observed data y. Hence the ML estimates cannot be obtained by directly maximize the likelihood function of y. In this case, the expectation & maximization (EM) algorithm is helpful to calculate the ML estimates in an iterative procedure.

The log-likelihood function of complete data can be expressed as

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\sum_{v=1}^{r}(x'_{ij}\beta_v+t'_{i,j-1}\gamma_v)t_{ij(v)}-log\{1+\sum_{v=1}^{r}exp(x'_{ij}\beta_v+t'_{i,j-1}\gamma_v)\}\right] \\
&\quad+\sum_{i=1}^{I}\sum_{j=1}^{J}[\sum_{u=1}^{s}y_{ij(u)}(1-\mathbf{1}'t_{ij})log(\pi_{ur+1})+(1-\mathbf{1}'y_{ij})(1-\mathbf{1}'t_{ij})log(1-\mathbf{1}'\pi_{r+1})] \\
&\quad+\sum_{i=1}^{I}\sum_{j=1}^{J}[\sum_{u=1}^{s}\sum_{v=1}^{r}y_{ij(u)}t_{ijv}log\pi_{uv}+\sum_{v=1}^{r}(1-\mathbf{1}'y_{ij(u)})t_{ij(v)}log(1-\mathbf{1}'\pi_v)] \qquad (4.6)
\end{aligned}
$$

In this function, the values $t_{ij}, i=1,\ldots,I$ and $j=1,\ldots,J$ are not observable. The EM algorithm starts with an initial value $\theta^{(0)}$. Denoting $\theta^{(k)}$ as the estimate of θ at the kth iteration, the $(k+1)$th iteration of the EM can be developed below.

E-step: Find the expected complete-data log-likelihood function if θ were $\theta^{(k)}$: $Q(\theta|\theta^{(k)},y)=E_T(\ell(\theta)|Y=y,\theta^{(k)})$

M-step: Determine $\theta^{(k+1)}$ by maximizing this expected log-likelihood function $Q(\theta|\theta^{(k)},y)$.

If the regularity conditions for π_{uv}'s are violated, for example, $\pi_{uv} = 0$, we have $y_{ij(u)}t_{ij(v)}log(\pi_{uv}) = 0$, in the log-likelihood function $\ell(\theta)$. Therefore the ML estimating procedure is still applicable.

We consider two special cases in the following sections. In Section 4.2, we conduct the analysis of the misclassified longitudinal binary data. In section 4.4, the misclassified binary data with non-ignorable missing information are analyzed.

4.2 Misclassified Longitudinal Binary Data

In epidemiologic studies such as the child asthma prevention and control program, a child's disease status is often determined based on the information provided by some proforma questionnaires completed by the parents. Questionnaires are widely used because they are relatively simple and economical when compared to the clinical examination of each child. However, it is impossible to design and conduct a completely reliable questionnaire due to the complexities of wide range of severity of the disease, triggers, and lack of medical knowledge among the public [Jenkins, et al. (1996)]. In addition, it is challenging sometimes for parents to identify symptoms of wheezing from cold symptoms. All of these reasons result in classification errors on children asthma statuses. In this section, we develop statistical models and methods to deal with misclassified longitudinal binary data to obtain more reliable inference by taking measurement errors into account.

Table 4.1: Misclassified Asthma Status

	Asthma (T_{ij})	
Diagnosis of test (Y_{ij})	Infected (1)	Healthy (0)
Positive (1)	π^+	ϵ^-
Negative (0)	ϵ^+	π^-

4.2.1 Model description

Let Y_{ij} denote the diagnosed status (positive=1, negative=0) for the ith child at the jth time point, and T_{ij} denote the corresponding true status (infected=1, healthy=0), for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. The possible diagnoses for a child's asthma status are shown in table 4.1. In this table, $\pi^- = Pr(Y_{ij} = 0|T_{ij} = 0)$ is the specificity which represents the probability of the correct diagnosis for a healthy individual. ϵ^- is the probability of type I error of a diagnostic test, that is the probability of misclassifying a healthy subject as a disease case. It is obvious that $\pi^- + \epsilon^- = 1$. Similarly, π^+ refers to the sensitivity which is defined as the probability of the correct diagnosis for an infected subject, that is, $Pr(Y_{ij} = 1|T_{ij} = 1)$. ϵ^+ is the probability of type II error of a diagnostic test, that is, the probability of misclassifying a patient to be free of the disease. It is apparent that $\pi^+ + \epsilon^+ = 1$. The corresponding FMC-matrix is given by

$$\widehat{\Pi} = \begin{pmatrix} \pi^+ & \epsilon^- \\ \epsilon^+ & \pi^- \end{pmatrix}. \tag{4.7}$$

To accommodate the biomedical background, we develop the misclassification model for binary data based on the two quantities, namely, the sensitivity π^+ and the specificity π^-.

It is assumed that $T_{ij} \sim b(1, \eta_{ij})$, where $\eta_{ij} = P(T_{ij} = 1)$ is a function of some risk factors or covariates. The classification error model is expressed as

$$Y_{ij} = \pi^+ * T_{ij} + (1 - \pi^-) * (1 - T_{ij}), \tag{4.8}$$

where $*$ is the binomial thinning operation described in Chapter 2. (Y_{ij}, T_{ij}) may take one of the four pairs of values: (1, 1), (0, 0), (0, 1) and (1, 0). The former two cases indicate the two correct diagnoses, respectively, while the latter two cases imply two type of misdiagnoses. This model can completely and explicitly describe the misclassification from T_{ij} to Y_{ij}.

Based on the model (4.8), it is easy to calculate the required moments for the development of the estimation approaches. The expectation and variance of Y_{ij} are given by

$$E(Y_{ij}) \;=\; \mu_{ij} = 1 - \pi^- + (\pi^- + \pi^+ - 1)\eta_{ij},$$

and

$$\begin{aligned} Var(Y_{ij}) &= \pi^+(1-\pi^+)\eta_{ij} + \pi^-(1-\pi^-)(1-\eta_{ij}) + (\pi^+ + \pi^- - 1)^2 Var(T_{ij}) \\ &= \mu_{ij}(1-\mu_{ij}). \end{aligned}$$

The covariance between Y_{ij} and Y_{iu} for $u < j$, is formulated by

$$
\begin{aligned}
Cov(Y_{ij}, Y_{iu}) &= E(Y_{ij}Y_{iu}) - E(Y_{ij})E(Y_{iu}) \\
&= E\{[\pi^{+}T_{ij} + (1-\pi^{-})(1-T_{ij})][\pi^{+}T_{iu} + (1-\pi^{-})(1-T_{iu})]\} \\
&\quad -1-\pi^{-}) + (\pi^{-} + \pi^{+} - 1)\eta_{ij}][(1-\pi^{-}) + (\pi^{-} + \pi^{+} - 1)\eta_{iu}] \\
&= (\pi^{-} + \pi^{+} - 1)^2 cov(T_{ij}, T_{iu}),
\end{aligned}
$$

where $(\pi^{-} + \pi^{+} - 1)$ is the Youden's index which captures the performance of a diagnostic test [Youden (1950)].

To develop the GQL and GQL2 approaches for the estimates of interested parameters, one needs to compute the moments of the response up to order four. Keeping in mind $Y_{ij}^2 = Y_{ij}$, the calculations involved are given by

$$
\begin{aligned}
E(Y_{ij}Y_{iu}) &= (1-\pi^{-})^2 + (1-\pi^{-})(\pi^{-} + \pi^{+} - 1)(\eta_{iu} + \eta_{ij}) + (\pi^{-} + \pi^{+} - 1)^2 E(T_{ij}T_{iu}), \\
E(Y_{iu}Y_{iv}Y_{il}) &= E\{\prod_{k=u,v,l} [1-\pi^{-} + (\pi^{-} + \pi^{+} - 1)T_{ik}]\} \\
&= (1-\pi^{-})^3 + (1-\pi^{-})^2(\pi^{-} + \pi^{+} - 1)(\eta_{iu} + \eta_{iv} + \eta_{il}) \\
&\quad + (1-\pi^{-})(\pi^{-} + \pi^{+} - 1)^2[E(T_{iu}T_{iv}) + E(T_{iu}T_{il}) + E(T_{iv}T_{il})] \\
&\quad + (\pi^{-} + \pi^{+} - 1)^3 E(T_{iu}T_{iv}T_{il}),
\end{aligned}
$$

and

$$
\begin{aligned}
& E(Y_{ij}Y_{iu}Y_{iv}Y_{il}) \\
= \; & E\{\prod_{k=j,u,v,l}[1-\pi^- - (1-\pi^- - \pi^+)T_{ik}]\} \\
= \; & (1-\pi^-)^4 - (1-\pi^-)^3(1-\pi^- - \pi^+)(\eta_{itj} + \eta_{iu} + \eta_{iv} + \eta_{il}) \\
& +(1-\pi^-)^2(\pi^- + \pi^+ - 1)^2 E(T_{ij}T_{iu} + T_{ij}T_{iv} + T_{ij}T_{il} + T_{iu}T_{iv} + T_{iu}T_{il} + T_{iv}T_{il}) \\
& +(1-\pi^-)(\pi^- + \pi^+ - 1)^3 E(T_{iu}T_{iv}T_{il} + T_{ij}T_{iv}T_{il} + T_{iu}T_{ij}T_{il} + T_{iu}T_{iv}T_{ij}) \\
& +(\pi^- + \pi^+ - 1)^4 E(T_{ij}T_{iu}T_{iv}T_{il}).
\end{aligned}
$$

It is clear that these moments of the observed response Y are linear combinations of the 1st to 4th order moments of the true response T. The corresponding quantities about the true response T involved in the formulas can be calculated under the assumed model (3.28) of T in Section 3.1.2.

The following covariances under model (4.8) are needed for the parameter estimation in the GQL2 (it is OGQL under the assumed model (3.28) of T) framework. They are given by

$$
\begin{aligned}
Cov(Y_{ij}, Y_{iu}Y_{iv}) \; = \; & (\pi^- + \pi^+ - 1)^2(1-\pi^-)Cov(T_{ij}, T_{iu} + T_{iv}) \\
& +(\pi^- + \pi^+ - 1)^3 Cov(T_{ij}, T_{iu}T_{iv}) \\
Cov(Y_{ij}Y_{iu}, Y_{iv}Y_{il}) \; = \; & (\pi^- + \pi^+ - 1)^4 Cov(T_{ij}T_{iu}, T_{iv}T_{il}) \\
& +(\pi^- + \pi^+ - 1)^3(1-\pi^-)[Cov(T_{ij} + T_{iu}, T_{iv}T_{il}) \\
& +(\pi^- + \pi^+ - 1)^2(1-\pi^-)^2 Cov(T_{ij} + T_{iu}, T_{iv} \\
& +T_{il}) + (1-\pi^-)^4 Cov(T_{ij}T_{iu}, T_{iv} + T_{il})]
\end{aligned}
$$

In addition, since $Y_{ij}^2 = Y_{ij}$, we have

$$\begin{aligned} Cov(Y_{ij}, Y_{iu}Y_{ij}) &= E(Y_{ij}Y_{iu})(1 - E(Y_{ij})), \\ Cov(Y_{ij}Y_{iv}, Y_{iu}Y_{ij}) &= E(Y_{ij}Y_{iu}Y_{iv}) - E(Y_{ij}Y_{iv})E(Y_{ij}Y_{iu}), \end{aligned}$$

which are linear combinations of the moments of the true responses.

Note that if we have a perfect specificity $\pi^- = 1$, that is, the completely exact diagnoses among the healthy population, then the 1st to 4th moments given above can be greatly simplified as follows:

$$\begin{aligned} \mu_{ij} &= \pi^+ \eta_{ij}, \\ E(Y_{ij}Y_{iu}) &= (\pi^+)^2 E(T_{ij}T_{iu}), \\ E(Y_{ij}Y_{iu}Y_{iv}) &= (\pi^+)^3 E(T_{ij}T_{iu}T_{iv}), \\ E(Y_{ij}Y_{iu}Y_{iv}Y_{il}) &= (\pi^+)^4 E(T_{ij}T_{iu}T_{iv}T_{il}). \end{aligned}$$

It follows that the covariances can be simplified by

$$\begin{aligned} Cov(Y_{ij}, Y_{iu}) &= (\pi^+)^2 Cov(T_{ij}, T_{iu}), \\ Cov(Y_{ij}, Y_{iu}Y_{ij}) &= (\pi^+)^3 Cov(T_{ij}, T_{iu}T_{ij}), \\ Cov(Y_{ij}, Y_{iu}Y_{iv}) &= (\pi^+)^3 Cov(T_{ij}, T_{iu}T_{iv}), \\ Cov(Y_{ij}Y_{iu}, Y_{iv}Y_{ij}) &= (\pi^+)^4 Cov(T_{ij}T_{iu}, T_{iv}T_{ij}), \\ Cov(Y_{ij}Y_{iu}, Y_{iv}Y_{il}) &= (\pi^+)^4 Cov(T_{ij}T_{iu}, T_{iv}T_{il}). \end{aligned}$$

In addition, the variance of Y_{ij} can be simplified as follows:

$$\begin{aligned} Var(Y_{ij}) &= \pi^+ \eta_{ij}(1 - \pi^+ \eta_{ij}) \\ &= (\pi^+)^2 Var(T_{ij}) + \pi^+(1 - \pi^+)\eta_{ij} \end{aligned}$$

Therefore, in case of perfect specificity, the estimating equations approaches become much simpler.

To be specific in developing estimation approaches, we assume that the true response T_{ij} follow the transition model (3.28). We rewrite it as

$$\begin{aligned} \eta_{ij}^{c} &= P(T_{ij} = 1 | T_{i,j-1} = t_{i,j-1}) \\ &= \frac{exp(x_{ij}'\beta + t_{i,j-1}\gamma)}{1 + exp(x_{ij}'\beta + t_{i,j-1}\gamma)}, \text{ for j=1,2,}\ldots\text{, m,} \end{aligned} \tag{4.9}$$

The 1st-4th moments of T under this model can be found in Section 3.1.2.

4.2.2 Estimation of the model effects

To develop the methodology, we assume that the sensitivity π^{+} and the specificity π^{-} are known, mainly for the simplicity of the derivation. In fact, in epidemiological studies, even if π^{+} and π^{-} are unknown, some reasonable estimates can be obtained from previously similar studies or from independent validation studies of the classification scheme, or from some more exact clinic examination of a relatively small sample [Roy, Banerjee and Maiti (2005)]. We further assume that the true response T_{ij} follows the transition model (4.9) with baseline observations $t_{i0} = 0$.

4.2.2.1 GQL method

When the sensitivity π^{+} and the specificity π^{-} are known, the parameters of interest are $\theta = (\beta', \gamma)'$ from the transition model (4.9), where $\beta = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients, and γ is the dynamic dependence parameter.

To estimate the model parameters based on the GQL method, we solve the following estimating equations

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{-1}(y_i - \mu_i) = 0, \tag{4.10}$$

where $y_i = (y_{i1}, \ldots, y_{iJ})'$, $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ})'$, and $\partial \mu_i / \partial \theta$ is the first order derivative matrix of μ_i with respect to θ of dimension $(p+1) \times J$ for known π^+ and π^-. The jkth element of $\partial \mu_i / \partial \theta$ is given by

$$\begin{aligned} \frac{\partial \mu_{ij}}{\partial \beta_k} &= (\pi^- + \pi^+ - 1)\{\tilde{\eta}_{ij}(1 - \tilde{\eta}_{ij})\eta_{i,j-1} + \tilde{\eta}_{ij}(1 - \tilde{\eta}_{ij})(1 - \eta_{i,j-1})\} x_{ij(k)} \\ &\quad + (\tilde{\eta}_{ij} - \bar{\eta}_{ij}) \frac{\partial \mu_{i,j-1}}{\partial \beta_k} \end{aligned} \tag{4.11}$$

for $k = 1, \ldots, p$, $j = 1, \ldots, J$, and

$$\frac{\partial \mu_{ij}}{\partial \gamma} = (\pi^- + \pi^+ - 1) \sum_{u=1}^{j} \eta_{iu-1} \bar{\eta}_{iu}(1 - \tilde{\eta}_{iu}) \prod_{k=u+1}^{j} (\tilde{\eta}_{ik} - \bar{\eta}_{ik}), \tag{4.12}$$

for $j = 2, \ldots, J$ and $\partial \eta_{i1} / \partial \gamma = 0 \Rightarrow \partial \mu_{i1} / \partial \gamma = 0$. Σ_i is the variance-covariance matrix of Y_i. If Σ_i is replaced with a general "working" covariance matrix W_i in the case that Σ_i is unknown, the GQL approach becomes the GEE approach [Liang and Zeger (1986)].

Once we have the estimate $\hat{\theta}_{GQL}$, the corresponding consistent estimate of the covariance matrix of $\hat{\theta}_{GQL}$ is given by

$$\hat{V}(\hat{\theta}_{GQL}) = \left(\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \theta} \right)^{-1} |_{\theta = \hat{\theta}_{GQL}}. \tag{4.13}$$

4.2.2.2 Maximum likelihood method

In this subsection, we develop the maximum likelihood (ML) approach under regularity conditions, such as imperfect sensitivity and specificity, that is $0 < \pi^+, \pi^- < 1$.

The complete likelihood function given observations $Y = y$ and $T = t$ is formulated by

$$
\begin{aligned}
L(\theta|y,t) &= \prod_{i=1}^{I} f(y_i, t_i) nonumber & (4.14)\\
&= \prod_{i=1}^{I} f(y_i|t_i) f(t_i)\\
&= \prod_{i=1}^{I}\prod_{j=1}^{J} f(y_{ij}|t_{ij}) \prod_{i=1}^{I}\prod_{j=1}^{J} g_{i,j|j-1}, & (4.15)
\end{aligned}
$$

where $g_{ij|j-1}$ are given in Section 3.1.2, and

$$
\begin{aligned}
f(y_{ij}|t_{ij}) &= [(\pi^+)^{y_{ij}}(1-\pi^+)^{1-y_{ij}}]^{t_{ij}}[(\pi^-)^{1-y_{ij}}(1-\pi^-)^{y_{ij}}]^{1-t_{ij}}\\
&= (\pi^+)^{y_{ij}t_{ij}}(1-\pi^+)^{(1-y_{ij})t_{ij}}(\pi^-)^{(1-y_{ij})(1-t_{ij})}(1-\pi^-)^{(1-t_{ij})y_{ij}}. \quad (4.16)
\end{aligned}
$$

The log-likelihood function can be expressed as

$$
\begin{aligned}
\ell(\theta;t,y) &= \sum_{i=1}^{I} \ell_i(\theta;t_i,y_i)\\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}(x_{ij}'\beta + \gamma t_{i,j-1}) - \sum_{i=1}^{I}\sum_{j=1}^{J} log[1 + exp(x_{ij}'\beta + \gamma t_{i,j-1})]\\
&+ \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}(1-t_{ij}) log(1-\pi^-) + \sum_{i=1}^{I}\sum_{j=1}^{J}(1-y_{ij}) t_{ij} log(1-\pi^+)\\
&+ \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} t_{ij} log\pi^+ + \sum_{i=1}^{I}\sum_{j=1}^{J}(1-y_{ij})(1-t_{ij}) log\pi^-. \quad (4.17)
\end{aligned}
$$

In $\ell(\theta;t,y)$, the values $t_{ij}, i = 1,\ldots,I,\ j = 1,\ldots,J$ are not observable. We, therefore, apply the EM algorithm to find the ML estimates of the model parameters. Given an initial value $\theta^{(0)}$, we denote $\theta^{(k)}$ as the estimate of θ at the kth iteration. The $(k+1)$th iteration of the EM algorithm can be derived as follows.

E-step: Find the expected complete-data log-likelihood function if θ were $\theta^{(k)}$:

$$
\begin{aligned}
& Q(\theta|\theta^{(k)}, y) \\
= \; & E[\ell(\theta; T, Y)|Y = y, \theta^{(k)}] \\
= \; & \sum_{i=1}^{I}\sum_{j=1}^{J}[t_{ij}^{(k+1)} x_{ij}'\beta + \gamma(t_{ij}t_{i,j-1})^{(k+1)}] \\
& -\sum_{i=1}^{I}\sum_{j=1}^{J}\{log[1+exp(x_{ij}'\beta+\gamma)]t_{i,j-1}^{(k+1)} + log[1+exp(x_{ij}'\beta)](1-t_{i,j-1}^{(k+1)})\} \\
& +\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}t_{ij}^{(k+1)}log\pi^{+} + \sum_{i=1}^{I}\sum_{j=1}^{J}(1-y_{ij})(1-t_{ij}^{(k+1)})log\pi^{-} \\
& +\sum_{i=1}^{I}\sum_{j=1}^{I} y_{ij}(1-t_{ij}^{(k+1)})log(1-\pi^{-}) + \sum_{i=1}^{I}\sum_{j=1}^{J}(1-y_{ij})t_{ij}^{(k+1)}log(1-\pi^{+}), \quad (4.18)
\end{aligned}
$$

where

$$
t_{ij}^{(k+1)} = E_{\hat{\theta}^{(k)}}(T_{ij}|Y_i = y_i) = \frac{P_{\theta^{(k)}}(Y_i = y_i, T_{ij} = 1)}{P_{\theta^{(k)}}(Y_i = y_i)} \quad (4.19)
$$

and

$$
(t_{ij}t_{i,j-1})^{(k+1)} = E_{\theta^{(k)}}(T_{ij}T_{i,j-1}|Y_i = y_i) = \frac{P_{\theta^{(k)}}(Y_i = y_i, T_{ij} = 1, T_{i,j-1} = 1)}{P_{\theta^{(k)}}(Y_i = y_i)}. \quad (4.20)
$$

Let $\Omega^{\#} = \{1, 2, \ldots, m\}$, $\Omega_j^{\#} = \Omega^{\#} \setminus \{j\}$ and $\Omega_{j,j-1} = \Omega^{\#} \setminus \{j, j-1\}$. For given i, we denote $A = \{k : t_{ik} = 1, k \in \Omega^{\#}\}$, $A_j = \{k : t_{ik} = 1, k \in \Omega_j^{\#}\}$ and $A_{j,j-1} = \{k : t_{ik} = 1, k \in \Omega_{j,j-1}^{\#}\}$, where t_i is the value of T_i, the probabilities involved in (4.18) and (4.19) can be calculated as

$$
\begin{aligned}
& P(Y_i = y_i) \\
= \; & \sum_{A \in 2^{\Omega^{\#}}} P(Y_i = y_i|T_i = t_i)P(T_{i1} = t_{i1}) \prod_{j=2}^{J} P(T_{ij} = t_{ij}|T_{i,j-1} = t_{i,j-1})P(T_{i,j-1} = t_{i,j-1}) \\
= \; & \sum_{A \in 2^{\Omega^{\#}}} (\pi^{+})^{\sum_{k\in A} y_{ik}} (1-\pi^{+})^{|A| - \sum_{k\in A} y_{ik}} (\pi^{-})^{J - \sum_{k\in\Omega^{\#}\setminus A} y_{ik}} (1-\pi^{-})^{\sum_{k\in\Omega^{\#}\setminus A} y_{ik}} \cdot \\
& \prod_{j\in\Omega^{\#}\setminus \bar{A}} (1-\bar{\eta}_{ij}) \prod_{k\in A} (1-\bar{\eta}_{i,k+1})^{1-I_{\{k+1\in A\}}} \left(\frac{\bar{\eta}_{i,k+1}}{\bar{\eta}_{i,k+1}}\right)^{I_{\{k+1\in A\}}} \bar{\eta}_{ik},
\end{aligned}
$$

$$
\begin{aligned}
& P(Y_i = y_i, T_{ij} = 1) \\
= & \sum_{A_j \in 2^{\Omega_j^{\#}}} P\{Y_i = y_i, (T_{i1}, \ldots, T_{i,j-1}, T_{ij}, T_{i,j+1}, \ldots, T_{iJ}) = (t_{i1}, \ldots, t_{i,j-1}, 1, t_{i,j+1}, \ldots, t_{iJ})\} \\
= & \sum_{A_j \in 2^{\Omega_j^{\#}}} (\pi^+)^{\sum\limits_{k\in A} y_{ik}} (1-\pi^+)^{|A| - \sum\limits_{k\in A} y_{ik}} (\pi^-)^{J - \sum\limits_{k\in\Omega^{\#}\setminus A} y_{ik}} (1-\pi^-)^{\sum\limits_{k\in\Omega^{\#}\setminus A} y_{ik}} \\
& \prod_{l\in\Omega^{\#}\setminus\bar{A}} (1-\bar{\eta}_{il}) \prod_{k\in A} (1-\tilde{\eta}_{i,k+1})^{1-I_{\{k+1\in A\}}} \left(\frac{\tilde{\eta}_{i,k+1}}{\bar{\eta}_{i,k+1}}\right)^{I_{\{k+1\in A\}}} \bar{\eta}_{ik},
\end{aligned}
$$

and

$$
\begin{aligned}
& P(Y_i = y_i, T_{ij} = 1, T_{i,j-1} = 1) \\
= & \sum_{A_{j,j-1} \in 2^{\Omega_{j,j-1}^{\#}}} P\{Y_i = y_i, (T_{i1}, \ldots, T_{i,j-1}, T_{ij}, T_{i,j+1}, \ldots, T_{iJ}) = (t_{i1}, \ldots, 1, 1, t_{i,j+1}, \ldots, t_{iJ})\} \\
= & \sum_{A_{j,j-1} \in 2^{\Omega_{j,j-1}^{\#}}} (\pi^+)^{\sum\limits_{k\in A} y_{ik}} (1-\pi^+)^{|A| - \sum\limits_{k\in A} y_{ik}} (\pi^-)^{J - \sum\limits_{k\in\Omega^{\#}\setminus A} y_{ik}} (1-\pi^-)^{\sum\limits_{k\in\Omega^{\#}\setminus A} y_{ik}} . \\
& \prod_{l\in\Omega^{\#}\setminus\bar{A}} (1-\bar{\eta}_{il}) \prod_{k\in A} (1-\tilde{\eta}_{i,k+1})^{1-I_{\{k+1\in A\}}} \left(\frac{\tilde{\eta}_{i,k+1}}{\bar{\eta}_{i,k+1}}\right)^{I_{\{k+1\in A\}}} \bar{\eta}_{ik},
\end{aligned}
$$

where $|A|$ denotes the size of A, $\bar{A} = A \cup \{k+1 : k \in A\}$, and $I_{\{k+1\in A\}} = 1$ if $k+1 \in A$, 0, otherwise. $2^{\Omega^{\#}} = \{B : B \subseteq \Omega^{\#}\}$ consisting of all subsets of $\Omega^{\#}$.

M-step: Determine $\theta^{(k+1)}$ by maximizing the expected log-likelihood. To do so, we solve equations:

$$
S^*(\theta) = \frac{\partial Q}{\partial \theta} = \begin{pmatrix} \frac{\partial Q(\theta|\theta^{(k)}, y)}{\partial \beta} \\ \frac{\partial Q(\theta|\theta^{(k)}, y)}{\partial \gamma} \end{pmatrix} = 0,
$$

where

$$
\frac{\partial Q(\theta|\theta^{(k)}, y)}{\partial \beta} = \sum_{i=1}^{I}\sum_{j=1}^{J} [t_{ij}^{(k+1)} - \bar{\eta}_{ij} - (\tilde{\eta}_{ij} - \bar{\eta}_{ij}) t_{i,j-1}^{(k+1)}] x_{ij}, \tag{4.21}
$$

$$
\frac{\partial Q(\theta|\theta^{(k)}, y)}{\partial \gamma} = \sum_{i=1}^{I}\sum_{j=1}^{J} [(t_{ij} t_{i,j-1})^{(k+1)} - \tilde{\eta}_{ij} t_{i,j-1}^{(k+1)}]. \tag{4.22}
$$

Once the ML estimates $\hat{\theta}_{ML}$ are achieved from the EM algorithm, we can estimate the variance-covariance matrix of the estimator $\hat{\theta}_{ML}$ by the inverse of the observed Fisher information matrix I_Y. Denote the first and second order derivatives of the log-likelihood functions $\ell(\theta;t,y)$ and $\ell_i(\theta;t_i,y_i)$ as $S(\theta;t,y) = \frac{\partial \ell(\theta;t,y)}{\partial \theta}$, $B(\theta;t,y) = -\frac{\partial^2 \ell(\theta;t,y)}{\partial \theta \partial \theta'}$, $S_i(\theta;t_i,y_i) = \frac{\partial \ell_i(\theta;t_i,y_i)}{\partial \theta}$, and $B_i(\theta;t_i,y_i) = -\frac{\partial^2 \ell_i(\theta;t_i,y_i)}{\partial \theta \partial \theta'}$, respectively. Then we have $S(\theta;t,y) = \sum_{i=1}^{I} S_i(\theta;t_i,y_i)$ and $B(\theta;t,y) = \sum_{i=1}^{I} B_i(\theta;t_i,y_i)$. Louis (1982) suggested a formula for the observed information

$$
\begin{aligned}
I_Y &= E_{\hat{\theta}}[B(\theta;Y,T)|y] - E_{\hat{\theta}}[S(\theta;T,Y)S'(\theta;T,Y)|y] + E_{\hat{\theta}}[S(\theta;T,Y)|y]E_{\hat{\theta}}[S(\theta;T,Y)|y] \\
&= E_{\hat{\theta}}[B(\theta;Y,T)|y] - \sum_{i=1}^{I} E_{\hat{\theta}}[S_i(\theta;T_i,Y_i)S_i'(\theta;T_i,Y_i)|y_i] \\
&\quad + \sum_{i=1}^{I} E_{\hat{\theta}}[S_i(\theta;T_i,Y_i)|y_i]E_{\hat{\theta}}[S_i'(\theta;T_i,Y_i)|y_i].
\end{aligned}
$$

For multinomial variables, I_Y reduces to

$$
\sum_{i=1}^{I} E_{\hat{\theta}}[S_i(\theta;T_i,Y_i)|y_i]E_{\hat{\theta}}[S_i'(\theta;T_i,Y_i)|y_i] \tag{4.23}
$$

due to the fact that

$$
E[B_i(\theta;Y_i,T_i)] = E[S_i(\theta;T_i,Y_i)S_i'(\theta;T_i,Y_i)]. \tag{4.24}
$$

However, simulations show that estimated standard errors derived from formula (4.23) tend to underestimate the true standard errors the MLE's based EM algorithm. It may be because the difference between the observed values of the two sides of equation (4.24) is not ignorable. We next introduce a corrected estimator of the Fisher information based on function (4.17). From the full likelihood function (4.14), it can be seen that the conditional density $f(y_i|t_i)$ do not contain the parameter of interest θ. This leads to $\frac{\partial Q(\theta|y)}{\partial \theta} = S(\theta;t,y)$, hence $\frac{\partial^2 Q(\theta|y)}{\partial \theta \partial \theta'} = B(\theta;t,y)$. We define the complete information matrix as the expected complete-data log-likelihood function (4.17), that is,

$$
\begin{aligned}
I_{T,Y} &= -E\left(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\begin{bmatrix} x_{ij}[\tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})\eta_{i,j-1}+\tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})(1-\eta_{i,j-1})]x'_{ij} & \tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})\eta_{i,j-1}x_{ij} \\ \tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})\eta_{i,j-1}x'_{ij} & \tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})\eta_{i,j-1} \end{bmatrix},
\end{aligned}
$$

which can be evaluated at the final estimate $\hat{\theta}_{ML}$. Furthermore, we define another quantity as

$$
\begin{aligned}
I_Q &= -E\left(\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right) \\
&= \sum_{i=1}^{I}\sum_{t=1}^{T}\begin{bmatrix} x_{ij}[\tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})(1-t_{i,j-1}^{(k+1)})+\tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})t_{i,j-1}^{(k+1)}]x'_{ij} & \tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})t_{i,j-1}^{(k+1)}x_{ij} \\ \tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})t_{i,j-1}^{(k+1)}x'_{ij} & \tilde{\eta}_{ij}(1-\tilde{\eta}_{ij})t_{i,j-1}^{(k+1)} \end{bmatrix}.
\end{aligned}
$$

In EM algorithm, I_Q can be estimated by employing $t_{ij}^{(k+1)}$ from the last E-step and final $\hat{\theta}$. Let S_i^* be the analogy of S^* for subject i, which is actually equal to $S_i(\hat{\theta}_{ML}; t_i, y_i)$ by replacing t_{ij} and $t_{ij}t_{i,j-1}$ with $t_{ij}^{(k+1)}$ and $(t_{ij}t_{i,j-1})^{(k+1)}$ from the last E-step, respectively. Our simulations reveal that the estimated $I_{T,Y}$ and I_Q are very close. However their difference is significant far from 0, which can be used to correct the bias of Louis estimator (4.23). Therefore, our new estimate of the observed information is given by

$$
I_Y^c = (I_{T,Y} - I_Q + \sum_{i=1}^{I} S_i^* S_i^*)|_{\hat{\theta}_{ML}}. \tag{4.25}
$$

Therefore, the estimate of $\hat{\theta}_{ML}$ can be given by

$$
\widehat{\mathrm{Var}}(\hat{\theta}_{ML}) = (I_Y^c)^{-1}. \tag{4.26}
$$

It is shown from the simulation in the next section that the new estimator (4.26) can consistently estimate true $V(\hat{\theta}_{ML})$ in EM algorithm. In addition, this new estimator is in good concordance with $\hat{V}(\hat{\theta}_{OGQL})$ which will be given in the Section 4.2.2.3.

Remark:

1. Suppose that we have the observations of $T_{ij} = t_{ij}$ and $Y_{ij} = y_{ij}$. As mentioned in Section 4.1, if we have perfect specificity, then $(1-t_{ij})y_{ij} = 0$ and $1-\pi^- = 0$. And we define that $0^0 = 1$. The conditional likelihood part in the full likelihood function then can be modified to accommodate this case, and it is given by

$$\begin{aligned} f(y_{ij}|t_{ij}) &= (\pi^+)^{y_{ij}t_{ij}}(1-\pi^+)^{(1-y_{ij})t_{ij}}(\pi^-)^{(1-y_{ij})(1-t_{ij})}(1-\pi^-)^{(1-t_{ij})y_{ij}} \\ &= (\pi^+)^{y_{ij}t_{ij}}(1-\pi^+)^{(1-y_{ij})t_{ij}} \end{aligned} \quad (4.27)$$

Then the corresponding log-likelihood function can be expressed as

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{I}\sum_{j=1}^{J} t_{ij}(x'_{ij}\beta + \gamma t_{i,j-1}) - \sum_{i=1}^{I}\sum_{j=1}^{J} log[1 + exp(x'_{ij}\beta + \gamma t_{i,j-1})] \\ &+ \sum_{i=1}^{I}\sum_{j=1}^{I} y_{ij}t_{ij}log\pi^+ + \sum_{i=1}^{I}\sum_{j=1}^{I}(1-y_{ij})t_{ij}log(1-\pi^+) \end{aligned} \quad (4.28)$$

The EM algorithm can be developed based on the log-likelihood function (4.28).

2. In the case of perfect sensitivity, a similar EM procedure can be developed.

3. The GQL approach developed in Section 4.2.2.1 automatically accommodates these two special cases.

4.2.2.3 GQL2 (OGQL) method

Here we develop the GQL2 approach, that is, the optimal GQL (OGQL) approach in Sutradhar and Farrell (2007) to estimate the model parameters under the first-order logistic transition model (4.9) for dynamic binary data. This approach can also accommodate the case of perfect sensitivity or specificity. As mentioned by Sutradhar and Farrell (2007), under the dynamic dependence model (4.9), the estimating equations based on both the ML and OGQL approach involve the first and the second order response. They demonstrated that the estimators obtained by OGQL are almost as efficient as the ML estimators but with fewer assumptions on the model. However, their conclusion was reached under the assumption that there is no misclassification in the data.

In this section, we develop the GQL2 approach by exploiting the first and second order statistics of the binary responses with misclassification being taken into account. It will be shown that the GQL2 method behaves similar to the maximum likelihood approach in simulation. Therefore, in presence of misclassification, the GQL2 approach tends to be also the OGQL approach under the model (4.9). And the OGQL is much simpler as far as methodology development and calculations are concerned.

Let f_i be the observation of $F_i = (Y_i', S_i')'$, where $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$, and $S_i = (Y_{i1}Y_{i2}, \ldots, Y_{iu}Y_{iv}, \ldots, Y_{i,J-1}Y_{iJ})'$, then $\delta_i = E(F_i) = (\mu_i', \nu_i')'$, where $\nu_i = E(S_i)$. Further, let

$$\Omega_i = \begin{pmatrix} Cov(Y_i) & Cov(Y_i, S_i) \\ Cov(S_i, Y_i) & Cov(S_i) \end{pmatrix}$$

be the $J(J+1)/2 \times J(J+1)/2$ covariance matrix. The estimating equations take the same form as those in [Sutradhar and Farrell (2007)]. But the meaning of the components in our setting is different from the ones in [Sutradhar and Farrell (2007)].

Our OGQL approach is given as

$$\sum_{i=1}^{I} \frac{\partial \delta_i'}{\partial \theta} \Omega_i^{-1} (f_i - \delta_i) = 0 \tag{4.29}$$

The quantities required in this equation such as μ_i, ν_i, $Cov(Y_i)$, $Cov(S_i, Y_i)$ and $Cov(S_i)$ can be calculated based on the moments given in section 4.2.1.

Now, let ζ_i denote the expected value of the true response S_i, with elements $\zeta_{iuv} = E(T_{iu}T_{iv})$. The elements of the first order derivative of δ_i with respect to θ are formulated by

$$\frac{\partial \delta_{iuv}}{\partial \theta} = (1 - \pi^+ - \pi^-)^2 \frac{\partial \zeta_{iuv}}{\partial \theta} - (1 - \pi^-)(1 - \pi^- - \pi^+)(\frac{\partial \eta_{iu}}{\partial \theta} + \frac{\partial \eta_{iv}}{\partial \theta}),$$

where $\frac{\partial \eta_{iu}}{\partial \theta}$'s are given by expressions (2.2) and (2.3) in the paper [Sutradhar and Farrell (2007)]. $\frac{\partial \zeta_{iuv}}{\partial \theta}$'s are given in the following equations

$$\frac{\partial \zeta_{iuv}}{\partial \beta} = \sum_{t_{iu}, t_{iv} \notin S^*} [g_{i1}^* \prod_{j=2}^{m} g_{i,j|j-1}^* \{\sum_{k=2}^{m} (t_{ik} - \lambda_{ik|k-1}^*) x_{ik} + (t_{i1} - \eta_{i1}) x_{i1}\}]_{t_{iu}=1, t_{iv}=1}, \tag{4.30}$$

$$\frac{\partial \zeta_{iuv}}{\partial \gamma} = \sum_{t_{iu}, t_{iv} \notin S^*} [g_{i1}^* \prod_{j=2}^{m} g_{i,j|j-1}^* \{\sum_{k=2}^{m} (t_{ik} - \lambda_{ik|k-1}^*) t_{ik}\}]_{t_{iu}=1, t_{iv}=1} \tag{4.31}$$

[Sutradhar and Farrell (2007)].

Once the estimate $\hat{\theta}_{OGQL}$ is obtained, the corresponding covariance matrix $V(\hat{\theta}_{OGQL})$ can be consistently estimated by

$$\hat{V}(\hat{\theta}_{OGQL}) = \left(\sum_{i=1}^{I} \frac{\partial \delta_i'}{\partial \theta} \Omega_i^{-1} \frac{\partial \delta_i}{\partial \theta} \right)^{-1} |_{\theta = \hat{\theta}_{OGQL}}. \tag{4.32}$$

4.2.3 Simulation Studies

As those in some other estimating-equations-based approaches, the estimators proposed in this thesis are all consistent and are asymptotically normally distributed.

In this subsection, we study the properties of the proposed estimators, namely, the GQL, OGQL and ML estimators, through Monte Carlo simulation. Random samples with a size similar to that of a subset of the H6SC study were generated based on different designs of covariates. Model parameters are then estimated by using the three proposed approaches.

4.2.3.1 Covariate designs

In the simulation study, we consider $I = 560$ independent subjects each with $J = 4$ repeated observations. The true data t_{ij}, $i = 1, 2, \ldots, 560$ and $j = 1, 2, 3, 4$, are generated following the dynamic model (4.9), and the observed data y_{ij} are generated following the misclassification model (4.8). To be specific, y_{ij} can be generated following the procedure described below.

(1) Once we have t_{ij}, we can firstly generate a binomial variable U from $binomial(t_{ij}, \pi^+)$.

(2) Secondly, we generate another binomial variable V from $binomial(1-t_{ij}, 1-\pi^-)$.

(3) Lastly, we get $y_{ij} = U + V$.

Sutradhar and Farrell (2007) conducted simulations for error-free binary data based on the following three covariate designs. In our simulation studies in this section, we also consider these three designs for the analysis of mis-measured longitudinal binary data. The three designs are given by:

Design 1: $x_{ij(1)} = 1$ and $x_{ij(2)} = j/4$ for $i = 1, \ldots, 560$ and $j = 1, \ldots, 4$.

Design 2: $x_{ij(1)} = 1,\ (j = 1, 2);\ x_{ij(1)} = 0,\ (j = 3, 4),\ i = 1, \ldots, 140;$

$x_{ij(1)} = 1,\ j = 1, \ldots, 4,\ i = 141, \ldots, 420;$

$x_{ij(1)} = 0,\ (j = 1, 2);\ x_{ij(1)} = 1,\ (j = 3, 4),\ i = 421, \ldots, 560.$

$x_{it(2)} = j/4,\ j = 1, \ldots, 4,\ i = 1, \ldots, 560.$

Design 3: $x_{ij(1)} = 1,\ (j = 1, 2);\ x_{ij(1)} = 0,\ (j = 3, 4),\ i = 1, \ldots, 140,$

$x_{ij(1)} = 1,\ j = 1, \ldots, 4,\ i = 141, \ldots, 420,$

$x_{ij(1)} = 1,\ (j = 1, 2);\ x_{ij(1)} = -1,\ (j = 3, 4),\ i = 421, \ldots, 560.$

$x_{ij(2)} = -0.5,\ (j = 1, 2);\ x_{ij(2)} = 0.5,\ (j = 3, 4),\ i = 1, \ldots, 140;$

$x_{ij(2)} = j/4,\ j = 1, \ldots, 4,\ i = 141, \ldots, 420;$

$x_{ij(2)} = 0.5(j - 2),\ j = 1, \ldots, 4,\ i = 421, \ldots, 560.$

Simulation results corresponding to different settings of the covariate designs and the choices of parameter values are presented in the subsequent subsections in the form of tables.

4.2.3.2 Estimation of the model parameters

In this subsection, we carry out a simulation study to examine the finite sample behaviors of the GQL, OGQL, and ML approaches discussed in the Section 4.2.2 for the estimates of the model parameters $\theta = (\beta', \gamma)'$ involved in the model (4.9) and (4.8). For this purpose, we first generate data $y_i = (y_{i1}, y_{i2}, \ldots, y_{i4})'$ for $i = 1, 2, \ldots, 560$ and $J = 4$ by model (4.9) and (4.8), in each of 1000 simulations, using the three covariate designs as mentioned in the previous subsection. As far as the model parameters are concerned, we have chosen

$\beta = (\beta_1, \beta_2)' = (1, 1)';$

$\gamma = 1.5, 0, \text{and} - 1;$

$(\pi^+, \pi^-) = (0.95, 0.90)$ and $(0.75, 0.80)$,

where $\gamma = 0$ implies dynamic independence. The higher values of sensitivity π^+ and the specificity π^-, that is, $(0.95, 0.90)$ in the simulation, means less classification error for an excellent diagnostic test, whereas the lower values of this two quantities, for example, (0.75, 0.80) in the simulation, implies more classification error for a poor diagnostic test.

The simulated means (SM), simulated standard errors (SSE), estimated standard errors (ESE), and coverage probabilities (CPr) of the 95% confidence intervals for the interested parameters β and γ are reported in Tables 4.2-4.4 when $(\pi^+, \pi^-) = (0.95, 0.90)$ under the covariates designs 1, 2, and 3, respectively. Similarly, the computed SM, SSE, ESE, CPr's of β and γ are reported in Tables 4.5-4.7 when $(\pi^+, \pi^-) = (0.75, 0.80)$ under the covariates designs 1, 2, and 3, respectively. It therefore follows that a confidence interval can be constructed by using $\hat{\theta}_i \pm z_{\frac{\alpha}{2}} ESE(\hat{\theta}_i)$ given a significant level α, where θ_i denotes the ith element of θ, $i = 1, 2, 3$.

Furthermore, in each of Tables 4.2-4.7, three different estimates are computed under each of the GQL, OGQL and ML approaches. Namely, they are the ideal estimates (1) when the observations of the error-free response T are used, the naive estimates (2) when the data $\{y_{ij}\}$ are used with ignoring misclassification, and the corrected estimates (3) taking classification error the data $\{y_{ij}\}$ into consideration. The same model (4.9) can be used for both the ideal estimates and the naive estimates, and the estimating equations under the GQL, OGQL, and ML approaches are given by equations (3.32), (3.38), and (3.35-3.36) in Section 3.1.2. The corrected estimates of θ are calculated based on the estimating equations under the three proposed estimation approaches which are given in Section 4.2.2 in this chapter.

Obviously, the ideal estimates (1) can be expected to perform the best when observations of the latent responses are available. However, this is not always the case in practice. Therefore, it is our main interest in this subsection to examine the performance of the naive estimates and the corrected estimates under the three estimation approaches.

First, we check the performance of the GQL, OGQL, and ML estimates under the ideal situation, where the latent responses are assumed to be directly observable. It is clear from Tables 4.2-4.7 that all of the three approaches yield ignorable biases on the estimates of $\theta = (\beta', \gamma)'$. The OGQL method seems to produce almost the same results as the ML method does. The SSE's under the GQL method are larger than those under the OGQL and ML methods. However, generally speaking, these three approaches are highly competitive, as far as the closeness of the ESE's to the corresponding SSE's, and the CPr's are concerned. Similar conclusions are reached by Sutradhar and Farrell (2007). For instance, the case when $(\pi^+, \pi^-) = (0.95, 0.90)$, $\gamma = -1.0$ under covariate design 1 is a good evidence of this conclusion. However, the performance of the ideal estimates is not of our main interest, due to the limited application of the ideal estimates in practice.

One of our main objectives is to compare the performance of the naive estimates under the three estimation approaches. Generally, all of the three approaches produce highly biased estimates through the three covariate designs, and the situation get worse when (π^+, π^-) take low values, i.e. $(\pi^+, \pi^-) = (0.75, 0.80)$. To be specific, the naive estimates of $\theta = (\beta', \gamma)'$, when $(\pi^+, \pi^-) = (0.75, 0.80)$, are greatly underestimated. In addition, the computed coverage probabilities are basically lower than the expected value 0.95 due to the bias of the estimates of the parameters. For example, when $(\pi^+, \pi^-) = (0.75, 0.80)$, the CPr's are all nearly 0.000 in the case that $\gamma = 1.5$. It is, however, interesting to see that from Tables 4.2-4.7, the GQL method does not necessarily perform worse than the other two methods. In some cases, the GQL method even performs better than the OGQL and ML methods. For example, under covariate design 2, when $(\pi^+, \pi^-) = (0.95, 0.90)$ (Table 4.3), for $\theta = (1, 1, -1)'$ and $\gamma = -1$, we get $\hat{\theta}_{GQL} = (0.879, 0.973, -0.874)'$, along with the respective CPr's given by $(0.725, 0.964, 0.938)$, we also get $\hat{\theta}_{OGQL} = (0.862, 0.853, -0.725)'$ with a corresponding CPr vector given by $(0.653, 0.808, 0.292)$. In this example, the GQL estimates are less biased than the OGQL estimates, and the corresponding CPr's are closer to 0.95 than those under the OGQL method, which may be due to the less error information used by naive GQL approach compared with the naive OGQL and ML approaches. The naive OGQL estimates perform similarly to the naive ML estimates, like in the ideal case. It can be understood that the classification errors plays an important role in the process. The estimates are all attenuated by it.

Now, we move on to examine the performance of the corrected estimates under the GQL, OGQL, and ML approaches. Form Tables 4.2-4.7, we can see that all of the three approaches produce approximately unbiased estimates, with coverage probabilities almost equal to the nominal level of 0.95. The corrected OGQL and ML approaches tend to gain more efficiency than the corrected GQL approach because the OGQL and ML methods use more information from the data than GQL does. This can be demonstrated by the smaller SSE's under the OGQL and ML methods than those under the GQL. Furthermore, it continues to be seen that the OGQL method yields efficient estimation results almost identical to the MLE's. This is demonstrated by the results in the case that $\gamma = 0$ when $(\pi^+, \pi^-) = (0.95, 0.90)$ under design 3 in Table 4.4. The corrected GQL estimates $\hat{\theta}_{GQL} = (1.004, 1.003, -1.011)'$ with CPr's $(0.955, 0.954, 0.956)$ and the corrected OGQL estimates $\hat{\theta}_{OGQL} = (1.004, 1.004, -1.012)'$ with CPr's $(0.944, 0.953, 0.952)$ are all very close to the true value $\theta = (1, 1, -1)'$ with the true nominal level 0.95, whereas the SSE's of $\hat{\theta}_{GQL}$ is given by $(0.091, 0.145, 0.191)'$, and the SSE's of $\hat{\theta}_{OGQL}$ are given by $(0.080, 0.121, 0.135)'$, it can be found that the SSE's under the OGQL method are smaller than the SSE's under the GQL method.

At last but not least, we compare the ideal estimates with the corresponding corrected estimates under the three estimation approaches. It can be seen that both of the ideal estimates and corrected estimates are almost unbiased, and their coverage probabilities are near 0.95. However, the SSE's of the ideal estimates appear to be slightly smaller than the corrected estimates. All of this seems quite reasonable, because the ideal estimates are supposed to be the best among the three kinds of estimates. However, the corrected estimates are almost as efficient as the ideal ones, and they are more practical since misclassification errors may often happen in most of cases.

Now, before we reach our final conclusion, there are a few important points to be mentioned.

1. The OGQL and ML approaches perform almost identically in most of cases as far as the unbiasedness, estimated standard errors, and coverage probabilities of the 0.95 confidence intervals are concerned. This is because the OGQL approach utilizes as much information as the ML approach does, that is, both approaches use the first and second order statistics of the responses in their estimating equations. Because of this similarity of theoretic development, these two approaches are found

to perform as poor as each other in the naive estimation. This further explains that the OGQL and ML approaches may be surpassed by the GQL method in some cases of the naive estimation because the GQL approach uses only the first order statistics of the error-prone responses, hence less error information is used in the estimation.

2. For the ideal and naive estimates of the parameters $\theta = (\beta', \gamma)'$, the ordinary ML approach is applied, whereas for the corrected estimates under the ML approach, the EM algorithm is utilized. Accordingly, the covariance matrices of $Var(\hat{\theta}_{ML})$ for the ideal and naive estimates are consistently estimated by the inverse of the observed information matrices I_T^{-1} and I_Y^{-1}, respectively. For the corrected $\hat{\theta}_{ML}$ obtained through the EM algorithm, the corresponding covariance matrix is consistently estimated by (4.26).

In conclusion, the simulation results in Tables 4.2-4.7 show that the three proposed estimation approaches, namely, the GQL, OGQL, and ML approaches, are highly competitive in both situations where the latent responses are known, or the misclassification of responses is corrected, whereas the GQL approach appears to be slightly less efficient than the other two approaches, as far as the simulated standard errors (SSE's) are concerned. On the other hand, under the naive situation, where misclassification of responses is ignored, all three approaches are found to perform poorly, with the GQL approach being slightly better than the other two approaches. Throughout the simulations in this section, the OGQL and ML methods produce quite similar results, due to the similarity of their theoretic development. However, considering the complexity of the EM algorithm used for the corrected ML method, we recommend the use of the corrected OGQL estimates in practice.

Table 4.2: Simulation results under Design 1 with $(\pi^{+}, \pi^{-}) = (0.95, 0.90)$ and the true values of parameters $\beta = (1, 1)$

		GQL			OGQL			ML		
γ	Quantity	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	SM(β_1)	1.015	0.909	1.007	1.032	0.983	0.992	1.031	0.983	0.995
	SSE	0.171	0.164	0.314	0.160	0.144	0.251	0.160	0.144	0.250
	ESE	0.171	0.165	0.310	0.165	0.142	0.244	0.165	0.142	0.242
	CPr	0.954	0.174	0.940	0.963	0.850	0.944	0.962	0.854	0.946
	SM(β_2)	1.002	0.696	1.011	0.989	0.771	1.052	0.989	0.771	1.054
	SSE	0.464	0.349	0.702	0.427	0.323	0.732	0.427	0.323	0.631
	ESE	0.465	0.352	0.704	0.425	0.318	0.716	0.425	0.318	0.616
	CPr	0.952	0.492	0.950	0.958	0.807	0.948	0.958	0.805	0.945
	SM(γ)	1.505	1.402	1.512	1.582	0.852	1.501	1.582	0.850	1.501
	SSE	0.385	0.277	0.515	0.236	0.181	0.450	0.235	0.180	0.450
	ESE	0.379	0.275	0.511	0.240	0.177	0.437	0.239	0.177	0.435
	CPr	0.936	0.890	0.954	0.954	0.071	0.947	0.955	0.070	0.947
0	SM(β_1)	1.009	0.980	1.010	1.005	0.975	1.005	1.005	0.975	1.005
	SSE	0.136	0.126	0.164	0.130	0.125	0.161	0.130	0.125	0.161
	ESE	0.132	0.128	0.167	0.132	0.127	0.164	0.132	0.127	0.164
	CPr	0.956	0.952	0.948	0.957	0.945	0.952	0.956	0.945	0.952
	SM(β_2)	0.989	0.727	0.986	1.007	0.760	1.015	1.007	0.760	1.015
	SSE	0.389	0.350	0.486	0.273	0.241	0.360	0.273	0.241	0.359
	ESE	0.375	0.344	0.479	0.267	0.242	0.360	0.267	0.242	0.362
	CPr	0.942	0.869	0.954	0.956	0.827	0.953	0.955	0.824	0.952
	SM(γ)	0.008	0.0280	0.013	-0.009	-0.003	-0.013	-0.009	-0.003	-0.013
	SSE	0.285	0.275	0.355	0.153	0.137	0.213	0.153	0.137	0.213
	ESE	0.275	0.266	0.343	0.148	0.135	0.212	0.148	0.135	0.213
	CPr	0.954	0.951	0.951	0.943	0.948	0.954	0.943	0.947	0.952
-1	SM(β_1)	1.007	0.955	1.010	1.006	0.949	1.008	1.006	0.949	1.008
	SSE	0.121	0.121	0.151	0.120	0.118	0.150	0.120	0.118	0.150
	ESE	0.123	0.120	0.151	0.122	0.119	0.150	0.122	0.119	0.151
	CPr	0.954	0.926	0.960	0.953	0.925	0.957	0.953	0.925	0.956
	SM(β_2)	0.994	0.838	1.000	1.000	0.706	1.008	1.000	0.706	1.007
	SSE	0.269	0.269	0.325	0.213	0.206	0.276	0.213	0.206	0.275
	ESE	0.268	0.264	0.322	0.213	0.204	0.272	0.213	0.203	0.273
	CPr	0.946	0.894	0.952	0.950	0.684	0.949	0.950	0.682	0.950
	SM(γ)	-1.002	-0.861	-1.009	-1.009	-0.717	-1.017	-1.009	-0.717	-1.017
	SSE	0.209	0.214	0.249	0.122	0.117	0.169	0.122	0.117	0.169
	ESE	0.209	0.213	0.245	0.120	0.114	0.168	0.120	0.114	0.166
	CPr	0.952	0.908	0.942	0.954	0.296	0.946	0.955	0.295	0.947

Table 4.3: Simulation results under Design 2 with $(\pi^+, \pi^-) = (0.95, 0.90)$ and the true values of parameters $\beta = (1, 1)$

		GQL			OGQL			ML		
γ	Quantity	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	SM(β_1)	1.005	0.805	1.008	0.998	0.797	0.999	0.998	0.797	0.999
	SSE	0.124	0.103	0.161	0.113	0.099	0.143	0.112	0.099	0.143
	ESE	0.126	0.107	0.160	0.117	0.104	0.147	0.117	0.104	0.148
	CPr	0.950	0.553	0.959	0.956	0.482	0.967	0.954	0.486	0.969
	SM(β_2)	0.997	1.153	1.010	1.019	1.217	1.027	1.019	1.217	1.030
	SSE	0.446	0.367	0.570	0.234	0.207	0.323	0.234	0.207	0.323
	ESE	0.440	0.368	0.564	0.238	0.205	0.330	0.239	0.209	0.333
	CPr	0.947	0.920	0.945	0.959	0.821	0.960	0.958	0.831	0.960
	SM(γ)	1.541	0.851	1.559	1.498	0.777	1.505	1.498	0.777	1.501
	SSE	0.483	0.352	0.643	0.184	0.150	0.300	0.184	0.150	0.299
	ESE	0.475	0.365	0.627	0.191	0.156	0.313	0.192	0.158	0.313
	CPr	0.945	0.540	0.947	0.958	0.003	0.961	0.960	0.003	0.964
0	SM(β_1)	1.003	0.875	1.006	1.002	0.873	1.003	1.002	0.873	1.003
	SSE	0.093	0.093	0.116	0.093	0.093	0.116	0.093	0.093	0.116
	ESE	0.094	0.092	0.114	0.094	0.091	0.114	0.094	0.091	0.114
	CPr	0.961	0.711	0.948	0.957	0.702	0.942	0.957	0.703	0.947
	SM(β_2)	0.981	0.958	0.984	0.996	0.939	1.001	0.996	0.939	1.001
	SSE	0.277	0.261	0.331	0.173	0.166	0.225	0.173	0.166	0.225
	ESE	0.263	0.256	0.320	0.166	0.159	0.217	0.167	0.160	0.218
	CPr	0.941	0.949	0.950	0.934	0.925	0.953	0.935	0.928	0.952
	SM(γ)	0.024	-0.028	0.025	0.006	-0.008	0.003	0.006	-0.008	0.004
	SSE	0.271	0.258	0.328	0.134	0.121	0.179	0.134	0.121	0.179
	ESE	0.260	0.257	0.315	0.130	0.123	0.180	0.130	0.123	0.181
	CPr	0.940	0.945	0.951	0.938	0.954	0.952	0.936	0.954	0.950
-1	SM(β_1)	1.002	0.879	1.001	1.002	0.862	1.001	1.002	0.862	1.001
	SSE	0.091	0.089	0.110	0.087	0.084	0.107	0.087	0.084	0.107
	ESE	0.093	0.090	0.113	0.090	0.087	0.110	0.090	0.087	0.111
	CPr	0.957	0.725	0.955	0.961	0.653	0.961	0.961	0.652	0.960
	SM(β_2)	0.999	0.973	0.992	1.003	0.853	0.999	1.003	0.853	0.999
	SSE	0.192	0.199	0.231	0.135	0.133	0.172	0.134	0.133	0.173
	ESE	0.202	0.206	0.239	0.141	0.137	0.181	0.141	0.137	0.181
	CPr	0.962	0.964	0.961	0.959	0.808	0.957	0.958	0.809	0.957
	SM(γ)	-1.000	-0.874	-0.992	-1.005	-0.725	-1.001	-1.005	-0.724	-1.001
	SSE	0.212	0.213	0.248	0.109	0.105	0.147	0.109	0.105	0.147
	ESE	0.214	0.219	0.250	0.111	0.108	0.153	0.111	0.108	0.153
	CPr	0.953	0.938	0.952	0.958	0.292	0.962	0.959	0.292	0.964

Table 4.4: Simulation results under Design 3 with $(\pi^+, \pi^-) = (0.95, 0.90)$ and the true values of parameters $\beta = (1, 1)$

		GQL			OGQL			ML		
γ	Quantity	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	SM(β_1)	1.004	0.807	1.004	1.003	0.847	1.003	1.003	0.847	1.003
	SSE	0.072	0.065	0.093	0.069	0.061	0.088	0.069	0.061	0.088
	ESE	0.071	0.065	0.090	0.069	0.062	0.086	0.069	0.062	0.087
	CPr	0.942	0.162	0.943	0.942	0.299	0.948	0.942	0.300	0.947
	SM(β_2)	1.001	0.697	1.002	1.003	0.885	1.001	1.003	0.885	1.002
	SSE	0.169	0.153	0.206	0.143	0.119	0.178	0.143	0.119	0.178
	ESE	0.165	0.152	0.204	0.139	0.123	0.179	0.139	0.124	0.181
	CPr	0.945	0.492	0.947	0.935	0.855	0.946	0.937	0.858	0.943
	SM(γ)	1.506	1.402	1.512	1.502	1.128	1.511	1.502	1.128	1.511
	SSE	0.184	0.175	0.214	0.123	0.107	0.167	0.123	0.107	0.166
	ESE	0.179	0.175	0.211	0.122	0.109	0.167	0.122	0.108	0.167
	CPr	0.943	0.901	0.952	0.951	0.074	0.951	0.953	0.073	0.943
0	SM(β_1)	1.005	0.868	1.005	1.004	0.886	1.005	1.004	0.886	1.005
	SSE	0.068	0.062	0.079	0.061	0.056	0.073	0.061	0.056	0.073
	ESE	0.066	0.063	0.080	0.060	0.057	0.074	0.060	0.057	0.074
	CPr	0.937	0.454	0.948	0.946	0.474	0.950	0.946	0.474	0.954
	SM(β_2)	0.999	0.817	0.999	0.999	0.870	1.001	0.999	0.870	1.001
	SSE	0.131	0.125	0.154	0.108	0.103	0.134	0.108	0.103	0.134
	ESE	0.128	0.126	0.154	0.106	0.102	0.133	0.106	0.102	0.133
	CPr	0.938	0.680	0.951	0.941	0.736	0.956	0.941	0.735	0.956
	SM(γ)	-0.001	0.132	0.001	-0.001	-0.056	-0.002	-0.001	-0.056	-0.002
	SSE	0.145	0.143	0.169	0.095	0.093	0.123	0.095	0.093	0.123
	ESE	0.142	0.143	0.169	0.095	0.093	0.124	0.095	0.093	0.124
	CPr	0.949	0.837	0.954	0.954	0.908	0.961	0.954	0.904	0.958
-1	$\hat{\beta}_1$	1.003	0.854	1.004	1.002	0.848	1.004	1.002	0.848	1.004
	SSE	0.075	0.070	0.091	0.063	0.059	0.080	0.063	0.059	0.080
	ESE	0.074	0.068	0.091	0.060	0.057	0.079	0.060	0.057	0.079
	CPr	0.942	0.432	0.955	0.942	0.245	0.944	0.942	0.245	0.946
	$\hat{\beta}_2$	1.003	0.854	1.003	1.003	0.844	1.004	1.003	0.844	1.004
	SSE	0.120	0.119	0.145	0.097	0.093	0.121	0.097	0.093	0.121
	ESE	0.122	0.119	0.147	0.095	0.093	0.122	0.095	0.093	0.123
	CPr	0.956	0.754	0.954	0.944	0.613	0.953	0.945	0.940	0.952
	$\hat{\gamma}$	-1.010	-0.708	-1.011	-1.009	-0.690	-1.012	-1.009	-0.690	-1.012
	SSE	0.157	0.157	0.191	0.099	0.095	0.135	0.099	0.095	0.135
	ESE	0.161	0.155	0.195	0.097	0.093	0.135	0.097	0.092	0.135
	CPr	0.961	0.519	0.956	0.941	0.098	0.952	0.942	0.094	0.955

Table 4.5: Simulation results under Design 1 with $(\pi^+, \pi^-) = (0.75, 0.80)$ and the true values of parameters $\beta = (1, 1)$

		GQL			OGQL			ML		
γ	Quantity	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	SM(β_1)	0.999	0.327	1.018	0.999	0.381	1.017	0.999	0.381	1.013
	SSE	0.069	0.053	0.171	0.135	0.049	0.166	0.066	0.049	0.161
	ESE	0.071	0.055	0.167	0.139	0.051	0.161	0.069	0.051	0.160
	CPr	0.949	0.000	0.950	0.958	0.000	0.953	0.958	0.000	0.958
	SM(β_2)	0.990	0.293	1.012	0.996	0.494	1.010	0.996	0.494	1.005
	SSE	0.164	0.116	0.367	0.135	0.088	0.347	0.135	0.088	0.336
	ESE	0.165	0.121	0.367	0.139	0.091	0.347	0.139	0.092	0.347
	CPr	0.956	0.000	0.957	0.962	0.000	0.955	0.962	0.000	0.962
	SM(γ)	1.520	0.686	1.533	1.512	0.361	1.536	1.512	0.361	1.530
	SSE	0.184	0.148	0.335	0.124	0.084	0.316	0.124	0.084	0.307
	ESE	0.180	0.159	0.357	0.122	0.089	0.328	0.122	0.092	0.326
	CPr	0.951	0.000	0.959	0.953	0.000	0.958	0.952	0.000	0.966
0	SM(β_1)	1.006	0.451	1.024	1.003	0.434	1.021	1.003	0.434	1.019
	SSE	0.063	0.056	0.136	0.056	0.049	0.1306	0.056	0.049	0.130
	ESE	0.066	0.058	0.139	0.060	0.051	0.133	0.060	0.051	0.133
	CPr	0.962	0.000	0.962	0.964	0.000	0.963	0.964	0.000	0.962
	SM(β_2)	1.006	0.427	1.013	1.001	0.381	1.009	1.001	0.381	1.001
	SSE	0.127	0.112	0.262	0.103	0.085	0.244	0.103	0.085	0.243
	ESE	0.128	0.113	0.263	0.106	0.087	0.245	0.106	0.087	0.246
	CPr	0.953	0.001	0.960	0.953	0.000	0.963	0.953	0.000	0.966
	SM(γ)	-0.009	-0.117	-0.016	-0.002	-0.036	-0.010	-0.002	-0.036	-0.007
	SSE	0.143	0.155	0.278	0.095	0.089	0.246	0.095	0.089	0.244
	ESE	0.142	0.154	0.275	0.095	0.086	0.240	0.095	0.086	0.241
	CPr	0.951	0.875	0.955	0.950	0.927	0.947	0.951	0.929	0.952
-1	SM(β_1)	0.998	0.402	1.000	0.997	0.399	1.000	0.997	0.399	0.997
	SSE	0.129	0.112	0.273	0.129	0.107	0.270	0.129	0.107	0.268
	ESE	0.123	0.107	0.272	0.122	0.106	0.269	0.122	0.106	0.269
	CPr	0.936	0.000	0.953	0.939	0.000	0.948	0.938	0.000	0.950
	SM(β_2)	1.004	0.471	1.017	1.012	0.220	1.021	1.012	0.220	1.006
	SSE	0.272	0.251	0.558	0.216	0.171	0.514	0.216	0.171	0.506
	ESE	0.2687	0.244	0.557	0.213	0.170	0.514	0.213	0.170	0.515
	CPr	0.948	0.436	0.947	0.949	0.006	0.947	0.949	0.006	0.947
	SM(γ)	-0.997	-0.579	-1.005	-1.006	-0.237	-1.011	-1.006	-0.237	-0.996
	SSE	0.214	0.246	0.407	0.121	0.094	0.337	0.121	0.094	0.329
	ESE	0.210	0.248	0.403	0.120	0.095	0.342	0.120	0.095	0.344
	CPr	0.942	0.619	0.956	0.958	0.000	0.960	0.959	0.000	0.965

Table 4.6: Simulation results under Design 2 with $(\pi^+, \pi^-) = (0.75, 0.80)$ and the true values of parameters $\beta = (1, 1)$

γ	Quantity	GQL			OGQL			ML		
		(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	SM(β_1)	1.005	0.804	1.008	0.996	0.794	1.004	0.996	0.795	1.005
	SSE	0.122	0.100	0.161	0.112	0.095	0.130	0.112	0.096	0.130
	ESE	0.127	0.108	0.160	0.116	0.104	0.132	0.117	0.104	0.131
	CPr	0.958	0.955	0.960	0.962	0478	0.954	0.962	0.472	0.952
	SM(β_2)	1.003	1.1613	1.009	1.037	1.231	1.011	1.037	1.230	1.003
	SSE	0.439	0.357	0.571	0.227	0.119	0.342	0.227	0.200	0.331
	ESE	0.431	0.369	0.535	0.240	0.215	0.339	0.240	0.210	0.331
	CPr	0.944	0.914	0.948	0.958	0.826	0.959	0.958	0.826	0.962
	SM(γ)	1.550	0.852	1.579	1.501	0.776	1.534	1.492	0.774	1.540
	SSE	0.492	0.354	0.661	0.181	0.149	0.434	0.180	0.149	0.432
	ESE	0.457	0.365	0.591	0.192	0.158	0.436	0.192	0.158	0.431
	CPr	0.942	0.546	0.946	0.956	0.002	0.959	0.946	0.002	0.953
0	SM(β_1)	0.998	0.424	1.020	0.997	0.423	1.014	0.997	0.423	1.012
	SSE	0.094	0.080	0.197	0.093	0.080	0.197	0.093	0.080	0.195
	ESE	0.094	0.082	0.202	0.094	0.082	0.200	0.094	0.082	0.201
	CPr	0.944	0.000	0.955	0.953	0.000	0.954	0.953	0.000	0.956
	SM(β_2)	0.988	0.343	0.971	1.006	0.371	1.017	1.006	0.371	0.995
	SSE	0.272	0.232	0.592	0.168	0.123	0.465	0.168	0.123	0.442
	ESE	0.263	0.226	0.565	0.167	0.127	0.456	0.167	0.127	0.458
	CPr	0.944	0.175	0.953	0.947	0.002	0.950	0.948	0.002	0.962
	SM(γ)	0.020	0.039	0.047	-0.001	0.000	-0.007	-0.001	0.000	0.014
	SSE	0.272	0.285	0.604	0.132	0.097	0.437	0.132	0.097	0.411
	ESE	0.259	0.273	0.562	0.132	0.099	0.422	0.130	0.099	0.424
	CPr	0.944	0.948	0.951	0.944	0.957	0.949	0.944	0.957	0.958
-1	SM(β_1)	1.004	0.442	1.007	1.003	0.425	1.005	1.003	0.425	1.002
	SSE	0.092	0.085	0.196	0.089	0.080	0.190	0.089	0.080	0.189
	ESE	0.093	0.083	0.197	0.090	0.081	0.193	0.090	0.081	0.194
	CPr	0.953	0.000	0.955	0.958	0.000	0.955	0.958	0.000	0.959
	SM(β_2)	0.992	0.406	0.999	1.000	0.202	1.014	1.000	0.202	1.001
	SSE	0.203	0.198	0.402	0.138	0.119	0.341	0.138	0.119	0.335
	ESE	0.202	0.197	0.393	0.141	0.120	0.343	0.141	0.119	0.344
	CPr	0.949	0.160	0.955	0.952	0.000	0.952	0.954	0.000	0.951
	SM(β_2)	-0.990	-0.573	-0.994	-1.001	-0.262	-1.012	-1.001	-0.262	-0.997
	SSE	0.218	0.255	0.423	0.109	0.095	0.313	0.109	0.095	0.306
	ESE	0.214	0.251	0.406	0.111	0.094	0.319	0.111	0.094	0.320
	CPr	0.947	0.618	0.944	0.958	0.000	0.961	0.959	0.000	0.963

Table 4.7: Simulation results under Design 3 with $(\pi^{+}, \pi^{-}) = (0.75, 0.80)$ and the true values of parameters $\beta = (1, 1)$

		GQL			OGQL			ML		
γ	Quantity	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	SM(β_1)	0.999	0.327	1.018	0.999	0.381	1.017	0.999	0.381	1.013
	SSE	0.069	0.053	0.171	0.066	0.049	0.166	0.066	0.049	0.161
	ESE	0.071	0.055	0.167	0.069	0.051	0.160	0.069	0.051	0.160
	CPr	0.949	0.000	0.950	0.958	0.000	0.953	0.958	0.000	0.958
	SM(β_2)	0.990	0.293	1.012	0.996	0.494	1.010	0.996	0.494	1.005
	SSE	0.164	0.116	0.367	0.135	0.088	0.347	0.135	0.088	0.336
	ESE	0.165	0.121	0.367	0.139	0.091	0.347	0.139	0.092	0.347
	CPr	0.956	0.000	0.957	0.962	0.000	0.955	0.962	0.000	0.962
	SM(γ)	1.520	0.686	1.533	1.512	0.361	1.536	1.512	0.361	1.530
	SSE	0.184	0.148	0.335	0.124	0.084	0.316	0.124	0.084	0.307
	ESE	0.180	0.159	0.347	0.122	0.089	0.328	0.122	0.088	0.326
	CPr	0.951	0.000	0.959	0.953	0.000	0.958	0.952	0.000	0.966
0	SM(β_1)	1.006	0.451	1.024	1.003	0.434	1.021	1.003	0.434	1.019
	SSE	0.063	0.056	0.136	0.056	0.049	0.131	0.056	0.049	0.130
	ESE	0.066	0.058	0.139	0.060	0.051	0.133	0.060	0.051	0.133
	CPr	0.962	0.000	0.962	0.964	0.000	0.963	0.964	0.000	0.962
	SM(β_2)	1.006	0.427	1.013	1.001	0.381	1.009	1.001	0.381	1.006
	SSE	0.127	0.112	0.262	0.103	0.085	0.244	0.103	0.085	0.243
	ESE	0.128	0.113	0.263	0.106	0.087	0.245	0.106	0.087	0.246
	CPr	0.953	0.001	0.960	0.953	0.000	0.963	0.953	0.000	0.966
	SM(γ)	-0.009	-0.117	-0.016	-0.002	-0.036	-0.010	-0.002	-0.036	-0.007
	SSE	0.143	0.155	0.278	0.095	0.089	0.246	0.095	0.089	0.244
	ESE	0.142	0.154	0.275	0.095	0.086	0.240	0.095	0.086	0.241
	CPr	0.951	0.875	0.955	0.95	0.927	0.947	0.951	0.929	0.952
-1	$\hat{\beta}_1$	1.001	0.457	1.013	0.998	0.379	1.009	0.998	0.379	1.005
	SSE	0.074	0.065	0.161	0.059	0.052	0.148	0.059	0.052	0.146
	ESE	0.074	0.062	0.155	0.060	0.050	0.145	0.060	0.051	0.149
	CPr	0.948	0.000	0.939	0.961	0.000	0.955	0.961	0.000	0.954
	$\hat{\beta}_2$	1.001	0.487	1.016	0.999	0.312	1.013	0.999	0.312	1.008
	SSE	0.124	0.117	0.251	0.096	0.083	0.226	0.096	0.083	0.223
	ESE	0.122	0.112	0.245	0.095	0.083	0.222	0.095	0.083	0.223
	CPr	0.950	0.008	0.949	0.943	0.000	0.951	0.943	0.000	0.954
	$\hat{\gamma}$	-1.002	-0.706	-1.025	-0.999	-0.357	-1.020	-0.999	-0.357	-1.011
	SSE	0.163	0.172	0.321	0.096	0.083	0.261	0.096	0.083	0.255
	ESE	0.161	0.170	0.318	0.097	0.085	0.264	0.097	0.086	0.264
	CPr	0.950	0.563	0.947	0.944	0.000	0.955	0.945	0.000	0.957

4.2.3.3 Insight to robustness: a continued simulation study

Recall that in the previous subsection, we checked the performance of the ideal estimates, naive estimates, and corrected estimates of $\theta = (\beta', \gamma)'$ under the GQL, OGQL, and ML approaches, and found that the corrected estimates suit the practical situation best. In this subsection, we continue to examine the robustness of the corrected estimates under the three approaches. Note that in the previous subsection, we simply assumed that the sensitivity π^+ and specificity π^- are known. This assumption, however, may not applicable in practice. Some authors suggest that estimates of these two parameters, π^+ and π^-, may be derived based on some independent validation studies or from historical knowledge. These estimates or historical values may be biased from the true values. Therefore, it is our main interest in this subsection to investigate the robustness of the corrected estimates under the situations that slightly biased estimates of π^+ and π^- are used when they are unknown.

For the regression parameter β and the dynamic dependence parameter γ, we chose $\beta = (1, 1)'$, and two values for γ: 1 and 0. As far as the true values of π^+ and π^- are concerned, we choose the same pairs of values used in the previous subsection, which are $(\pi^+, \pi^-) = (0.95, 0.90)$ and $(\pi^+, \pi^-) = (0.75, 0.80)$. To be specific, in the case of $(\pi^+, \pi^-) = (0.95, 0.90)$, three pairs of biased estimates of (π^+, π^-) are used when $\gamma = 1$ or 0, they are (0.96, 0.91), (0.94, 0.89) and (0.97, 0.92), and two extra pairs (0.965, 0.915) and (0.935, 0.885), are used when $\gamma = 0$. 500 simulations are conducted when the true values $(\pi^+, \pi^-) = (0.95, 0.90)$. Similarly, for the true values $(\pi^+, \pi^-) = (0.75, 0.80)$, three pairs of biased estimates of (π^+, π^-) are used for both cases where $\gamma = 1$ or 0, they are (0.76, 0.81), (0.74, 0.79) and (0.77, 0.82). Two more pairs (0.765, 0.815) and (0.735, 0.785) are used when $\gamma = 0$. 500 simulations are conducted in the case of $(\pi^+, \pi^-) = (0.75, 0.80)$. The simulation results are reported in Table 4.8 for $(\pi^+, \pi^-) = (0.95, 0.90)$, and Table 4.9 for $(\pi^+, \pi^-) = (0.75, 0.80)$, where in both tables, covariate design 3 given in Subsection 4.2.3 is used.

It can be seen from Table 4.8 and Table 4.9 that the OGQL approach continues to perform similarly to the ML approach in most of cases. Therefore in the following part of this paragraph, we focus on the comparison between the GQL approach with the OGQL approach, keeping in mind that the GQL approach is being compared with the ML approach at the same time. It is clear from the simulation results that all three approaches produce very small biases on the estimates of $\theta = (\beta', \gamma)'$ when the biases of of estimated (π^+, π^-) are very small. However, the estimate of parameters and confidence intervals for some parameters may not perform well when the bias of π^+ and π^- gets larger. For example, for true values of $(\pi^+, \pi^-) = (0.95, 0.90)$ when the values $(0.97, 0.92)$ are used in the case that $\gamma = 1$ (Table 4.8), $\hat{\gamma}_{GQL} = 0.895$ (the CPr's 0.934), and $\hat{\gamma}_{OGQL} = 0.886$ (the CPr's 0.860). They are significantly smaller than the true value $\gamma = 1$ (the nominal level 0.95). Therefore, when the biases of the estimated sensitivity π^+ and specificity π^- become larger, the estimate of model parameters θ produce more and more biases from their true values.

Furthermore, the simulation results in Tables 4.8 and 4.9 show that the OGQL approach tends to performs better than the GQL approach as far as the simulated standard errors (SSE) are concerned. For example, when $(\pi^+, \pi^-) = (0.95, 0.90)$, $\gamma = 1$ (Table 4.8), and the "working" values $(\pi^+, \pi^-) = (0.94, 0.89)$ are used, the computed SSE's of $\hat{\theta}_{GQL}$ are given by $(0.089, 0.191, 0.422)$, each one of them is greater than the corresponding SSE of $\hat{\theta}_{OGQL}$, which are given by $(0.080, 0.124, 0.171)$. Whereas under both methods, the ESE's of $\hat{\theta}$ are quite close to the respective SSE's. However, the biases of the OGQL estimates of θ may be greater than the GQL estimates in the case that the "working" sensitivity and specificity have much biases. For example, $\hat{\gamma}_{OGQL} = 0.844$ more biased from the true value 1 than $\hat{\gamma}_{GQL} = 0.883$. This is because the OGQL approach may use more error information than GQL approach does when the much biased sensitivity and specificity used,.

Another interesting finding is that when the working values of (π^+, π^-) are slightly overestimated, the corresponding simulation results, especially, the SSE's and CPr's, show to be better as compared with the situation where the working values are slightly underestimated. For instance, for $(\pi^+, \pi^-) = (0.75, 0.80)$, $\gamma = 0$ (Table 4.9), when $(0.76, 0.81)$ are used as the working values of (π^+, π^-), the SSE's of $\hat{\theta}_{OGQL} = (\beta', \gamma)'$ are given by $(0.108, 0.215, 0.311)$, which are smaller than the corresponding SSE's given by $(0.125, 0.251, 0.365)$ when the underestimated values $(\pi^+, \pi^-) = (0.74, 0.79)$ are used. The same finding holds for the GQL and ML approaches.

In summary, based on the simulation results in Tables 4.8 and 4.9, we recommend the use of slightly overestimated sensitivity π^+ and specificity π^-, to ensure satisfactorily robust estimation results. Also, we suggest the use of the OGQL approach, since it produces tiny biases on the estimates of model parameters and smaller SSE's than the GQL method when the "working" sensitivity and specificity have small biases. In addition the, OGQL approach is also less complicated than the ML approach when computation is concerned.

Table 4.8: Robustness about estimated (π^+, π^-) based on 500 simulations under Design 3 with true values $(\pi^+, \pi^-) = (0.95, 0.90)$, $\beta = (-1, 1)$, $\gamma = 1, 0$

			GQL			OGQL			ML		
γ	(π^+, π^-)	Quantity	β_1	β_2	γ	β_1	β_2	γ	β_1	β_2	γ
1	(0.95,0.90)	SM	-1.003	1.013	1.004	-1.001	1.008	1.001	-1.001	1.006	1.001
		SSE	0.085	0.183	0.394	0.076	0.119	0.160	0.076	0.119	0.159
		ESE	0.083	0.186	0.389	0.075	0.123	0.161	0.075	0.124	0.161
		CPr	0.938	0.950	0.960	0.956	0.960	0.950	0.950	0.960	0.942
	(0.96,0.91)	SM	-0.968	0.985	0.948	-0.966	0.983	0.941	-0.966	0.982	0.942
		SSE	0.081	0.175	0.370	0.073	0.114	0.149	0.073	0.114	0.149
		ESE	0.079	0.178	0.364	0.071	0.118	0.150	0.071	0.119	0.150
		CPr	0.936	0.952	0.948	0.938	0.952	0.930	0.937	0.948	0.930
	(0.94,0.89)	SM	-1.040	1.043	1.062	-1.038	1.032	1.065	-1.038	1.032	1.065
		SSE	0.089	0.191	0.422	0.080	0.124	0.171	0.080	0.124	0.171
		ESE	0.088	0.194	0.417	0.078	0.129	0.173	0.079	0.129	0.173
		CPr	0.930	0.942	0.964	0.930	0.960	0.952	0.930	0.960	0.954
	(0.97,0.92)	SM	-0.934	0.960	0.895	-0.933	0.959	0.886	-0.933	0.959	0.886
		SSE	0.077	0.168	0.348	0.070	0.109	0.140	0.070	0.109	0.139
		ESE	0.075	0.171	0.342	0.068	0.114	0.141	0.068	0.114	0.141
		CPr	0.848	0.954	0.934	0.804	0.946	0.860	0.808	0.948	0.864
0	(0.95,0.90)	SM	-1.008	1.007	-0.007	-1.005	1.005	-0.007	-1.005	1.005	-0.007
		SSE	0.070	0.180	0.327	0.070	0.122	0.156	0.070	0.122	0.156
		ESE	0.072	0.182	0.328	0.071	0.123	0.151	0.071	0.124	0.152
		CPr	0.950	0.944	0.950	0.950	0.944	0.950	0.950	0.948	0.950
	(0.96,0.91)	SM	-0.976	0.974	-0.009	-0.974	0.971	-0.007	-0.974	0.971	-0.007
		SSE	0.067	0.173	0.314	0.067	0.116	0.147	0.067	0.116	0.147
		ESE	0.069	0.175	0.315	0.068	0.118	0.143	0.068	0.119	0.143
		CPr	0.932	0.948	0.950	0.928	0.938	0.948	0.930	0.938	0.950
	(0.94,0.89)	SM	-1.042	1.043	-0.006	-1.039	1.041	-0.007	-1.039	1.041	-0.007
		SSE	0.074	0.188	0.341	0.073	0.128	0.165	0.073	0.1283	0.165
		ESE	0.076	0.190	0.343	0.075	0.130	0.160	0.075	0.131	0.161
		CPr	0.932	0.950	0.946	0.934	0.938	0.952	0.940	0.936	0.946
	(0.965,0.915)	SM	-0.961	0.959	-0.010	-0.959	0.955	-0.007	-0.959	0.955	-0.007
		SSE	0.066	0.170	0.307	0.065	0.114	0.143	0.065	0.114	0.143
		ESE	0.068	0.172	0.308	0.066	0.115	0.139	0.067	0.116	0.139
		CPr	0.918	0.944	0.950	0.902	0.922	0.948	0.902	0.918	0.950
	(0.935,0.885)	SM	-1.060	1.063	-0.005	-1.056	1.061	-0.007	-1.056	1.061	-0.007
		SSE	0.076	0.192	0.349	0.075	0.132	0.170	0.075	0.132	0.170
		ESE	0.078	0.194	0.351	0.076	0.133	0.165	0.077	0.134	0.166
		CPr	0.906	0.952	0.948	0.906	0.936	0.952	0.906	0.934	0.944
	(0.97,0.92)	SM	-0.947	0.943	-0.011	-0.945	0.939	-0.007	-0.945	0.939	-0.007
		SSE	0.065	0.166	0.301	0.064	0.111	0.139	0.064	0.111	0.139
		ESE	0.066	0.169	0.302	0.065	0.113	0.135	0.065	0.113	0.135
		CPr	0.872	0.932	0.950	0.868	0.908	0.950	0.868	0.906	0.950

Table 4.9: Robustness about estimated (π^+, π^-) based on 500 simulations under Design 3 with true values $(\pi^+, \pi^-) = (0.75, 0.80)$, $\beta = (-1, 1)$, $\gamma = 1, 0$

			GQL			OGQL			ML		
γ	(π^+, π^-)	Quantity	β_1	β_2	γ	β_1	β_2	γ	β_1	β_2	γ
1	(0.75,0.80)	SM	-1.012	1.035	1.058	-1.007	1.027	0.996	-1.002	1.035	0.993
		SSE	0.137	0.309	0.672	0.121	0.216	0.362	0.119	0.213	0.358
		ESE	0.138	0.300	0.677	0.124	0.218	0.366	0.124	0.219	0.366
		CPr	0.954	0.960	0.956	0.962	0.952	0.960	0.960	0.952	0.966
	(0.76,0.81)	SM	-0.963	0.964	0.997	-0.955	0.978	0.943	-0.952	0.983	0.944
		SSE	0.135	0.280	0.642	0.116	0.203	0.338	0.115	0.202	0.322
		ESE	0.129	0.279	0.618	0.116	0.204	0.332	0.116	0.205	0.332
		CPr	0.922	0.954	0.956	0.924	0.948	0.938	0.928	0.956	0.932
	(0.74,0.79)	SM	-1.065	1.088	1.093	-1.061	1.070	1.065	-1.052	1.083	1.058
		SSE	0.145	0.330	0.752	0.129	0.232	0.399	0.126	0.227	0.387
		ESE	0.148	0.322	0.749	0.133	0.233	0.406	0.133	0.235	0.404
		CPr	0.954	0.960	0.958	0.942	0.948	0.964	0.956	0.956	0.969
	(0.77,0.82)	SM	-0.917	0.949	0.883	-0.912	0.953	0.844	-0.910	0.955	0.837
		SSE	0.122	0.269	0.570	0.108	0.190	0.302	0.107	0.189	0.298
		ESE	0.121	0.265	0.565	0.109	0.193	0.304	0.109	0.194	0.305
		CPr	0.856	0.936	0.930	0.846	0.942	0.928	0.846	0.944	0.934
0	(0.75,0.80)	SM	-1.021	0.995	-0.047	-1.012	1.007	0.005	-1.012	0.998	0.007
		SSE	0.118	0.313	0.570	0.116	0.233	0.337	0.116	0.229	0.323
		ESE	0.123	0.297	0.567	0.119	0.224	0.331	0.120	0.225	0.323
		CPr	0.972	0.928	0.944	0.964	0.940	0.960	0.962	0.940	0.962
	(0.76,0.81)	SM	-0.972	0.946	0.041	-0.964	0.956	0.002	-0.964	0.951	0.007
		SSE	0.110	0.298	0.542	0.108	0.215	0.311	0.108	0.215	0.303
		ESE	0.115	0.294	0.538	0.111	0.209	0.306	0.111	0.210	0.307
		CPr	0.956	0.939	0.948	0.944	0.944	0.962	0.944	0.944	0.968
	(0.74,0.79)	SM	-1.075	1.051	0.055	-1.065	1.064	0.003	-1.064	1.050	0.011
		SSE	0.127	0.329	0.607	0.125	0.251	0.365	0.124	0.243	0.358
		ESE	0.133	0.323	0.596	0.128	0.247	0.360	0.129	0.242	0.362
		CPr	0.940	0.940	0.946	0.944	0.940	0.958	0.948	0.946	0.960
	(0.765,0.815)	SM	-0.950	0.923	0.038	-0.942	0.933	0.003	-0.942	0.929	0.010
		SSE	0.107	0.286	0.515	0.105	0.211	0.300	0.105	0.209	0.294
		ESE	0.112	0.267	0.497	0.108	0.202	0.295	0.108	0.203	0.296
		CPr	0.930	0.912	0.948	0.922	0.924	0.958	0.926	0.920	0.961
	(0.735,0.785)	SM	-1.099	1.080	0.187	-1.094	1.095	0.003	-1.092	1.078	0.037
		SSE	0.134	0.349	0.608	0.130	0.251	0.381	0.129	0.251	0.379
		ESE	0.138	0.322	0.605	0.134	0.252	0.376	0.134	0.251	0.377
		CPr	0.912	0.940	0.946	0.922	0.938	0.960	0.928	0.944	0.964
	(0.77,0.82)	SM	-0.928	0.901	0.035	-0.921	0.910	0.001	-0.922	0.907	0.008
		SSE	0.104	0.268	0.509	0.101	0.204	0.289	0.101	0.203	0.285
		ESE	0.108	0.266	0.493	0.104	0.196	0.284	0.105	0.197	0.285
		CPr	0.892	0.886	0.946	0.862	0.902	0.952	0.866	0.898	0.964

4.3 Application to Children Asthma Data

H6CS is a large population-based longitudinal study designed to assess the effect of indoor air pollution on the respiratory health of residents of selected cities in the United States, which began in 1974. As a part of H6CS, a study was conducted in Steubenville, Ohio to evaluate the effect of passive smoking on children asthma. This study has collected complete information from 537 children. Each child was visited at home annually from age 7 to 10. At each home visit, parents were interviewed about the symptoms and diagnoses relevant to asthma and allergy history of their child(ren) through a proforma questionnaire [Ware, et al. (1984)]. A Child's asthma status (Positive=1, Negative=0) was decided based on the information collected in the interview. Some other information, such as diet, mother's smoking status (yes=1, no=0) which was determined at the first interview, are also collected.

Compared with clinical examination of each child, questionnaires are relatively easier and more economical to conduct. Unfortunately, because of the complexities of wide range of severity, triggers, and lack of medical knowledge among the public, it is impossible to formulate completely reliable questionnaires [Jenkins, et al. (1996)]. Besides the inherent shortcomings of questionnaires, the challenge for parents to distinguish the symptoms of wheezing from cold symptoms, along with the great overlap of measurements of healthy children and those with previous wheezing, results in reduced accuracy of diagnosis based only on reported symptoms by questionnaires [Dundas and McKenzie (2006)]. Therefore the reported asthma data are likely to be contaminated with diagnosis errors. However, the mother's smoking status, reported by the mothers at the beginning of the survey on a yearly basis, can be reasonably assumed to be free of classification errors.

In this study, mother's smoking habit is considered to be an important risk factor of children asthma. Many studies have reported adverse effects of parental smoking on children respiratory health. H6CS reported an significant increase in the frequency of coughing and wheezing in children living in parental smoking families [Friebele (1996)]. In addition, the 1986 study conducted in Tecumseh, Michigan, reported that parental smoking significantly accounted for increased prevalence and risk of asthma in children [Friebele (1996)]. It was also pointed out by Gilliand et al. (2001) and Pattenden et al. (2006) that the prevalence of wheezing in childhood is strongly associated with exposure to maternal smoking. In order to evaluate the effect of mother's smoking habit on children asthma, some simple preliminary analysis based on the 537 children's observations is conducted, and the result is shown in Table 4.10.

It can be seen from Table 4.10 that among those children living with smoking mothers, 36.9% of them have had at least 1 asthma attack in the past 4 years, as compared to the rate 32.3% among the children with nonsmoking mothers. We have conducted the Pearson Chi-square test and likelihood ratio test to check whether there is a significant difference between these two percentages. The Pearson Chi-square test (p-value=0.282) and likelihood ratio test (p-value=0.283), however, indicate nonsignificant effect of maternal smoking. There are many reasons those cause this controversy in the analysis. One possible reason is the possible misclassification, especially between the children at lower risk of asthma (asthma attack =1) with those healthy children (asthma attacks = 0) reduces the detectibility of the effect of maternal smoking effect. When re-classifying the children with more than 1 reported asthma attack in the past 4 years into the high risk group, while others in the low

Table 4.10: Exploratory Analysis of Asthma Data of 537 Children from Steubenville, Ohio in H6CS

		Maternal smoking					
Status	Asthma attacks	0	Percentage	1	Percentage	Total	Percentage
Heathy	0	237	67.7%	118	63.1%	355	66.1%
Infected	1	65	18.6%	32	17.1%	97	18.1%
	2	25	7.1%	19	10.2%	44	8.2%
	3	12	3.4%	11	5.9%	33	6.2%
	4	11	3.1%	7	3.7%	18	3.4%
	Subtotal	113	32.3%	69	36.9%	182	33.9%
	Total	350		187		537	

risk group, which includes children never attacked by asthma and those attacked only once from age 7-10, the Pearson Chi-square test (p-value=0.066) and likelihood ratio test (p-value=0.070) indicate a non-ignorable association between passive smoking and children asthma.

Lots of studies have been done to evaluate the adverse effect of mother's smoking on children asthma based on this data set. For example, Zeger and Qaqish (1988) analyzed the H6CS data by using the generalized estimating equations (GEE) approach based on a random effect model. Fitzmaurice and Laird (1993) developed likelihood inference based on the mothers' smoking status, the children's age and the interaction between the two. Due to the strong association between the current asthma status and the reported previous attack (Fuhlbrigge et al., 2001), the linear transition model (2.1) appears to be a reasonable choice to analyze the data, and it was used by Sutradhar and Farrell (2007) to examine the effect of mother's smoking and the previous attack on children asthma. Sutradhar and Farrell (2007) developed the generalized qusi-likelihood (GQL), optimal GQL (OGQL) and maximum likelihood (ML) approaches to estimate the effect of mother's smoking habit and the previous asthma attack on the current asthma status of children in a dynamic model set-up [Sutradhar (2003); Sutradhar and Farrell (2007)].

However, all of these analysis were carried out under the assumption that the observed data are free of measurement errors and the data collected by questionnaires provide completely accurate information about the children's asthma status. We have mentioned before that data collected by questionnaires are prone to classification errors, and the simulation study conducted in the previous section also indicate that ignoring measurement errors in the data leads to biased estimates of the unknown parameters. Therefore, in this section, we reanalyze the asthma data by using the corrected GQL, OGQL and ML approaches taking the misclassification into consideration to avoid misleading conclusion.

Recall that the inherent asthma status of the ith child in the jth year, which was denoted by T_{ij}, may not be directly observable. Instead, his/her manifest status Y_{ij} can be easily obtained from the information provided by the parent-reported H6CS questionnaires. Therefore, the relationship between Y_{ij} and T_{ij} can be described by the misclassification model given by

$$Y_{ij} = \pi^+ * T_{ij} + (1 - \pi^-) * (1 - T_{ij}), \text{ for } i = 1, \dots, 537 \text{ and } j = 1, 2, 3, 4, \quad (4.33)$$

where π^+ is the sensitivity and π^- is the specificity of the H6CS questionnaires. We further assume that the dynamic asthma status of a child can be characterized by the non-linear transition model:

$$\begin{aligned}\lambda^*_{i,j|j-1} &= P(T_{ij} = 1|T_{ij-1} = t_{i,j-1}) \\ &= \frac{exp(\beta_1 + \beta_2 MS_{ij} + \gamma t_{i,j-1})}{1 + exp(\beta_1 + \beta_2 MS_{ij} + \gamma t_{i,j-1})},\end{aligned} \tag{4.34}$$

where MS_{ij} is the maternal smoking status, and $t_{i0} = 0$ are the baseline observations.

We assume that the sensitivity π^+ and specificity π^- of the questionnaire used in this study are constants over time and they are independent of subjects and covariates. To our best knowledge, the sensitivity and specificity of this study are not yet well estimated. However, Yang et al. (1998) conducted a survey to assess the effect of indoor environment on children asthma in Taiwan based on questionnaires which are similar to those used in the H6CS and they reported a sensitivity of 0.80 and a specificity of 0.95. Therefore, we use their results $(\pi^+, \pi^-) = (0.80, 0.95)$ as a close estimate of the true sensitivity and specificity in our study.

Recall that in the simulation study conducted in Section 4.2.3, the naive estimates of $\theta = (\beta', \gamma)'$ were computed by assuming that $(\pi^+, \pi^-) = (1.0, 1.0)$ and that the observations are errors-free. Whereas, in fact, it may be not this case, which implies that $\pi^+ < 1$ and $\pi^- < 1$. Similarly, in this subsection, we choose to compare the naive estimates $((\pi^+, \pi^-) = (1.0, 1.0))$ with the corrected estimates where (π^+, π^-) was chosen to be $(0.80, 0.90)$, through the analysis of the asthma data. The estimation results are listed in Table 4.11.

It can be seen from Table 4.11 that, when taking misclassification into account, the estimates of the model parameters are very different from those obtained by ignoring misdiagnosis, such as the estimates provided in [Sutradhar and Farrell (2007)].

To be specific, the corrected OGQL estimates of β_2 and γ are larger than the corresponding naive estimates. Notice that in Table 4.11, $\hat{\gamma}_{GQL}$ reveals a negative dynamic dependence which is quite unreasonable, which may be due to the loss of information from the higher order responses, It can be verified by the fact that $\hat{\gamma}_{OGQL}$ and $\hat{\gamma}_{ML}$ including more information from data produce a significant positive association with the prior asthma status which is in good accordance with medical practice. Even more, the OGQL and ML approaches demonstrate higher efficiency over the GQL approach. We therefore prefer to accept the results provided by the OGQL and ML approaches.

Although the maternal smoking effect β_2 and the dynamic dependence effect γ are the main interest of statisticians, the odds ratio (OR) of asthma with respect to maternal smoking, and the odds ratio with respect to previous asthma attack are the main focuses in epidemiological studies, The OR with respect to maternal smoking reveals directly the relative risk of asthma attack of children living with smoking mothers compared with those living with non-smoking mothers. It can be seen from Table 4.11 that the OR with respect to maternal smoking, the corrected OGQL estimate of e^{β_2} is equal to 1.3606 which is greater than the naive estimates 1.2468. This is in total agreement with the fact that the corrected estimates of β_2 is greater, than the corresponding naive estimates of β_2 under all three estimation approaches.

Another quantity of interest in this study is e^{γ}, the odds ratio of developing asthma with respect to whether or not the child had an asthma attack in the previous year. As mentioned above the GQL estimate shows an unreasonable negative effect of a previous asthma attack, while the OGQL and ML estimates indicate strongly positive effect. Therefore, we prefer to use the results under the OGQL and ML approaches. From Table 4.11, we can see that the corrected OGQL estimate of this odds ratio of developing further asthma attack is 43.6822 (0.95 CI is (18.9339, 100.7787)) when $(\pi^+, \pi^-) = (0.80, 0.95)$, while the odds ratio is only 7.0669 (0.95 CI is (5.2340, 9.05417)) when misclassification is ignored. Analysis taking diagnosis errors into consideration reveals stronger association between the current asthma status with the a previous attack, which is in concordance with the medical practice. Therefore, the report of a prior asthma attack is of importance in diagnosing a child's current asthma status.

Recall that in the simulation study in Section 4.2.3, we conducted a robustness examination on the corrected estimates of $\theta = (\beta', \gamma)'$ with respect to different choices of π^+ and π^-. In the analysis of the asthma data, we have also examined the robustness of the corrected estimates of β_2 and γ, together with the robustness of the estimates of the odds ratio, e^{β_2} and e^{γ}, namely. The parameters are plotted versus the sensitivity π^+ and specificity π^- in Figures 4.1-4.3, whereas the estimated odds ratios are plotted in Figures 4.4-4.5. An interesting finding is that it is always more robust when estimating the intercept β_1, which functions as the baseline information of asthma incidence. Different values of estimated sensitivity and specificity have milder influence on the estimation of β_1 than on the other model parameters.

It can be seen from Figure 4.4 (a), (b), (c) that when both the sensitivity and specificity increase, the estimated OR $e^{\hat{\beta}_2}$ decreases. This indicates the attenuation effect of overestimated sensitivity and specificity on the estimation of model parameters. Furthermore, the decreasing rate of the OGQL and Ml estimates of OR ($e^{\hat{\beta}_2}$) along the sensitivity is higher in the case of a low sensitivity than that in the case of a high specificity. Also the decrease of estimated relative risk along specificity is much faster than that along the direction of sensitivity, which implies that the estimated specificity has much stronger influence on the estimation of OR ($e^{\hat{\beta}_2}$) than that of the estimated sensitivity. Furthermore, from Figure 4.5 (a) and (b), it is apparent that the estimate of e^{β_2} under the OGQL or ML approaches increases with decreasing sensitivity and specificity. On the contrary, the estimated e^{γ} under GQL has an opposite tendency and the variation is also slighter (Figure 4.5 (c)).

Based on Figures 4.1-4.5 we conjecture that the second or higher order correlation among the responses of the actual asthma data may carry important information about the interested parameters, especially the dynamic dependence parameter γ. This may also explain the failure of GQL approach in estimating γ, which involves only the first order responses. Therefore the OGQL and ML approaches are recommended to analyze the asthma data.

In summary, in this section, we analyzed the children asthma data from Steubenville, Ohio based on the misclassification model (4.33) and the lag 1 dependence model (4.34), the corrected estimates of interested effects under the GQL, OGQL and ML methods were obtained. We not only calculated the interested estimates of θ, the odds ratio e^{β_2} and e^{γ}, which were reported in Table 4.11, but we also checked the robustness of the estimates and odds ratios, which were plotted in Figures 4.1-4.3 and 4.4-4.5. Based on Table 4.11 and Figures 4.1-4.5, we can see that the OGQL approach produced similar results to those of the ML approach. Under OGQL and ML approaches, the corrected estimates reveals stronger association of children asthma with the passive smoking from mothers than the results ignoring measurement errors. Also, the corrected estimates indicate more importance of the history of asthma attacks in predicting the current asthma statuses. By the comparison of the GQL approach with the OGQL and ML approaches, the second order statistics of the responses include important information about the interested effects, especially the power of the previous asthma attack in predicting the future status. So we recommend the OGQL and ML approach to do statistical inference. Furthermore, due to strong dependence of the ML approach on model assumptions, the OGQL approach is therefore preferable when real life data sets are concerned.

Finally, we give a little discussion about the sensitivity π^+ and specificity π^- to conclude this section. As mentioned previously in this thesis, we assume constant sensitivity and specificity which are independent of covariates, time, and subject groups. But in practice, these two quantities may change over these factors. For example, in some other parts of H6CS, asthma of children are reported by parents up to age 9, and self reported after that [Speizer (1990)]. Jenkins (1996) reported the difference in sensitivity and specificity between the parental-report group and self-report group. It would be very interesting to estimate the relative risk of maternal smoking, the odds ratio of previous attack and the sensitivity, specificity simultaneously for those data set.

Table 4.11: Analysis of Asthma Data of 537 Children from Steubenville, Ohio in H6CS Taking Misdiagnosis into Consideration

		GQL			OGQL			ML		
(π^+, π^-)	Quantity	β_1	β_2	γ	β_1	β_2	γ	β_1	β_2	γ
(1,1)	Estimate	-1.7737	0.2843	-0.4959	-2.1886	0.2205	1.9554	-2.1886	0.2205	1.9554
	Ste[a]	0.1185	0.1264	1.2068	0.0891	0.1323	0.1532	0.0893	0.1307	0.1444
	$p-$ value*	0.0000	0.0245	0.6811	0.0000	0.0955	0.0000	0.0000	0.0916	0.0000
	OR[b]		1.3288	0.6090		1.2468	7.0669		1.2468	7.0669
	LB[c]		1.0372	0.0572		0.9620	5.2340		0.9650	5.3252
	UB[d]		1.7024	6.4838		1.6159	9.5417		1.6108	9.3782
(0.8,0.95)	Estimate	-1.9567	0.4071	-0.5419	-2.7858	0.3079	3.7769	-2.7771	0.3206	3.7296
	Ste	0.1686	0.1902	1.9938	0.1628	0.1992	0.4265	0.1187	0.1622	0.2308
	p-value	0.0000	0.0323	0.7858	0.0000	0.1221	0.0000	0.0000	0.0481	0.0000
	OR		1.5024	0.5816		1.3606	43.6822		1.3779	41.6622
	LB		1.0349	0.0117		0.9209	18.9339		1.0027	26.5040
	UB		2.1812	28.9577		2.0103	100.7787		1.8936	65.4896

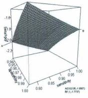
[a]Standard error

[b]Odds ratio of maternal smoking and a previous asthma attack given by $exp(\hat{\beta}_2)$ and $e^{\hat{\pi}}$

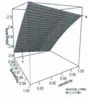
[c]Lower bound of 95% confidence interval of OR given by $e^{\hat{\beta}_2 - z_{0.975} ste_{\hat{\beta}_2}}$ and $e^{\hat{\pi} - z_{0.975} ste_{\hat{\pi}}}$

[d]Upper bound of 95% confidence interval of OR given by $e^{\hat{\beta}_2 + z_{0.975} ste_{\hat{\beta}_2}}$ and $e^{\hat{\pi} + z_{0.975} ste_{\hat{\pi}}}$

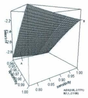
*p-values from Wald type test for two sided hypotheses $\beta_1 = 0$; $\beta_2 = 0$ and $\pi = 0$



(a) GQL estimate of β_1

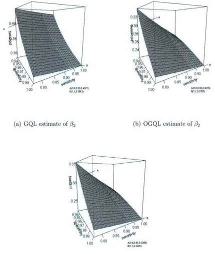


(b) OGQL estimate of β_1



(c) ML estimate of β_1

Figure 4.1: Estimates of the Intercept β_1 in Model (4.34) for Asthma Data of 537 Children from Steubenville, Ohio in H6CS Taking Misdiagnosis into Consideration.

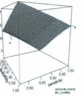


(a) GQL estimate of β_2

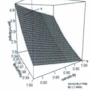
(b) OGQL estimate of β_2

(c) ML estimate of β_2

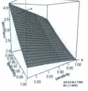
Figure 4.2: Estimates of the Effect of Mother's Smoking Status β_2 in Model (4.34) for Asthma Data of 537 Children from Steubenville, Ohio in H6CS Taking Misdiagnosis into Consideration.



(a) GQL estimate of γ

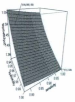


(b) OGQL estimate of γ

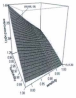


(c) ML estimate of γ

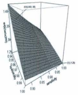
Figure 4.3: Estimates of the Dynamic Dependence Parameter γ in Model (4.34) for Asthma Data of 537 Children from Steubenville, Ohio in H6CS Taking Misdiagnosis into Consideration.



(a) GQL estimate of e^{β_2}

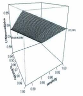


(b) OGQL estimate of e^{β_2}

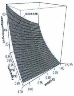


(c) ML estimate of e^{β_2}

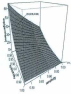
Figure 4.4: Estimates of the Odds Ratio about Mother's Smoking Status for Asthma Data of 537 Children from Steubenville, Ohio in H6CS.



(a) GQL estimate of e^{γ}



(b) OGQL estimate of e^{γ}



(c) ML estimate of e^{γ}

Figure 4.5: Estimates of the Odds Ratio about Prior Asthma Status for Asthma Data of 537 Children from Steubenville, Ohio in H6CS.

4.4 Joint Modeling the Misclassified Data with Missing Information Due to "Unsure" Responses

4.4.1 Model description

The classical misclassification models for categorical data are defined between the latent categorical variable T and the observed categorical Y. T and Y generally have an equal number of classes, for example the misclassification described by Table 2.2 in Chapter 2. But, in practice, there are some cases that the manifest variable Y may have more categories than the inherent variable T does, of which an example is given in Table 2.3 of Chapter 2. In sociological, psychological and epidemiologic studies, it is often the case that data are collected through some proforma questionnaires. Due to the limited knowledge among the public, some people may give an answer like "unsure" or "I don't know". For instance, in studies of children asthma, a question may be set as follows:

> Do you think that your child had asthma in the past 12 months?
>
> (1) Yes; (2) No; (3) $Unsure$.

The possible misclassification can be shown by Table 4.12. This situation is named the unbalanced misclassification in Chapter 2. A child's asthma status may be reported by his or her parents as "infected" (answer: yes), "healthy" (answer: no), or "undecided" (answer: unsure), while his or her true status can only be one of the two statuses: infected and healthy. The information about those children who are reported by responses "unsure" is partially missing. Therefore, the "unsure" responses

Table 4.12: Unbalanced misclassification of children asthma

	Asthma status (T)	
Reported status (Y)	Infected (1)	Healthy (2)
Yes (1)	π_{11}	π_{12}
No (2)	π_{21}	π_{22}
Unsure(3)	π_{31}	π_{32}

are often treated as a kind of missing values in practice. Generally, the quantity π_{31} in Table 4.12, the probability that the parents of an actually infected child give an "unsure" answer, is different from π_{32} which is the probability that the parents of a healthy child give an "unsure" response. This implies that whether a subject has an "unsure" response is related to his/her true status. Therefore, it is missing at random (MAR) but not missing completely at random (MCAR). In fact, even these "Yes" or "No" responses may involve classification errors. Although the true statuses of all subjects are impossible to know, the probabilities in Table 4.12, including the two probabilities π_{31} and π_{32} related to the subjects with "unsure" responses, can be estimated through some validation studies.

In this section, we apply the unbalanced misclassification model to describe this kind of data subject to both misclassification and missing information. Statistical inference based on the GQL approach about the unbalancingly misclassified data is developed.

We define the observed response Y_{ij} as follows:

$$Y_{ij} = \begin{cases} (1,0)', & \text{the } i\text{th child's parents answer "yes" at the } j\text{th time point;} \\ (0,1)', & \text{the answer is "no";} \\ (0,0)', & \text{the answer is "unsure".} \end{cases},$$

where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. The true response T_{ij} takes the value 1, if the true status is "infected", 0 otherwise. So Y_{ij} is a trinomial variable, while T_{ij} is a binary variable.

Let $\widehat{\Pi}$ denote the FMC-matrix which is defined by

$$\widehat{\Pi} = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{pmatrix}. \tag{4.35}$$

Then the MC-matrix is given by

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix} = [\pi_1, \pi_2], \tag{4.36}$$

where $\pi_1 = (\pi_{11}, \pi_{21})'$, and $\pi_2 = (\pi_{12}, \pi_{22})'$. Based on the MC-matrix, the unbalanced misclassification model describing the relationship between Y_{ij} and T_{ij}, according to Section 4.1, can be written as

$$Y_{ij} = \Pi * \widehat{T}_{ij} = \pi_1 * T_{ij} + \pi_2 * (1 - T_{ij}), \tag{4.37}$$

where $\widehat{T}_{ij} = (T_{ij}, 1 - T_{ij})'$.

It is easy to derive some useful moments of the observed response Y_{ij}. The expectation of Y_{ij} is given by

$$\mu_{ij} = E(Y_{ij}) = \Pi \begin{pmatrix} \eta_{ij} \\ 1 - \eta_{ij} \end{pmatrix} = \pi_2 + (\pi_1 - \pi_2)\eta_{ij},$$

where $\eta_{ij} = E(T_{ij})$ is the expectation of the true response T_{ij} which is a scalar. The variance-covariance matrix of Y_{ij} is given by

$$\begin{aligned}
Var(Y_{ij}) &= Var[E(Y_{ij}|T_{ij})] + E[Var(Y_{ij}|T_{ij})] \\
&= \eta_{ij}V_{\pi_1} + (1-\eta_{ij})V_{\pi_2} \\
&\quad +(\pi_1 - \pi_2)Var(T_{ij})(\pi_1 - \pi_2)'.
\end{aligned}$$

As discussed in Section 4.1, $Y_{ij} \sim Multinomial(1, \mu_{ij})$, hence the covariance matrix of Y_{ij} can be written in an alternative form

$$\Sigma_{ijj} = Var(Y_{ij}) = V_{\mu_{ij}} = diag(\mu_{ij}) - \mu_{ij}\mu_{ij}'. \tag{4.38}$$

For $u < j$, the covariance between Y_{ij} and Y_{iu} is given by

$$\begin{aligned}
\Sigma_{iju} &= Cov(Y_{ij}, Y_{iu}) \\
&= \Pi Cov(\widehat{T}_{ij}, \widehat{T}_{iu})\Pi' \\
&= \Pi \begin{pmatrix} cov(T_{ij}, T_{iu}) & cov(T_{ij}, 1 - T_{iu}) \\ cov(1 - T_{ij}, T_{iu}) & cov(1 - T_{ij}, 1 - T_{iu}) \end{pmatrix} \Pi' \\
&= (\pi_1 - \pi_2)Cov(T_{it}, T_{iu})(\pi_1 - \pi_2)'
\end{aligned}$$

It can be seen that Σ_{iju} is singular and of rank 1.

Based on the previous development, we can write the covariance matrix of Y_i in the form of

$$\Sigma_i = \begin{pmatrix} \Sigma_{i11} & \Sigma_{i12} & \cdots & \Sigma_{i1J} \\ \Sigma_{i21} & \Sigma_{i22} & \cdots & \Sigma_{i2J} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{iJ1} & \Sigma_{iJ2} & \cdots & \Sigma_{iJJ} \end{pmatrix}_{2J \times 2J}. \tag{4.39}$$

4.4.2 Estimation of model effects

There are several methods to deal with the imperfect categorical data subject to both misclassification and missing values due to the "unsure" responses

Case I: *Deleting the "unsure" responses and ignoring misclassification.* In this case, the observed response has only two categories, "Yes" and "No", which leads to a binary variable $\tilde{Y}_{ij}$. Actually, $\tilde{Y}_{ij}$ is the first element of original response Y_{ij}. Similarly, the corresponding expectation $\tilde{\mu}_{ij}$ of $\tilde{Y}_{ij}$ is the first element of μ_{ij}, and the covariance $Var(\tilde{Y}_{ij})$ is the $(1,1)$th element of $Cov(Y_{ij})$. So, we use a balanced misclassification model to describe the relationship between $\tilde{Y}_{ij}$ and T_{ij}

$$\tilde{Y}_{ij} = \pi_w^+ * T_{ij} + (1 - \pi_w^-) * (1 - T_{ij}), \tag{4.40}$$

and the "working" FMC matrix is

$$\widehat{\Pi}_w = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

therefore the "working" sensitivity and specificity $(\pi_w^+, \pi_w^-) = (1, 1)$. Notice that, in this case, subjects may have unequally-spaced observations due to deleting the "unsure" values.

Case II: *Deleting the "unsure" values but taking misclassification into consideration.* In this case, we use the same misclassification model as (4.40) due to deleting the third category "unsure" from observed response Y_{ij}. Accordingly, the probabilities in a FMC matrix will be proportionally reassigned as follows

$$\widehat{\Pi}_w = \begin{pmatrix} \frac{\pi_{11}}{\pi_{11}+\pi_{21}} & \frac{\pi_{12}}{\pi_{12}+\pi_{22}} \\ \frac{\pi_{21}}{\pi_{11}+\pi_{21}} & \frac{\pi_{22}}{\pi_{12}+\pi_{22}} \end{pmatrix}$$

which implies that the "working" sensitivity and specificity $(\pi_w^+, \pi_w^-) = (\frac{\pi_{11}}{\pi_{11}+\pi_{21}}, \frac{\pi_{22}}{\pi_{12}+\pi_{22}})$. The subjects may also have unequally-spaced observations similar to those in *Case I*.

Case III: *Ignoring these subjects with at least one missing values but taking misclassification into consideration.* In the case that we delete all the subjects with at least one "unsure" response from the study, the misclassification model (4.40) can still be used because there are no "unsure" responses in the data any more At the same time, the probabilities in a FMC matrix will be proportionally reassigned as follows

$$\widehat{\Pi}_w = \begin{pmatrix} \frac{\pi_{11}}{\pi_{11}+\pi_{21}} & \frac{\pi_{12}}{\pi_{12}+\pi_{22}} \\ \frac{\pi_{21}}{\pi_{11}+\pi_{21}} & \frac{\pi_{22}}{\pi_{12}+\pi_{22}} \end{pmatrix},$$

and the "working" sensitivity and specificity $(\pi_w^+, \pi_w^-) = (\frac{\pi_{11}}{\pi_{11}+\pi_{21}}, \frac{\pi_{22}}{\pi_{12}+\pi_{22}})$. Different from *Case II*, the data in this case do not involve unequally-spaced observations.

Case IV: *Ignoring the misclassification but taking the "unsure" values into consideration.* If we take "unsure" responses into consideration, there will be three observed categories. In this situation, the response Y_{ij} is a trinomial variable of which the value $(0,0)'$ implies that we get an "unsure" answer. Ignoring misclassification means that the probabilities of Type I and type II errors are assumed to be 0's. Therefore, the "working" FMC matrix is

$$\widehat{\Pi}_w = \begin{pmatrix} \pi_{11}+\pi_{21} & 0 \\ 0 & \pi_{12}+\pi_{22} \\ \pi_{31} & \pi_{32} \end{pmatrix}.$$

In this case, all subjects have complete data at time $j = 1, 2, \ldots, J$. Then we use the unbalanced misclassification model (4.37) with the assumed FMC matrix $\widehat{\Pi}_w$ which is given above.

Case V: *Taking both the missing values and misclassification into consideration.* One may take both the "unsure" values and classification errors into consideration to obtain more reliable statistical inference. In this situation, we use the model (4.37) with the true FMC matrix to develop statistical inference. The true FMC matrix is

$$\widehat{\Pi}_w = \widehat{\Pi} = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{pmatrix}.$$

In this case, we have equally-spaced observations for all subjects at time $j = 1, 2, \ldots, J$.

Case VI: *Ideal case that the data $\{t_{ij}\}$ are available.* Suppose that we know the data t_{ij} of the true response, and use the data to conduct statistical inference. The estimation can be based on the model in Section 3.2. In an alternative way, we can use the model (4.26) with the FMC

$$\widehat{\Pi}_w = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

for the data t_{ij}, $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots, J$. In this case, the sensitivity π_w^+ and specificity π_w^- are (1, 1). It should be pointed out that, this case is seldom possible in practice, and it is only applicable in simulation studies.

In the following part, we discuss the estimation of model parameters based on the GQL approach in the six cases. For this purpose, we assume, like that in Section 4.2, that the true response T_{ij} follows the nonlinear transition model (4.9). Therefore, the moments of the true responses T_{ij} can be easily obtained from Section 3.1.2.

An "unsure" response is often treated as a missing value in practice as mentioned in Section 4.4.1. Therefore, if we simply delete this kind of missing values in cases *I*, *II* and *III*, we can use $\widetilde{Y}_{ij}$ in the development of GQL estimation. The expectation $\tilde{\mu}_{ij}$ and covariance $Var(\tilde{Y}_{ij})$ of $\tilde{Y}_{ij}$ are similar to those in (4.12) and their computations can follow the similar development in Section 4.2.1. In cases *IV* and *V*, we use the original response Y_{ij}. In fact the following development can be generalized to some other cases of missing values with MAR mechanism which can be modeled by the unbalanced misclassification.

4.4.2.1 Ignoring the "unsure" responses

Generally, there are two mechanisms to delete the missing values among longitudinal data, which are described in *Case II* and *Case III*. *Case II* is to delete the subjects with at least one "unsure", which makes the analysis simple but leads to loss of efficiency. The other one, as that in *Case III*, is to only delete those "unsure" responses and use all valid observations ("yes" or "no"), which results in loss of less information than the first mechanism. However, the second way may lead to more complexity in simulation studies of GQL approach. The estimating equations only based on the available data can be written as

$$\sum_{i=1}^{I^o} \left(\frac{\partial \tilde{\mu}_i^o}{\partial \theta}\right)' (\widetilde{\Sigma}_i^o)^{-1}(\tilde{y}_i^o - \tilde{\mu}_i^o) = 0. \tag{4.41}$$

In the observed GQL estimating equations (4.41), $\tilde{y}_i^o$ denotes the observed values of subject i, and $\tilde{\mu}_i^o$ and $\widetilde{\Sigma}_i^o$ represent the corresponding expectation and covariance matrix, respectively. In addition, I^o denotes the total number of subjects with at least one valid observation. In simulation studies, missing values ("unsure" responses indicated by $(0,0)'$ for y_{ij}'s) are randomly assigned for both subjects and time points. So, one has to identify the $\tilde{y}_i^o$, $\tilde{\mu}_i^o$ and $\widetilde{\Sigma}_i^o$ for every subject and to construct the observed estimating equations in each simulation run. Furthermore, $\tilde{y}_i^o$, $\tilde{\mu}_i^o$ and $\widetilde{\Sigma}_i^o$ vary in different simulation runs. This leads to considerable difficulty in carrying out a large number of simulations. Therefore, in this subsection, we consider a simple method to construct the observed GQL estimating equations. In this simple way, we use the complete data $\tilde{y}_i$ and its expectation $\tilde{\mu}_i$ and an adjusted covariance matrix $\widetilde{\Sigma}_i^A$ but not $\widetilde{\Sigma}_i^o$, nor $\widetilde{\Sigma}_i$. The adjusted GQL estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \tilde{\mu}_i'}{\partial \theta} (\widetilde{\Sigma}_i^A)^{+}(\tilde{y}_i - \tilde{\mu}_i) = 0, \tag{4.42}$$

where the adjusted covariance matrix $\widetilde{\Sigma}_i^A$ is built by introducing a missing indicator matrix (MIM) Λ_i to the complete covariance matrix $\widetilde{\Sigma}_i$. $(\widetilde{\Sigma}_i^A)^{+}$ denotes the Moore-Penrose inverse of $\widetilde{\Sigma}_i^A$. In the following part, we will show how to construct the adjusted covariance matrix $\widetilde{\Sigma}_i^A$.

For simplicity in notation, we here consider one subject with a series of repeated measurements, which can be easily generalized to longitudinal studies with multiple participants. To be specific, we suppose that the complete data are $\tilde{y} =$

$(\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_J)'$, the observed data are $\tilde{y}^o = (\tilde{y}_{j_1}, \tilde{y}_{j_2}, \ldots, \tilde{y}_{j_s})'$, at the time points $j_1 < j_2 < \ldots < j_s$, We also assume that the subject gives "unsure" answers at time $k_1 < k_2 < \ldots < k_m$, where $m + s = J$. Similar to the previous paragraph, we let $\widetilde{\Sigma}$ and $\widetilde{\Sigma}^o$ denote the variance-covariance matrix of complete $\widetilde{Y}$ and the observed $\widetilde{Y}^o$, respectively. It is easy to see that $\widetilde{\Sigma}^o$ is composed by elements at the intersection of the $(j_1, j_2, \ldots, j_s)$th rows and the $(j_1, j_2, \ldots, j_s)$th columns of $\widetilde{\Sigma}$. We define the MIM as $\Lambda = (\lambda_{uv})_{J\times J}$ of which the(u, v)th element is given by

$$\lambda_{uv} = \begin{cases} 1, & \text{for } u = v \text{ and } \tilde{y}_u \text{ is a valid answer ("yes" or "no");} \\ 0, & \text{for } u = v \text{ and } \tilde{y}_u \text{ is an "unsure" answer;} \\ 0, & \text{for } u \neq v. \end{cases}$$

So the diagonal vector of Λ is the missing indicator vector. It is obvious that Λ is an identity matrix for the complete data $\tilde{y}$, and is singular in presence of "unsure" values with $\Lambda^+ = \Lambda$. Then the adjusted covariance matrix is defined by

$$\widetilde{\Sigma}^A = \Lambda \widetilde{\Sigma} \Lambda', \tag{4.43}$$

the elements at the $(k_1, k_2, \ldots, k_m)$th rows and $(k_1, k_2, \ldots, k_m)$th columns of $\widetilde{\Sigma}^A$ are all zeros, and $\widetilde{\Sigma}^A$ is singular.

Following the development of the adjusted covariance matrix $\widetilde{\Sigma}^A$, we can prove that the adjusted GQL estimating equations (4.42) based on $\widetilde{\Sigma}^A$ are equivalent to the observed GQL estimating equations (4.41).

It is easy to see that $\widetilde{\Sigma}^A$ can be non-singularly transformed into $\begin{pmatrix} \widetilde{\Sigma}^o_{s\times s} & \mathbf{0}_{s\times m} \\ \mathbf{0}_{m\times s} & \mathbf{0}_{m\times m} \end{pmatrix}$ by exchanging the non-zero rows and columns $(j_1, j_2, \ldots, j_s)$ with the zero rows and columns $(k_1, k_2, \ldots, k_m)$. The covariance matrix $\widetilde{\Sigma}^o$ of the observed data $\tilde{y}^o$ is a non-singular submatrix of $\widetilde{\Sigma}^A$. This implies that

$$\widetilde{\Sigma}^A = I_{1,j_1} I_{2,j_2} \cdots I_{s,j_s} \begin{pmatrix} \widetilde{\Sigma}^o & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} I_{s,j_s} \cdots I_{2,j_2} I_{1,j_1}. \tag{4.44}$$

A matrix I_{uv} is defined by

$$\begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0(u) & \cdots & 1 & & & & \\ & & & & 1 & & & & & \\ & & & \vdots & \ddots & \vdots & & & & \\ & & & & & 1 & & & & \\ & & & 1 & \cdots & 0(v) & & & & \\ & & & & & & 1 & & & \\ & & & & & & & \ddots & & \\ & & & & & & & & 1 \end{pmatrix},$$

where the elements $0(u)$ and $0(v)$ in the diagonal line mean that the uth and vth elements are zeros. It is known that $I_{uv}^{-1} = I_{uv}$ and $I_{uv} = I_{vu} = I'_{uv}$. For any matrix B, $I_{uv}B$ represents exchanging the uth and vth rows of matrix B, while BI_{uv} implies exchanging uth and vth columns of B.

Keeping the following well-known conclusion in mind

$$\begin{pmatrix} B_{ss} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}^+ = \begin{pmatrix} B_{ss}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

it is easy to show that the Moore-Penrose inverse of Σ^A is

$$(\widetilde{\Sigma}^A)^+ = I_{1,j_1} I_{2,j_2} \cdots I_{sj_s} \begin{pmatrix} (\widetilde{\Sigma}^o)^{-1} & 0 \\ 0 & 0 \end{pmatrix} I_{s,j_s} \cdots I_{2,j_2} I_{1,j_1}. \tag{4.45}$$

By comparing (4.41) with (4.40), one can seen that $(\widetilde{\Sigma}^A)^+$ has the same zero rows and columns as $\widetilde{\Sigma}^A$.

We now show that the adjusted GQL estimating equations (4.42) is equivalent to the observed GQL estimating equations (4.41).

Denote

$$\frac{\partial \tilde{\mu}'}{\partial \theta} = (\frac{\partial \tilde{\mu}_1}{\partial \theta}, \frac{\partial \tilde{\mu}_2}{\partial \theta}, \cdots \frac{\partial \tilde{\mu}_J}{\partial \theta})$$

and $y - \tilde{\mu} = (\tilde{y}_1 - \tilde{\mu}_1, \tilde{y}_2 - \tilde{\mu}_2, \ldots, \tilde{y}_J - \tilde{\mu}_J)'$. We have

$$\begin{aligned}
\frac{\partial \tilde{\mu}'}{\partial \theta}(\widetilde{\Sigma}^A)^+(y - \tilde{\mu}) &= \frac{\partial \tilde{\mu}'}{\partial \theta} I_{1,j_1} I_{2,j_2} \cdots I_{s,j_s} \begin{pmatrix} (\widetilde{\Sigma}^o)^{-1} & 0 \\ 0 & 0 \end{pmatrix} I_{s,j_s} \cdots I_{2,j_2} I_{1,j_1}(\tilde{y} - \tilde{\mu}) \\
&= \left(\frac{\partial \tilde{\mu}_{j_1}}{\partial \theta}, \frac{\partial \tilde{\mu}_{j_2}}{\partial \theta}, \cdots \frac{\partial \tilde{\mu}_{j_s}}{\partial \theta}, 0, \ldots, 0\right) \begin{pmatrix} (\widetilde{\Sigma}^o)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{y}_{j_1} - \tilde{\mu}_{j_1} \\ \tilde{y}_{j_2} - \tilde{\mu}_{j_2} \\ \vdots \\ \tilde{y}_{j_s} - \tilde{\mu}_{j_s} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \left(\frac{\partial \tilde{\mu}_{j_1}}{\partial \theta}, \frac{\partial \tilde{\mu}_{j_2}}{\partial \theta}, \cdots \frac{\partial \tilde{\mu}_{j_s}}{\partial \theta}\right)(\widetilde{\Sigma}^o)^{-1} \begin{pmatrix} \tilde{y}_{j_1} - \tilde{\mu}_{j_1} \\ \tilde{y}_{j_2} - \tilde{\mu}_{j_2} \\ \vdots \\ \tilde{y}_{j_s} - \tilde{\mu}_{j_s} \end{pmatrix} \\
&= \left(\frac{\partial \tilde{\mu}^o}{\partial \theta}\right)'(\widetilde{\Sigma}^o)^{-1}(\tilde{y}^o - \tilde{\mu}^o),
\end{aligned}$$

where $\frac{\partial \tilde{\mu}'}{\partial \theta} I_{1,j_1} I_{2,j_2} \cdots I_{s,j_s} = (\frac{\partial \tilde{\mu}_{j_1}}{\partial \theta}, \frac{\partial \tilde{\mu}_{j_2}}{\partial \theta}, \cdots \frac{\partial \tilde{\mu}_{j_s}}{\partial \theta}, 0, \ldots, 0)$ and $I_{s,j_s} \cdots I_{2,j_2} I_{1,j_1}(\tilde{y} - \tilde{\mu}) = (\tilde{y}_{j_1} - \tilde{\mu}_{j_1}, \tilde{y}_{j_2} - \tilde{\mu}_{j_2}, \ \ldots, \tilde{y}_{j_s} - \tilde{\mu}_{j_s}, 0, \ldots, 0)'$. By generalizing this result to all participants, we can conclude that the adjusted GQL estimating equations (4.42) are equivalent to the observed estimating equations (4.41). Therefore, the adjusted GQL estimating equations can be rewritten as

$$\sum_{i=1}^{I} \frac{\partial \tilde{\mu}_i'}{\partial \theta}(\Lambda_i \widetilde{\Sigma}_i \Lambda_i)^+(\tilde{y}_i - \tilde{\mu}_i) = 0, \tag{4.46}$$

where Λ_i is the MIM of the subject i, for $i = 1, 2, \ldots, I$. By applying the adjusted GQL approach in simulation studies, we do not have to identify the $\tilde{y}_i^o$, $\tilde{\mu}_i^o$ and $\widetilde{\Sigma}_i^o$ for each sample unit, and we only need to find Λ_i. In the practical applications, we can simply assign any finite values to the "unsure" responses.

In addition, if one simply delete a subject with at least one "unsure" answer, he/she can still use the adjusted GQL approach by assigning a zero MIM for this subject. The MIM for a subject i can be defined as

$$\Lambda_i = \begin{cases} \mathbf{0}_{J\times J}, & \text{for subject } i \text{ with incomplete data;} \\ \mathbf{I}_{J\times J}, & \text{for subject } i \text{ with complete data,} \end{cases} \tag{4.47}$$

where $\mathbf{I}_{J\times J}$ represents the $J \times J$ identity matrix. Let i_u, $u = 1, 2, \ldots, I^c$ denote the individuals with complete observations during the studying period. It is apparent that the adjusted estimating equations are equivalent to

$$\sum_{u=1}^{I^c} \frac{\partial \tilde{\mu}'_{i_u}}{\partial \theta} \widetilde{\Sigma}^{+}_{i_u} (\tilde{y}_{i_u} - \tilde{\mu}_{i_u}) = 0. \tag{4.48}$$

4.4.2.2 Taking missing values into account

Although, some participants may give "unsure" answers, the values of the corresponding covariates may still be useful. In addition, we may have good knowledge about the probability in the unbalanced misclassification matrix Π in (4.36). In this situation, the missing values together with the corresponding covariates and the misclassification probabilities can also provide some useful information for statistical inference. Therefore, one may want to take these "unsure" answers into consideration. Notice that, as mentioned before, there are three observed categories "yes", "no" and "unsure". So we will use the original data y_{ij} as given in Section 4.4.1.

In *Case IV* and *Case V*, we construct the GQL estimation based on the unbalanced misclassification model (4.37) and the corresponding FMC matrices described in Section 4.4.2. The estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{+} (y_i - \mu_i) = 0, \tag{4.49}$$

μ_i and Σ_i can be computed based on the calculations in Section 4.4.1.

In fact, the misclassification model (4.37), the related FMC matrix in (4.35) and an adjusted version of the GQL estimating equations which are given by

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} (\Lambda_i \Sigma_i \Lambda_i)^{+} (y_i - \mu_i) = 0. \tag{4.50}$$

can accommodate the three cases where missing values are deleted in Section 4.4.2.1. In (4.50), Λ_i is defined as

$$\Lambda_i = \begin{pmatrix} \Lambda_{i1} & & & & \\ & \ddots & & & \\ & & \Lambda_{ij} & & \\ & & & \ddots & \\ & & & & \Lambda_{iJ} \end{pmatrix}, \tag{4.51}$$

where

$$\Lambda_{ij} = \begin{cases} I_{2\times 2}, & \text{the answer of the subject } i \text{ at the time } j \text{ is "yes" or "no";} \\ \mathbf{0}_{2\times 2}, & \text{otherwise.} \end{cases}$$

in *Case I* and *Case II*, whereas

$$\Lambda_i = \begin{cases} I_{2J\times 2J} & \text{there are no "unsure" anwser for the subject } i; \\ \mathbf{0}_{2J\times 2J}, & \text{the } i\text{th subject has at least one "unsure".} \end{cases}$$

in *Case III*

The working FMC involved in (4.37) is, in *Case I*,

$$\widehat{\Pi}^w = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

And it is

$$\widehat{\Pi}^w = \begin{pmatrix} \frac{\pi_{11}}{\pi_{11}+\pi_{12}} & \frac{\pi_{21}}{\pi_{21}+\pi_{22}} \\ \frac{\pi_{12}}{\pi_{11}+\pi_{12}} & \frac{\pi_{22}}{\pi_{21}+\pi_{22}} \\ 0 & 0 \end{pmatrix}$$

in both *Case II* and *Case III*. In equations (4.50), we use the data y_i and expectations μ_i and covariance matrices Σ_i instead of $\tilde{y}_i$ and its moments.

4.4.3 Simulation

In this section, we carry out simulation studies to check the performance of the GQL estimates in all the six cases described in Section 4.4.1. The ideal case means that we use the data with the true responses T, which follows the nonlinear transition model (4.9) to estimate parameters in this model.

4.4.3.1 Design

We apply the same covariate design as Design 3 described in Section 4.2.3.1. It is rewritten here by

$x_{ij(1)} = 1,\ (j = 1, 2);\ x_{ij(1)} = 0,\ (j = 3, 4),\ i = 1, \ldots, 140,$

$x_{ij(1)} = 1,\ j = 1, \ldots, 4,\ i = 141, \ldots, 420,$

$x_{ij(1)} = 1,\ (j = 1, 2);\ x_{ij(1)} = -1,\ (j = 3, 4),\ i = 421, \ldots, 560.$

$x_{ij(2)} = -0.5,\ (j = 1, 2);\ x_{ij(2)} = 0.5,\ (j = 3, 4),\ i = 1, \ldots, 140;$

$x_{ij(2)} = j/4,\ j = 1, \ldots, 4,\ i = 141, \ldots, 420,$

$x_{ij(2)} = 0.5(j-2),\ j = 1, \ldots, 4,\ i = 421, \ldots, 560.$

As far as the FMC matrix design is concerned, we consider three different settings which are given by

$$\textbf{(a)}.\ \widehat{\Pi} = \begin{pmatrix} 0.8 & 0.02 \\ 0.03 & 0.9 \\ 0.17 & 0.08 \end{pmatrix}$$

$$\textbf{(b)}.\ \widehat{\Pi} = \begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.75 \\ 0.2 & 0.15 \end{pmatrix}$$

$$\textbf{(c)}.\ \widehat{\Pi} = \begin{pmatrix} 0.8 & 0.08 \\ 0.17 & 0.9 \\ 0.03 & 0.02 \end{pmatrix}$$

From setting (a), (b) to (c), the classification errors becomes severer and severer. On the other hand, among the three settings, setting (c) involves the fewest values of $(0,0)'$ for y_{ij}, which correspond to "unsure" answers, while setting (b) generate the most values of $(0,0)'$ for y_{ij}.

In each setting, the true binary data t_{ij}, $i = 1, 2, \ldots, 560$ and $j = 1, 2, \ldots, 6$ are generated following the model (4.9). The observed data y_{ij} are generated from the unbalanced misclassification model (4.33). The generation of y_{ij} follows the procedure described below.

(1) Once we have t_{ij}, we can firstly generate a trinomial variable U from $Trinomial(t_{ij}, \pi_1)$. In **R** package, it should be $U < -rmultinom(t[i,j], 1, \widehat{\pi}_1)$, where $\widehat{\pi}_1 = (\pi_1', 1 -$

$\mathbf{1}'\pi_1)$.

(2) Secondly, we generate another trinomial variable V from $Trinomial(1-t_{ij}, \pi_2)$. In **R** package, it should be $V < -rmultinom(1-t[i,j], 1, \tilde{\pi}_2)$, where $\tilde{\pi}_2 = (\pi_2', 1-\mathbf{1}'\pi_2)$. Then we let $W = U + V$.

(3) Finally, we get $y_{ij} = W[1:2]$, which means y_{ij} takes the vector consists of the first two element of W.

Notice that in **R** package, U, V and W are three dimensional vectors. Therefore, to accommodate the development in this thesis, our two dimensional response y_{ij} takes the first two elements of W, that is $y[i,j] < -W[1:2]$ in **R** package. When a y_{ij} takes $(0,0)'$, it means that this subject gives an "unsure" answer which can be further treated as a missing value.

We conduct 500 simulations in each case described in Section 4.4.2. In *Case I*, *II* and *III*, we apply the adjusted GQL estimating equations (4.42), and in *Case IV* and *V*, we apply the GQL estimating equations (4.49). For *Case VI*, we apply the GQL estimating equations (4.12) but use data t_i with perfect sensitivity and specificity, that is , $\pi_1 = 1$ and $\pi_0 = 1$, which is equivalent to the estimating equations (3.32) in Section 3.1.2.

4.4.3.2 Simulation results

The simulation results are presented in Table 4.13. It can be seen from the table that ignoring both misclassification and missing values (*Case I*) leads to nonignorable biases on estimates of model parameters $\theta = (\beta_1, \beta_2, \gamma)'$, and the biases become bigger when there are higher degree of misclassification in the three settings of FMC matrix from (a), (b) to (c). For example, in *Case I*, the GQL estimates of θ in three settings are, respectively, $(0.810, 0.806, 1.274)$, $(0.554, 0.491, 1.079)$ and $(0.452, 0.438, 0.821)$. These estimates are more and more biased from their true values (1, 1, 1.5) as the degree of misclassification increases. Accordingly, the coverage probabilities of 95% confidence intervals becomes lower and lower. When we only ignore misclassification and take missing values ("unsure") into consideration, the estimates becomes slightly better, which can be seen by the comparison of results in *Case I* and *Case IV* in the table. However, the estimates still have significant biases. By comparing the results in *Case I* and *IV* with those in *Case V*, we can conclude that ignoring misclassification leads to nonignorable biases on estimates of model parameters. However the estimated standard errors (ESE's) keep very close to the corresponding SSE's.

Ignoring these "unsure" responses ($y_{ij} = (0,0)'$), or deleting those y_i of which at least one observation $y_{ij} = (0,0)'$ results in small biases of GQL estimates of model parameters as long as we take misclassification into consideration. For example, $\hat{\theta} = (1.032, 1.046, 1.526)'$ (the true values are $\theta = (1,1,1.5)'$) in *Case III* with setting (c) which has the severest measurement errors among the three settings, and $\hat{\theta} = (1.035, 1.1114, 1.403)'$ in *Case III* with setting (b) which has the most missing values. Although, the estimates in *Case II* and *Case III* have little biases, the CPr's of 95% CI tend to be much lower than the nominal level 0.95, especially in *Case III* where we delete those subjects with incomplete observations. For instance, in setting (b) which leads to most missing values, we have CPr's of θ: $(0.856, 0.882, 0.828)$. In addition, the estimates in *Case III* tend to have greater ESE's, SSE's and worse CPr's than the estimates in *Case II* when we deleting more values in the first two settings. This is because we ignore more information from missing values due to "unsure" responses in *Case III*. Moreover, it can be seen that, in *Case II* and *III*, estimates in setting (b) tends to be the worst and estimates in setting (c) exhibit the best performance as far as the CPr's and ESE's are concerned. Therefore, it can be concluded that ignoring missing values results in loss of efficiency And deleting all observations of those subjects with incomplete data causes severer loss.

To correct biases of the estimates caused by misclassification and loss of efficiency due to missing values, we use the corrected GQL estimating equations (4.49) to estimate the model parameters by taking missing values and misclassification into consideration. The results are reported *Case V* in Table 4.13. In this case, the results show that the corrected GQL approaches provide excellent parameter estimates, estimated standard errors and coverage probabilities. For example, we have estimates of parameter $\hat{\theta} = (1.026, 1.026, 1.514)'$ with CPr's (0.945, 0.961, 0.956) in setting (c) with most measurement errors, and $\hat{\theta} = (1.001, 0.988, 1.548)'$ with CPr's (0.958, 0.950, 0.942) in setting (b) with most "unsure" values. These estimates of parameters are very close to their true values $\theta = (1, 1, 1.5)'$ and the CPr's are also very close to the nominal level 0.95. In addition , the ESE's also have satisfactory performance compared with the corresponding SSE's.

Based on the above discussion and the comparisons, for example, Case I and IV with Case V, Case II and III with Case V, Case I and IV with Case II and III, we can conclude that ignoring misclassification leads to non-ignorable biases on estimates of model parameters, whereas neglecting missing values results in slight biases on estimates of parameters and significant biases on estimated standard errors of $\hat{\theta}$. Ignoring misclassification and deleting missing values of the subjects with incomplete data together lead to poor CPr's of confidence intervals, unless there are little misclas-

Table 4.13: Simulation results under GQL approach for imperfect data due to missing values and misclassification with the true value: $\theta = (1, 1, 1.5)$.

		(a)			(b)			(c)		
Case	Quantity	β_1	β_2	γ	β_1	β_2	γ	β_1	β_2	γ
I	SM	0.810	0.806	1.274	0.554	0.491	1.079	0.452	0.438	0.821
	SSE	0.117	0.273	0.312	0.102	0.259	0.314	0.097	0.217	0.256
	ESE	0.117	0.271	0.306	0.108	0.249	0.303	0.097	0.217	0.273
	CPr	0.618	0.888	0.864	0.016	0.468	0.684	0.000	0.244	0.302
II	SM	0.966	0.990	1.426	0.965	0.983	1.578	1.032	1.026	1.528
	SSE	0.144	0.329	0.351	0.193	0.470	0.482	0.218	0.467	0.420
	ESE	0.141	0.320	0.344	0.201	0.450	0.460	0.213	0.459	0.442
	CPr	0.941	0.940	0.936	0.962	0.942	0.930	0.944	0.948	0.961
III	SM	0.974	1.017	1.401	1.035	1.114	1.403	1.032	1.046	1.526
	SSE	0.199	0.414	0.450	0.321	0.680	0.798	0.219	0.505	0.493
	ESE	0.143	0.322	0.338	0.213	0.463	0.462	0.208	0.456	0.448
	CPr	0.862	0.884	0.832	0.856	0.882	0.828	0.956	0.936	0.928
IV	SM	0.860	0.809	1.381	0.581	0.488	1.147	0.456	0.437	0.833
	SSE	0.115	0.275	0.311	0.102	0.258	0.312	0.098	0.218	0.256
	ESE	0.116	0.271	0.306	0.107	0.250	0.304	0.097	0.217	0.273
	CPr	0.746	0.888	0.922	0.030	0.468	0.756	0.000	0.256	0.318
V	SM	1.022	1.008	1.529	1.001	0.988	1.548	1.026	1.026	1.514
	SSE	0.143	0.326	0.348	0.190	0.462	0.475	0.216	0.460	0.425
	ESE	0.142	0.325	0.341	0.199	0.450	0.457	0.206	0.451	0.435
	CPr	0.952	0.944	0.952	0.958	0.950	0.942	0.945	0.961	0.956
VI	SM	1.019	1.011	1.524	1.001	0.991	1.503	1.015	1.011	1.509
	SSE	0.119	0.277	0.313	0.120	0.289	0.322	0.120	0.278	0.281
	ESE	0.120	0.278	0.303	0.120	0.277	0.305	0.120	0.277	0.301
	CPr	0.948	0.953	0.954	0.956	0.949	0.946	0.948	0.956	0.952

sification and very few missing values. Our corrected GQL can effectively estimate model parameters and the corresponding standard errors as well as the confidence intervals with a specific nominal level.

Chapter 5

Modeling Mis-measured Longitudinal Count Data

5.1 Overview

In Chapter 2, we developed two models (2.15) and (2.17) to describe count errors in aggregated data. The two models can be used to characterize overcounted and undercounted data with imperfect sensitivity and specificity. Also, they can accommodate the perfect sensitivity or specificity or both of them. Actually, in some large population-based longitudinal studies, the follow-up observations of count responses are often contaminated with measurement errors. Analysis taking theses count errors into consideration in statistical inference is of great scientific interest. In this chapter, we apply models (2.15) and (2.17) to fit mis-measured count data from longitudinal processes.

5.2 Miscounted Binomial Count Data with Dynamic Population

5.2.1 Models

The model dealing with miscounted binomial data is given by

$$Y = \pi^+ * T + (1 - \pi^-) * (N - T), \tag{5.1}$$

where N is population size in an area, T is the true count of patients infected by an epidemic disease in this area, Y is the reported count of disease cases from a registration system, and π^+ and π^- are, respectively, the sensitivity and the specificity of the registration system. As mentioned before, in a longitudinal study lasting for several years, the annual population in an area generally changes over time due to birth, death, immigration and emigration. Therefore, it is reasonable to assume that the population size N is random, and it is following a longitudinal process. In this section, we assume that the population size follows a specified dynamic model.

Let n_{ij} be the observation of the population size N_{ij} of the ith district in the jth year. The true response T_{ij} describing the count of disease cases (e.g. infection by asthma) in this area can be assumed to follow the binomial model which is given by

$$T_{ij}|_{N_{ij}=n_{ij}} \sim b(n_{ij}, p_{ij}), \tag{5.2}$$

where

$$p_{ij} = \frac{exp(x'_{ij}\beta)}{1 + exp(x'_{ij}\beta)} \tag{5.3}$$

is the disease rate in the ith area during the jth year. The population size N_{ij} is from a dynamic model and its expectation and variance are, respectively, ϕ_{ij} and ς_{ij}^2. The inherent repones T_{ij} is a latent variable and it can not be observed directly. However, its surrogate Y_{ij} can be obtained from the registration system. The relationship between the true response T_{ij} and the observed response Y_{ij} can be described by the following expression:

$$Y_{ij} = \pi^+ * T_{ij} + (1 - \pi^-) * (N_{ij} - T_{ij}), \tag{5.4}$$

where π^+ is the sensitivity, and π^- is the specificity of the data collecting procedure. This expression can be used to model the count errors in longitudinal data Y_{ij}. Similar to the interpretation of model (2.15) in Chapter 2, the term $\pi^+ * T_{ij}$ in model (5.4) denotes the total of the infected people who are correctly reported as disease cases, and $(1 - \pi^-) * (N_{ij} - T_{ij})$ represents the number of people who are incorrectly reported as disease cases from the healthy population in the ith area during the jth year. $T_{ij} - \pi^+ * T_{ij}$ is the number of patients who are wrongly diagnosed as healthy, and $(N_{ij} - T_{ij}) - (1 - \pi^-) * (N_{ij} - T_{ij})$ is the total number of heathy people who are correctly counted into the healthy group.

In the binomial model (5.2-5.3), it is reasonable to assume that, for $1 \le j \ne u \le J$, and $i = 1, \ldots, I$, $T_{ij}|(N_{ij}, N_{iu}) \stackrel{d}{=} T_{ij}|N_{ij}$, and $T_{iu}|(N_{ij}, N_{iu}, T_{ij}) \stackrel{d}{=} T_{iu}|N_{iu}$. The latter assumption also means that, given N_{ij} and N_{iu}, T_{ij} will be independent of T_{iu}. In fact, Y_{ij} and Y_{iu} are also independent of each other conditional on N_{ij} and N_{iu}.

As discussed in Section 2.3, given the population size $N_{ij} = n_{ij}$, the observed response follows a binomial distribution $Y_{ij} \sim b(n_{ij}, q_{ij})$, where $q_{ij} = 1 - \pi^- + (\pi^- + \pi^+ - 1)p_{ij}$ is the reported disease rate of the ith area in the jth year based on the registration system. This means that we can skip the unobservable T_{ij} and build the direct relationship between Y_{ij} and N_{ij}. In fact, the dynamic patterns of both Y_{ij} and T_{ij} are determined by the longitudinal process of N_{ij}. Therefore, once we have defined a specific model for N_{ij}, it is easy to model Y_{ij} and T_{ij}. As a result, it is easy to calculate the moments of Y_{ij}. The expectation and variance of Y_{ij}, which are similar to the expressions (2.13) and (2.14) in Chapter 2, are given by

$$E(Y_{ij}|N_{ij}) = N_{ij}q_{ij} \tag{5.5}$$

$$\mu_{ij} = E(Y_{ij}) = \phi_{ij}q_{ij}, \tag{5.6}$$

and

$$Var(Y_{ij}) = \phi_{ij}q_{ij}(1-q_{ij}) + q_{ij}^2 Var(N_{ij}). \tag{5.7}$$

The expectation of pairwise product of Y_{ij} and Y_{iu} is formulated by

$$\begin{aligned} E(Y_{ij}Y_{iu}) &= E[E(Y_{ij}Y_{iu}|T_{iu}, T_{ij}, N_{iu}, N_{ij})] \\ &= E[E(Y_{ij}Y_{iu}|N_{ij}, N_{iu}] \\ &= E[q_{ij}q_{iu}N_{ij}N_{iu}] \\ &= q_{ij}q_{iu}E(N_{ij}N_{iu}). \end{aligned}$$

Hence, the covariance between Y_{ij} and Y_{iu} is

$$\begin{aligned} Cov(Y_{ij}, Y_{iu}) &= E(Y_{ij}Y_{iu}) - E(Y_{ij})E(Y_{iu}) \\ &= q_{ij}q_{iu}[E(N_{ij}N_{iu}) - \mu_{ij}\mu_{iu}] \\ &= q_{ij}q_{iu}Cov(N_{ij}, N_{iu}). \end{aligned} \tag{5.8}$$

These moments are very useful in the development of the GEE and GQL approaches for the estimation of model parameters.

To address the issue of estimation, it is assumed that the population sizes N_{ij} of district i during the studying period follows the linear transition model (LT) described in section 3.2.1.2 of Chapter 3. The LT model is given by

$$N_{ij}|N_{i,j-1} = n_{i,j-1} \sim Poisson(\phi_{i,j|j-1} = \xi_{ij} + \gamma n_{i,j-1}), \tag{5.9}$$

where

$$\xi_{ij} = exp(z_{ij}'\alpha). \tag{5.10}$$

z_{ij} represents some covariates related to the dynamics of the population size N_{ij}. The baseline observations are n_{i0}'s. From Subsection 3.2.3, it is easy to get the expectation and variance of N_{ij}.

$$\phi_{ij} = E(N_{ij}) = \gamma\phi_{ij-1} + \xi_{ij} = \sum_{v=u+1}^{j} \xi_{iv}\gamma^{j-v} + \gamma^{j-u}\phi_{iu} \tag{5.11}$$

with $\phi_{i0} = n_{i0}$, and

$$\varsigma_{ij}^2 = Var(N_{ij}) = \phi_{ij} + \gamma^2\varsigma_{ij-1}^2 = \sum_{v=u+1}^{j} \phi_{iv}\gamma^{2(j-v)} + \gamma^{2(j-u)}\varsigma_{iu}^2 \tag{5.12}$$

for $u < j$. $\varsigma_{i1}^2 = \phi_{i1}$ because $N_{i1} \sim$ Poisson(ϕ_{i1}). For convenience, we assume that $\varsigma_{i0}^2 = 0$. The covariance between N_{ij} and N_{iu} is

$$\mathrm{Cov}(N_{ij}, N_{iu}) = \gamma^2\varsigma_{iu}^2, \text{ for } u < j. \tag{5.13}$$

The correlation coefficient between N_{ij} and N_{iu} is given by

$$Corr(N_{iu}, N_{ij}) = min\{1,\ \gamma^{j-u}\frac{\varsigma_{iu}}{\varsigma_{ij}}\}.$$

5.2.2 Estimation of the model parameters

We apply the GEE and GQL methods to estimate the model parameters. In the GEE approach, we choose the independent correlation structure as the working correlation.

Suppose that the sensitivity π^+ and specificity π^- are known or their estimates can be obtained from prior experience or validation studies. The parameters of interest are $\theta = (\beta', \alpha', \gamma)'$, where β represents the effects of risk factors x_{ij} which are associated with the disease rate p_{ij} in (5.2), α describes the effects of covariates z_{ij} which is related to ϕ_i in the LT model (5.9) defined on the population size N_{ij}, and γ is the dynamic dependence parameter in the LT model.

To estimate the model parameters using the GEE approach, we solve the following estimating equations

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} W_i^{-1}(y_i - \mu_i) = 0, \tag{5.14}$$

where $y_i = (y_{i1}, \ldots, y_{iJ})'$ is the observation vector of response Y_i. In the equation (5.14), μ_i is the expectation of Y_i. W_i is the "working" covariance matrix and it can be written in the form of $W_i = \mathrm{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \ldots, \sigma_{iJ}^2)$, $\sigma_{ij}^2 = V(Y_{ij})$. $\partial \mu_i / \partial \theta$ is the first order derivative of μ_i with respect to θ.

The GEE approach is considered to be an effective procedure to estimate the model parameters in the situation that the true covariance matrix of Y_i, which is denoted by Σ_i, is unknown. However, if we know the true covariance matrix Σ_i in some cases, the GEE approach employing working covariance matrix W_i will lead to loss of efficiency, compared with the GQL approach that exploits the true covariance matrix Σ_i. The GQL estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{-1}(y_i - \mu_i) = 0, \tag{5.15}$$

and all the elements of Σ_i can be calculated based on expressions (5.6-5.8) and (5.10-5.13).

The expressions (5.6) and (5.11) lead to the following relationship between the expectations of Y_{ij} and Y_{ij-1}

$$\mu_{ij} = \gamma\mu_{ij-1} + q_{ij}\xi_{ij}. \tag{5.16}$$

Therefor, under the joint model (5.4) and (5.9), the elements of $\partial\mu_i/\partial\theta$ are given by

$$\begin{aligned}
\frac{\partial\mu_{ij}}{\partial\beta} &= (\pi^- + \pi^+ - 1)\frac{\partial p_{ij}}{\partial\beta}\phi_{ij} \\
&= (\pi^- + \pi^+ - 1)\phi_{ij}(1 - p_{ij})x_{ij}, & (5.17) \\
\frac{\partial\mu_{ij}}{\partial\alpha} &= q_{ij}\frac{\partial\phi_{ij}}{\partial\alpha} \\
&= q_{ij}\left(\gamma\frac{\partial\phi_{ij-1}}{\partial\alpha} + \xi_{ij}\right)z_{ij} \\
&= \left(\gamma\frac{\partial\mu_{ij-1}}{\partial\alpha} + q_{ij}\xi_{ij}\right)z_{ij} & (5.18)
\end{aligned}$$

for $j = 1, \ldots, J$, and

$$\begin{aligned}
\frac{\partial\mu_{ij}}{\partial\gamma} &= q_{ij}\frac{\partial\phi_{ij}}{\partial\gamma} \\
&= q_{ij}\left(\gamma\frac{\partial\phi_{ij-1}}{\partial\gamma} + \phi_{ij-1}\right) \\
&= \gamma\frac{\partial\mu_{ij-1}}{\partial\gamma} + \mu_{ij-1} & (5.19)
\end{aligned}$$

for $j = 1, \ldots, J$.

Once we have the estimates of θ under the GEE and GQL approaches, say $\hat{\theta}_{GEE}$ and $\hat{\theta}_{GQL}$, their corresponding covariance matrix can be consistently estimated by

$$\hat{V}(\hat{\theta}_{GEE}) = \left[\sum_{i=1}^{I}\frac{\partial\mu_i'}{\partial\theta}W_i^{-1}\frac{\partial\mu_i}{\partial\theta}\right]|_{\theta=\hat{\theta}_{GEE}}, \tag{5.20}$$

and

$$\hat{V}(\hat{\theta}_{GQL}) = \left[\sum_{i=1}^{I}\frac{\partial\mu_i'}{\partial\theta}\Sigma_i^{-1}\frac{\partial\mu_i}{\partial\theta}\right]|_{\theta=\hat{\theta}_{GQL}}, \tag{5.21}$$

respectively.

One may notice that even we do not know the specific dynamic model of $\{N_{ij}\}$, the GEE and GQL approaches are still feasible as long as the first and the second moments of $N_i = (N_{i1}, N_{i2}, \ldots, N_{im})'$ are known. More practical applications of those approaches can be expected due to this flexibility.

5.2.3 Simulation studies

We have two objectives of the simulation studies in this subsection. The first one is to examine the attenuations of the naive estimates of model parameters due to ignoring the measurement errors. The second one is to check the performance of the corrected GEE and GQL approaches in correcting the attenuation by taking account errors into consideration when using the measurement error model (5.4).

The parameters of interest include β in the binomial model (5.2-5.3), and (α', γ) in the LT model (5.9). The sensitivity and the specificity (π^+, π^-) in the count error model (5.4) are assumed to be known in simulations. The covariates will be selected as those in the simulation design in Section 5.2.3.1. Then the data of N, T, and Y can be sampled following the data generation procedure given in Section 5.2.3.2. The parameters will then be estimated by solving the GEE and GQL estimating equations (5.14) and (5.15), respectively. Finally, the GEE and GQL estimates are summarized in Table 6.1 and Table 6.2 from 500 simulation runs.

5.2.3.1 Covariate design

We consider $I = 100$ independent regions each with $J = 4$ repeated observations for the count responses N, T, and Y. As far as the time dependent covariates are concerned, we consider the following design. The covariates x_{ij} involved in the true disease rate model (5.3) are given by

$x_{ij(1)} = 1$ for $j = 1, 2, 3, 4$ and $i = 1, 2, \ldots, 100$,

$x_{ij(2)} = \sin(\frac{\pi}{6}j)$, for $j = 1, 2, 3, 4$ and $i = 1, 2, \ldots, 100$,

which implies that

$$p_{ij} = \frac{exp(\beta_1 + \beta_2 sin(\frac{\pi}{6}j))}{1 + exp(\beta_1 + \beta_2 sin(\frac{\pi}{6}j))}. \tag{5.22}$$

The covariates z_{ij} related to the population size in the LT model (5.9) are given by

$z_{ij(1)} \sim log(P(10 * j)),\ j = 1, 2, 3, 4, i = 1, 2, \ldots, 50;$

$z_{i\cdot(1)} \sim log(P(20 * j),\ j = 1, 2, 3, 4, i = 51, 52, \ldots, 100.$

$z_{i\cdot(2)} = (1, 1, 0, 1), i = 1, 2, \ldots, 50;$

$z_{i\cdot(2)} = (1, 0, 1, 1), i = 51, 52, \ldots, 100.$

$z_{ij(3)} \sim N((j-1)/4, 0.5^2), j = 1, 2, 3, 4,\ i = 1, 2, \ldots, 100.$

5.2.3.2 Data generation

We choose two different settings for the baseline observations of the population size n_{i0} in the LT model (5.9), that is, $\phi_0 = 200$ and 10. The first setting produces large population size n_{ij}, whereas the latter yields small values of n_{ij}. By assigning the true values of parameters as $\beta = (-2.50, 0.50)'$, $\alpha = (1.00, -1.00, 1.00)'$, $\gamma = 0.85$, and $(\pi^+, \pi^-) = (0.75, 0.90)$, the data of n_{ij}, t_{ij}, and y_{ij} can be generated from the procedure described below.

1. The baseline observations of population sizes are generated from the Poisson model, say $N_{i0} \overset{iid}{\sim} P(\phi_0)$, $i = 1, 2, \ldots, 100$.

2. n_{ij} are generated from the linear transition model (5.9): $N_{ij}|N_{i,j-1} = n_{i,j-1} \sim P(\gamma n_{i,j-1} + \xi_{ij})$, where $\xi_{ij} = exp(z_{ij}'\alpha)$.

3. t_{ij} is sampled from the binomial model $b(n_{ij}, p_{ij})$, where p_{ij} is given by (5.22).

4. y_{ij} is generated from the measurement error model (5.4). We first sample U_{ij} from $b(t_{ij}, \pi^+)$ and V_{ij} from $b(n_{ij} - t_{ij}, 1 - \pi^-)$, then we calculate y_{ij} by adding U_{ij} and V_{ij} up, that is $y_{ij} = U_{ij} + V_{ij}$.

5.2.3.3 Simulation results

In this subsection, we consider two settings of baseline observations of the population size n_{i0}. Simulation results under the first setting with $n_{i0} \overset{iid}{\sim} Poisson(\phi_0 = 200)$ are given in Table 5.1, and the results under the second setting with $n_{i0} \overset{iid}{\sim} Poisson(\phi_0 = 10)$ are presented in Table 5.2.

From the two tables, it can be seen that ignoring measurement errors in the count data leads to significant biases of the estimates of parameters β involved in the model (5.2-5.3) which directly defines the inherent response T. The biased estimates subsequently result in poor performance of coverage probabilities (CPr's) of CI's with confidence level 95%. For example, in Table 5.1, the simulated mean (SM) of the naive GEE estimate of β_1 is -1.7521 which is far away from its true value -2.5, and the CPr=0.000 is extremely poor compared with the nominal level 95%. Similar phenomena also happen in Table 5.2 with small population size, i.e. $\phi_0 = 10$. Therefore, we conclude that ignoring measurement error in count data leads to significant attenuation on the estimates of effects of covariates in the binomial model (5.2-5.3) of the true count response T_{ij}. As a result, misleading evaluation of the true disease rate p_{ij} is obtained.

It is surprising that the both the naive GEE and GQL estimates of the parameters α and γ which are from the LT model (5.9) and (5.10) defined for the population sizes N_{ij}'s, are very close to their true values. For instance, in Table 5.1, the naive GEE estimate of α is $\hat{\alpha}_{NGEE} = (1.0017, -1.0076, 0.9930)'$, and the naive GEE estimate of γ is 0.8484. They are quite close to their true values $\alpha = (1, -1, 1)'$ and $\gamma = 0.85$. The naive GQL method has similar results. To check the performance of the naive estimates of α in the situation that the count error $\mid t_{ij} - y_{ij} \mid$ is large relative to the total number of studying subjects n_{ij}, we conduct 500 simulations by setting up a small $\phi_0 = 10$ of which the results are reported in Table 5.2. In the simulation, when the ratios $\mid t_{ij} - y_{ij} \mid /n_{ij}$ sometimes are greater than 20%, the naive estimates of α and γ under the GEE and GQL approaches are still acceptable. Therefore, we can conclude that ignoring misclassification does not leads to biased estimates of parameters in models (5.9) and (5.10). This is because that whatever the error-prone data y_{ij} or the true count data t_{ij} are used in estimation, the total number of people n_{ij} in the study keeps the same, which can be expressed by

population size (n_{ij})

$=$*true count of disease cases*(t_{ij}) + true number of healthy *people*$(n_{ij} - t_{ij})$

$=$*reported count of disease cases*(t_{ij}) + *reported number of healthy people*$(n_{ij} - y_{ij})$.

As far as the corrected GEE and GQL approaches are concerned, Table 5.1 and Table 5.2 show that they can effectively estimate all the model parameters. The attenuation on the naive estimates of β in the function p_{ij} can be well adjusted under the corrected GEE and GQL approaches. As an example, in Table 5.1, $\hat{\beta}_{CGEE} = (-2.4937, 0.4933)$ and $\hat{\beta}_{CGQL} = (-2.4940, 0.4933)$, which are very close to their true values $\beta = (-2.5, 0.5)$. Similarly, the biases of corrected GEE and GQL estimates of α and γ are ignorable. In addition, it can also be seen from Table 5.1 and Table 5.2 that the estimated standard errors (ESE) of corrected GEE and GQL estimates based on (5.20) and (5.21), respectively, are very close to the corresponding simulated standard errors (SSE).

As mentioned in Section 5.2.2, the GEE approach borrowing the "working" independence correlation structure tends to yield less efficient estimates compared with the GQL approach employing the true correlation structure. Many researchers made this conclusion for error-free data. Most of our simulation results in Table 5.1 and Table 5.2 about the estimated snd simulated standard errors and coverage probabilities (CPr) of 95% confidence interval also support this conclusion. For example, in most of cases, the SEE's and ESE's of GEE estimates tend to be smaller than those of GQL estimate.

As far as the efficiencies of the corrected estimates are concerned, the corrected GEE approach also leads to loss of efficiency. It can be seen that in most cases of Table 5.1 and Table 5.2, the CPr's under the corrected GQL approach are closer to the true nominal level 0.95 than the CPr's under the corrected GEE approach. For example, in Table 5.1, the CPr's of β are (0.956, 0.970) under the corrected GEE approach, and they are (0.948, 0.966) under the corrected GQL approach. Actually, it can be seen that, in most cases, the naive estimates under the GQL approach also tend to have higher efficiency than the naive GEE estimates, especially when parameters α and γ are concerned.

Table 5.1: Simulation results of GEE and GQL approaches based on the count error model (6.4), the binomial model (6.2-6.3) and the LT model (6.9) with $(\beta_1, \beta_2) = (-2.50, 0.50)$, $\alpha = (1.00, -1.00, 1.00)$, $\gamma = 0.85$ and $(\pi^+, \pi^-) = (0.75, 0.90)$ under the setting $\phi_0 = 200$.

	GEE			GQL		
Quantity	Ideal	Naive	Corrected	Ideal	Naive	Corrected
SM(β_1)	-2.4995	-1.7521	-2.4937	-2.4995	-1.7522	-2.4940
SSE	0.0674	0.0544	0.1275	0.0672	0.0542	0.1272
ESE	0.0688	0.0588	0.1354	0.0670	0.0562	0.1311
CPr	0.960	0.000	0.956	0.952	0.000	0.948
SM(β_2)	0.5022	0.2219	0.4933	0.5022	0.2219	0.4933
SSE	0.1241	0.1070	0.2240	0.1227	0.1055	0.2213
ESE	0.1347	0.1188	0.2473	0.1285	0.1102	0.2287
CPr	0.960	0.344	0.970	0.960	0.278	0.966
SM(α_1)	1.0012	1.0017	1.0018	1.0011	1.0017	1.0017
SSE	0.0403	0.0339	0.0339	0.0392	0.0326	0.0326
ESE	0.0394	0.0325	0.0325	0.0408	0.0338	0.0338
CPr	0.936	0.930	0.930	0.950	0.952	0.950
SM(α_2)	-1.0084	-1.0076	-1.0073	-1.0085	-1.0081	-1.0076
SSE	0.1508	0.1243	0.1245	0.1474	0.1187	0.1189
ESE	0.1492	0.1238	0.1239	0.1516	0.1238	0.1257
CPr	0.944	0.945	0.947	0.952	0.953	0.952
SM(α_3)	0.9949	0.9930	0.9932	0.9951	0.9933	0.9935
SSE	0.1106	0.0913	0.0913	0.1086	0.0887	0.0888
ESE	0.1037	0.0862	0.0862	0.1118	0.0948	0.0949
CPr	0.920	0.932	0.932	0.948	0.964	0.965
SM(γ)	0.8487	0.8484	0.8488	0.8488	0.8485	0.8490
SSE	0.0301	0.0252	0.0251	0.0300	0.0247	0.0245
ESE	0.0311	0.0260	0.0258	0.0305	0.0254	0.0252
CPr	0.962	0.956	0.952	0.956	0.948	0.948

Table 5.2: Simulation results of GEE and GQL approaches based on the count error model (5.4), the binomial model (5.2-5.3) and the LT model (5.9) with $(\beta_1, \beta_2) = (-2.50, 0.50)$, $\alpha = (1.00, -1.00, 1.00)$, $\gamma = 0.85$ and $(\pi^+, \pi^-) = (0.75, 0.90)$ under the setting $\phi_0 = 10$.

	GEE			GQL		
Quantity	Ideal	Naive	Corrected	Ideal	Naive	Corrected
SM(β_1)	-2.5048	-1.7776	-2.5651	-2.5042	-1.7771	-2.5647
SSE	0.1834	0.1493	0.3721	0.1830	0.1495	0.3722
ESE	0.1886	0.1601	0.3938	0.1878	0.1556	0.3848
CPr	0.950	0.002	0.974	0.948	0.002	0.968
SM(β_2)	0.4940	0.2434	0.5393	0.4932	0.2428	0.5386
SSE	0.2396	0.1954	0.4270	0.2369	0.1948	0.4252
ESE	0.2483	0.2155	0.4789	0.2427	0.2079	0.4591
CPr	0.970	0.790	0.974	0.964	0.790	0.956
SM(α_1)	1.0026	0.9997	1.0006	1.0027	0.9998	1.0008
SSE	0.0411	0.0332	0.0329	0.0405	0.0324	0.0321
ESE	0.0383	0.0312	0.0308	0.0396	0.0326	0.0321
CPr	0.930	0.930	0.930	0.942	0.944	0.942
SM(α_2)	-1.0052	-0.9942	-0.9947	-1.0047	-0.9958	-0.9963
SSE	0.0922	0.0784	0.0782	0.0914	0.0773	0.0772
ESE	0.0940	0.0769	0.0767	0.0949	0.0776	0.0775
CPr	0.952	0.950	0.952	0.956	0.952	0.954
SM(α_3)	1.0007	0.9992	0.9993	1.0006	0.9992	0.0992
SSE	0.0554	0.0490	0.0490	0.0552	0.0487	0.0487
ESE	0.0507	0.0417	0.0417	0.0533	0.0447	0.0447
CPr	0.924	0.906	0.906	0.938	0.924	0.924
SM(γ)	0.8493	0.8491	0.8498	0.8493	0.8496	0.0850
SSE	0.0605	0.0512	0.0513	0.0600	0.0507	0.0508
ESE	0.0604	0.0496	0.0497	0.0599	0.0489	0.0490
CPr	0.954	0.940	0.938	0.954	0.944	0.942

In conclusion, ignoring measurement errors in count response results in remarkable attenuation on the estimates of the effects of covariates in the true disease rate p_{ij} (5.3), which is used to define the model of the true response T_{ij}. However, count errors do not influence the statistical inference on the effects of covariates associated with the population size N_{ij} in the LT model (5.9).

Based on the count error model (5.4), the binomial model (5.2-5.3) for true count response T_{ij}, and the LT model (5.2) for population size N_{ij}, our corrected GEE and GQL approaches can consistently estimate all model parameters. In addition, the independence GEE approach leads to loss of efficiency in estimation due to the choice of a "working" covariance structure. Therefore, if the true covariance matrix is available, the GQL approach using the true covariance structure can improve the performance of the inference.

5.3 Miscounted Longitudinal Data with Little Information about Population Size

5.3.1 The model

In some situation, it is impossible to know the population size of an area and its first order moment, for example, the population size in the years far from the censal years. Suppose that there are two censuses in years 1990 and 2000, respectively. The population sizes in the intercensal years close to 1990 or 2000, such as 1991, 1992, are easy to model based on the information of the censal years. But the population sizes in 1994, 1995, 1996 may be difficult to model. The information of censal year may not be helpful due to widely and complicated migration, birth and death. Although, in this case, it is reasonable to assume that the population size in one of these years follows a Poisson distribution, however, we do not know anything more. This means that we do not have any knowledge about the expectation of the distribution, and we also have no idea about the correlation structure between population sizes in different years. In this situation, model (5.4) is not valid for modeling miscounted disease cases among the population. And the corrected additive count error model (2.17) developed in Chapter 2 appears to be an appropriate alternative.

We assume that T_{ij} is the true count of subjects who are infected by a kind of epidemic disease in the ith area during the jth year, $i = 1, \dots, I$ and $j = 1, \dots J$. And Y_{ij} is the corresponding total of reported disease cases from a surveillance system. The corrected additive count error model describing the relationship between Y_{ij} and T_{ij} is given by

$$Y_{ij} = \pi^+ * T_{ij} + e_{ij}, \tag{5.23}$$

where π^+ is the sensitivity of the surveillance system. The term $\pi^+ * T_{ij}$ represents the total number of patients who are correctly reported by the surveillance system. The true size of the infected population T_{ij} is assumed to follow a specific dynamic model, and it may be associated with some risk factors x_{ij}, for example, the environmental exposures. e_{ij} denotes the number of healthy people who are incorrectly reported as disease cases, and it is assumed to be independent of T_{ij}, hence independent of $\pi^+ * T_{ij}$. However, given a specific district i, e_{ij}'s may be correlated due to the fact that they are from the same district. Furthermore, e_{ij} is assumed to be a Poisson variable with expectation $\psi_{ij} = exp(z_{ij}'\alpha)$, where z_{ij} are covariates related to the miscounted subjects from the healthy population, for example, the health care level of an area. In some cases, x_{ij} and z_{ij} may share some common covariates.

Similar to the calculations in Section 2.4.2, we can write the expectation of Y_{ij} in the following expression

$$E(Y_{ij}) = \mu_{ij} = \pi^+ \eta_{ij} + \psi_{ij}, \tag{5.24}$$

where η_{ij} is the expectation of the true count of disease cases T_{ij}. The variance of Y_{ij} is given by

$$Var(Y_{ij}) = \mu_{ij} + (\pi^+)^2 [Var(T_{ij}) - \eta_{ij}]. \tag{5.25}$$

The expectation of the pairwise product of Y_{ij} and Y_{iu} can be expressed by

$$\begin{aligned} E(Y_{ij}Y_{iu}) &= E[(\pi^+ * T_{ij} + e_{ij})(\pi^+ * T_{iu} + e_{iu})] \\ &= (\pi^+)^2 E(T_{ij}T_{iu}) + \pi^+(\eta_{ij}\psi_{iu} + \psi_{ij}\eta_{iu}) + E(e_{ij}e_{iu}). \end{aligned}$$

Therefore, the covariance between Y_{ij} and Y_{iu} is given by

$$Cov(Y_{ij}, Y_{iu}) = (\pi^+)^2 Cov(T_{ij}, T_{iu}) + Cov(e_{ij}, e_{iu}). \tag{5.26}$$

As far as the model of T_{ij} is concerned, we assume that the true count of disease cases T_{ij} follows the LT model described in Section 3.2.2 of Chapter 3, which is given by

$$T_{ij}|_{T_{i,j-1}=t_{i,j-1}} \sim Poisson(\eta_{ij}^c = \xi_{ij} + \gamma t_{i,j-1}), \tag{5.27}$$

with

$$\xi_{ij} = exp(x_{ij}'\alpha). \tag{5.28}$$

The baseline observations are t_{i0}'s. From Section 3.2.2, it is easy to get the expectation and variance of T_{ij}. They are given by

$$\eta_{ij} = E(T_{ij}) = \gamma\eta_{ij-1} + \xi_{ij} = \sum_{v=u+1}^{j} \xi_{iv}\gamma^{j-v} + \gamma^{j-u}\eta_{iu} \tag{5.29}$$

with $\eta_{i0} = t_{i0}$, and

$$\varsigma_{ij}^2 = Var(T_{ij}) = \eta_{ij} + \gamma^2\varsigma_{ij-1}^2 = \sum_{v=u+1}^{j} \eta_{iv}\gamma^{2(j-v)} + \gamma^{2(j-u)}\varsigma_{iu}^2 \tag{5.30}$$

for $u < j$. Let $\varsigma_{i0}^2 = 0$, and therefore $\varsigma_{i1}^2 = \eta_{i1}$. The covariance between T_{ij} and T_{iu}, is

$$\mathrm{Cov}(T_{ij}, T_{iu}) = \gamma^2\varsigma_{iu}^2, \text{ for } u < j, \tag{5.31}$$

and the correlation coefficient between T_{ij} and T_{iu} is given by

$$Corr(T_{iu}, T_{ij}) = min\{1,\ \gamma^{j-u}\frac{\varsigma_{iu}}{\varsigma_{ij}}\}.$$

5.3.2 Estimation of the model parameters

In this subsection, we apply the GEE and GQL methods to estimate the model parameters. The interested parameters are $\theta = (\beta', \gamma, \alpha')'$ in the case that the sensitivity π^+ is known or its estimate can be obtained from prior knowledge or validation studies, while the interested parameters are $\theta = (\beta', \gamma, \alpha', \pi^+)'$ in the case that the sensitivity π^+ is unknown. β represents the effects of risk factors x_{ij} which are associated with the true count of disease cases in district i during the jth year, and γ is the dynamic parameter in model (5.27-5.28). α describes the effects of covariates z_{ij} which is related to e_{ij}, the total of false disease cases reported by the surveillance system. π^+ is the sensitivity of the surveillance system.

In the GEE approach, we choose the independent correlation structure as the working correlation. The estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} W_i^{-1}(y_i - \mu_i) = 0, \tag{5.32}$$

where $y_i = (y_{i1}, \ldots, y_{iJ})'$ is the observation vector of Y_i, and μ_i is the expectation of Y_i. Under the independence correlation structure, $W_i = \text{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \ldots, \sigma_{iJ}^2)$, where $\sigma_{ij}^2 = Var(Y_{ij})$. $\partial\mu_i/\partial\theta$ is the first order derivative of μ_i with respect to θ.

To improve the efficiency of the estimation, we use the GQL approach by exploiting the true covariance matrix of Y_i which is denoted by Σ_i. The GQL estimating equations are given by

$$\sum_{i=1}^{I} \frac{\partial\mu_i'}{\partial\theta} \Sigma_i^{-1}(y_i - \mu_i) = 0, \tag{5.33}$$

where Σ_i can be calculated based on formulas (5.24-5.26) and (5.29-5.31)).

Under the model assumptions, we can calculate the first order derivative of μ_{ij} with respect to θ. Here, if the sensitivity π^+ is known or can be estimated from other studies, the parameters of interest are $\theta = (\beta', \alpha', \gamma)'$. If π^+ is unknown, then one may be interested in estimating $\theta = (\beta', \alpha', \gamma, \pi^+)'$. The first order derivatives are

$$\begin{aligned}
\frac{\partial\mu_{ij}}{\partial\beta} &= \pi^+\frac{\partial\eta_{ij}}{\partial\beta} = \pi^+[\gamma\frac{\partial\eta_{ij-1}}{\partial\beta} + \xi_{ij}x_{ij}] \\
&= \gamma\frac{\partial\mu_{ij-1}}{\partial\beta} + \pi^+\xi_{ij}x_{ij}, && (5.34) \\
\frac{\partial\mu_{ij}}{\partial\gamma} &= \pi^+\frac{\partial\eta_{ij}}{\partial\gamma} = \pi^+[\gamma\frac{\partial\eta_{ij-1}}{\partial\gamma} + \eta_{ij-1}] \\
&= \gamma\frac{\partial\mu_{ij-1}}{\partial\gamma} + \pi^+\eta_{ij-1}, && (5.35) \\
\frac{\partial\mu_{ij}}{\partial\alpha} &= \psi_{ij}z_{ij}, && (5.36) \\
\frac{\partial\mu_{ij}}{\partial\pi^+} &= \eta_{ij}, && (5.37)
\end{aligned}$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, I$. Notice that the zero baseline observations of T_{ij}, that is $t_{i0} = 0$, lead to $\partial\eta_{i1}/\partial\gamma = 0$, and therefore, $\partial\mu_{i1}/\partial\gamma = 0$.

Once we have the estimates from the GEE and GQL approaches, that is $\hat{\theta}_{GEE}$ and $\hat{\theta}_{GQL}$, the corresponding consistent estimators of their covariance matrices are given by

$$\hat{V}(\hat{\theta}_{GEE}) = \left[\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} W_i^{-1} \frac{\partial \mu_i}{\partial \theta}\right] |_{\theta=\hat{\theta}_{GEE}}, \tag{5.38}$$

and

$$\hat{V}(\hat{\theta}_{GQL}) = \left[\sum_{i=1}^{I} \frac{\partial \mu_i'}{\partial \theta} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \theta}\right] |_{\theta=\hat{\theta}_{GQL}}, \tag{5.39}$$

respectively.

5.3.3 Numerical study

Similar to the simulation study in Subsection 5.2.2, we conduct a finite sample simulation to examine the performance of the naive and corrected estimates based on the GEE and GQL approaches. To be specific, we apply the GEE estimating equations (5.32), and the GQL estimating equations (5.33) to the randomly generated data to estimate the model parameters θ. In the simulation, we consider two cases for θ. The first one is the case with known π^+, we estimate $\theta = (\beta', \gamma, \alpha')'$, whereas in the second case that π^+ is unknown, we estimate $\theta = (\beta', \gamma, \alpha', \pi^+)'$. We have 500 simulation runs for the data generation and parameter estimation.

5.3.3.1 Covariate design

We consider $I = 60$ independent districts each with $J = 4$ repeated count responses T and Y. As far as the choice of the true values of the parameters are concerned, we consider $\beta = (0.6, -1.0, 1.0)'$, $\alpha = (0.30, -0.50)'$, $\gamma = 0.8$ and 0.3, and $\pi^+ = 0.7$ and 0.85. The covariate values for the simulation are given in the following paragraphs.

The first covariates is assumed to be a variable related to the population size.

Suppose that we randomly divide the 60 districts into 6 groups each consists of 10 districts. Each group is assigned a set of growth rates of population (pgr) which are given by

$$\begin{aligned} pgr[1,] &= (1, 1.01, 1.01^2, 1.01^3), \\ pgr[2,] &= (1, 0.99, 0.99^2, 0.99^3), \\ pgr[3,] &= (1, 1.01, 0.99, 0.99^2), \\ pgr[4,] &= (1, 0.99, 1.01, 1.01^2), \\ pgr[5,] &= (1, 1.01, 0.99, 1.01), \\ pgr[6,] &= (1, 0.99, 1.01, 0.99). \end{aligned}$$

The time-varying population sizes are generated following the procedure described below:

$$pop[(g-1)10+k, j] \sim Poisson(1000(g+k)pgr[g, j]), g = 1, \ldots, 6, k = 1, \ldots, 10, j = 1, 2, 3, 4.$$

Then we randomly order the rows of matrix *pop* by the code

$$population < -pop[sample(60, 60),]$$

in statistical package **R**. Then the first covariate $x_{ij(1)}$ is defined as

$$x_{ij1} \sim log(population[i, j]), \text{ for } j = 1, 2, 3, 4, i = 1, 2, \ldots, 60.$$

Other covariates of x_{ij} are defined as

$$\begin{aligned} x_{i\cdot(2)} &= (1, 1, 0, 0), i = 1, 2, \ldots, 30; \\ x_{i\cdot(2)} &= (0, 0, 1, 1), i = 31, 32, \ldots, 60. \\ x_{ij(3)} &\sim N((j-1)/4, 0.5^2), j = 1, 2, 3, 4, \ i = 1, 2, \ldots, 60. \end{aligned}$$

The following covariates z_{ij} are defined on e_{ij}. We assume that z_{ij} shares the first element of x_{ij}. So, the design for z_{ij} is given as

$$z_{ij(1)} = x_{ij1}, j = 1,2,3,4, i = 1,2,\ldots,60.$$

$$z_{ij(2)} \sim N(1, 0.5^2), j = 1,2, i = 1,2,\ldots,30;$$

$$z_{ij(2)} \sim N(1, 0.65^2), j = 3,4, i = 1,2,\ldots,30;$$

$$z_{ij(2)} \sim N(1, 0.55^2), j = 1,2, i = 31,32,\ldots,60;$$

$$z_{ij(2)} \sim N(1, 0.7^2), j = 3,4, i = 31,32,\ldots,60.$$

5.3.3.2 Data generation

We set the baseline observations for the population sizes $t_{i0} = 0$ in the LT model (5.27). The data T_{ij}, and Y_{ij} can be generated following the procedure described below:

1. The true count t_{ij} are generated based on the linear transition model (5.27) with $t_{i0} = 0$: $T_{ij}|_{T_{ij-1}=t_{ij-1}}$ follows $Poisson(\gamma t_{ij-1} + exp(x'_{ij}\alpha))$.

2. y_{ij} is generated from the corrected additive error model (5.23).

   (1). First,we generate $U_{ij} \sim b(t_{ij}, \pi^+)$.

   (2). For simplicity, we generated independent additive errors e_{ij} given i, that is, $e_{ij} \sim P(exp(z'_{ij}\alpha))$.

   (3). The observed count data $y_{ij} = U_{ij} + e_{ij}$.

5.3.3.3 Simulation results

In this subsection, we consider two sets of values of parameters $\theta = (\beta', \gamma, \alpha', \pi^+)'$, where $\beta = (\beta_1, \beta_2, \beta_3)'$ are effects of covariates x_{ij} associated with the true count of disease cases in the LT model (5.27), γ is the dynamic dependence parameter in the LT model, $\alpha = (\alpha_1, \alpha_2)'$ represent the effects of explanatory variable z_{ij} in the count error model (5.23), and π^+ is the sensitivity in model (5.23). One set of values of θ is $\beta = (0.6, -1.0, 1.0)$, $\gamma = 0.8$, $\alpha = (0.3, -0.5)$ and $\pi^+ = 0.7$, and the other set is $\beta = (0.6, -1.0, 1.0)$, $\gamma = 0.3$, $\alpha = (0.3, -0.5)$ and $\pi^+ = 0.85$. For each set of values of parameters, we consider four types of estimates: ideal, naive, corrected1 and corrected2 estimates under both GEE and GQL approaches. Among these four types of estimates, the ideal estimates are obtained by using the data of true response T_{ij}, the naive estimates are based on the error-contaminated data y_{ij} ignoring the measurement errors; the corrected1 estimation implies that all parameters including the sensitivity π^+ are estimated by employing the observed data y_{ij} and taking errors into consideration under the assumption that π^+ is unknown, and the last one, i.e. the corrected2 estimates, means that all parameters except π^+ are estimated based on y_{ij} taking count errors into account for known π^+. Notice that, in the ideal and naive frameworks, the parameters need to estimate only consist of β and γ. The simulation results are presented in Table 5.3 and Table 5.4.

From Table 5.3 and Table 5.4, it is easy to see that both naive GEE and GQL estimates of model parameters have significant biases due to ignoring count errors in the observed data. For example, for the true value of $\beta_2 = -1.0$ in Table 5.3, we estimate it as $\hat{\beta}_{2(GEE)} = -0.9378$ and $\hat{\beta}_{2(GQL)} = -0.9412$ with bias more than 5%.

Especially, the coverage probabilities of 95% CI's for the naive estimates of β and γ are considerably biased from the nominal level 0.95. For instance, the CPr's for $(\hat{\beta}'_{NGEE}, \hat{\gamma}_{NGEE})$ are (0.000, 0.020, 0.012, 0.014), and the CPr's for $(\hat{\beta}'_{NGQL}, \hat{\gamma}_{NGQL})$ are (0.000, 0.018, 0.010, 0.024), while the specified nominal level is 0.95.

To improve the unsatisfactory naive estimates, we apply the corrected GEE and GQL approaches to estimate the model parameters and construct confidence intervals. Two types of estimates, the corrected1 estimates for unknown sensitivity π^+ and the corrected2 estimates for known π^+ are computed, and the simulation results are given in Table 5.3 and Table 5.4. It can be seen that both corrected1 and corrected2 estimates under the GEE and GQL approaches produce tiny biases which can be neglected. For example, for the parameters $(\beta', \gamma)'$ with true values $(0.6, -1.0, 1.0, 0.8)$ in Table 5.3, the first corrected GEE estimates are $(\hat{\beta}'_{CGEE1}, \hat{\gamma}_{CGEE1}) = (0.5994, -0.9995, 0.9997, 0.8001)$, and the second corrected GEE estimates are $(\hat{\beta}'_{CGEE2}, \hat{\gamma}_{CGEE2}) = (0.6000 - 0.9997, 0.9998, 0.8002)$. Similarly, the corrected1 GQL estimates are $(\hat{\beta}'_{CGQL1}, \hat{\gamma}_{CGQL1}) = (0.5994, -0.9993, 0.9998, 0.7999)$, and the corrected2 GQL estimates are $(\hat{\beta}'_{CGQL1}, \hat{\gamma}_{CGQL1}) = (0.6000, -0.9994, 0.9999, 0.8000)$. Under the corrected approaches, we also obtain excellent estimates of parameters $\alpha = (\alpha_1, \alpha_2)'$ which are defined on e_{ij}, the number of false disease cases miscounted from the healthy population. For instance, in Table 5.3, for the true $\alpha = (0.3, -0.5)$, we have $\hat{\alpha}_{CGEE1} = (0.2980, -0.5517)'$, $\hat{\alpha}_{CGEE2} = (0.2982, -0.5685)'$ under the GEE approach, and $\hat{\alpha}_{CGQL1} = (0.2981, -0.5557)'$, $\hat{\alpha}_{CGQL2} = (0.2983, -0.5714)'$ under the GQL approach. Beside the parameters mentioned above, we also obtained approximately unbiased estimates of the sensitivity π^+. For example, with the true value $\pi^+ = 0.7$ in Table 5.3, we get $\hat{\pi}_{CGEE1} = 0.7113$ and $\hat{\pi}_{CGQL1} = 0.7115$, which are very close to the true value 0.7. From these examples, we can also see that the second corrected estimates assuming a known π^+ often perform better than the first corrected estimates in the case of unknown π^+. It should be pointed out that the estimation about the error-related parameters α and π^+ is of greatly scientific interest. They can be used to evaluate the severity of measurement errors in the collected data and further to improve the reliability of the data collecting procedures such as surveillance programs and registration systems.

Similar to the discussions in the simulation study in Section 5.2.3.3, the ideal and the two types of corrected estimates under the GQL approach tend to have higher efficiency than those estimates under the GEE approach in estimating model parameters. This is because the corrected GQL approach uses the true correlation structure in its estimating equations (5.33), while the GEE method uses a working independence covariance matrix. It can be seen that, in most cases in Table 5.3 and Table 5.4, the CPr's under the corrected GQL approach are closer to the nominal level 0.95 than the CPr's under the corrected GEE approach. For example, the CPr's of β are (0.926, 0.956, 0.930) under the corrected1 GEE approach, and (0.964, 0.954, 0.942) under the corrected1 GQL approach in Table 5.4. Similarly, in the same table, the CPr's of β are (0.940, 0.954, 0.940) under the corrected2 GEE approach, and (0.948, 0.952,0.953) under the corrected2 GQL approach.

In summary, ignoring measurement errors in count data leads to non-ignorable biases of the estimates of effects of covariates associated with the true response T_{ij}. Our corrected GEE and GQL approaches, based on count error model (5.23) and the LT model (5.27) for true response T_{ij}, can estimate model parameters almost unbiasedly. In addition, the independence GEE approach leads to loss of efficiency

Table 5.3: Simulation results of GEE and GQL approaches based on the count error model (5.23) and the LT model (5.27) with $\beta = (0.6, -1.0, 1.0)$, $\gamma = 0.8$, $\alpha = (0.3, -0.5)$ and $\pi^+ = 0.7$).

	GEE				GQL			
Quantity	Ideal	Naive	Correct1	Correct2	Ideal	Naive	Correct1	Correct2
SM(β_1)	0.6000	0.5681	0.5994	0.6000	0.6000	0.5680	0.5994	0.6000
SSE	0.0010	0.0010	0.0152	0.0018	0.0010	0.0010	0.0151	0.0017
ESE	0.0009	0.0010	0.0128	0.0016	0.0009	0.0011	0.0142	0.0016
CPr	0.924	0.000	0.892	0.918	0.943	0.000	0.939	0.941
SM (β_2)	-0.9995	-0.9378	-0.9995	-0.9997	-0.9994	-0.9412	-0.9993	-0.9994
SSE	0.0146	0.0151	0.0233	0.0234	0.0134	0.0143	0.0213	0.0213
ESE	0.0132	0.0145	0.0206	0.0213	0.0131	0.0148	0.0202	0.0201
CPr	0.924	0.020	0.910	0.914	0.946	0.018	0.936	0.940
SM(β_3)	1.0002	0.9624	0.9997	0.9998	1.0003	0.9620	0.9998	0.9999
SSE	0.0084	0.0093	0.0127	0.0123	0.0076	0.0090	0.0119	0.0114
ESE	0.0075	0.0081	0.0109	0.0106	0.0072	0.0084	0.0110	0.0109
CPr	0.921	0.012	0.911	0.915	0.941	0.010	0.934	0.938
SM(γ)	0.8003	0.7682	0.8001	0.8002	0.8002	0.7721	0.7999	0.8000
SSE	0.0070	0.0077	0.0102	0.0099	0.0062	0.0074	0.0092	0.0090
ESE	0.0067	0.0076	0.0093	0.0091	0.0061	0.0071	0.0088	0.0088
CPr	0.940	0.014	0.924	0.934	0.948	0.024	0.940	0.946
SM(α_1)			0.2980	0.2982			0.2981	0.2983
SSE			0.0206	0.0198			0.0196	0.0191
ESE			0.0191	0.0187			0.0180	0.0178
CPr			0.944	0.946			0.954	0.951
SM(α_2)			-0.5517	-0.5685			-0.5557	-0.5714
SSE			0.2692	0.2699			0.2422	0.2378
ESE			0.2284	0.2214			0.2075	0.2064
CPr			0.940	0.942			0.954	0.952
SM(π^+)			0.7113				0.7115	
SSE			0.1013				0.1009	
ESE			0.0726				0.0936	
CPr			0.842				0.936	

Table 5.4: Simulation results of GEE and GQL approaches based on the count error model (5.23) and the LT model (5.27) with $\beta = (0.6, -1.0, 1.0)$, $\gamma = 0.3$, $\alpha = (0.3, -0.5)$ and $\pi^{+} = 0.85$).

	GEE				GQL			
Quantity	Ideal	Naive	Correct1	Correct2	Ideal	Naive	Correct1	Correct2
SM(β_1)	0.6000	0.5893	0.6006	0.6001	0.6000	0.5891	0.6005	0.6001
SSE	0.0009	0.0009	0.0116	0.0017	0.0008	0.0009	0.0115	0.0017
ESE	0.0008	0.0009	0.0104	0.0017	0.0009	0.0009	0.0118	0.0017
CPr	0.930	0.000	0.926	0.940	0.952	0.000	0.964	0.948
SM(β_2)	-0.9999	-0.9418	-0.9991	-0.9990	-0.9998	-0.9440	-0.9991	-0.9990
SSE	0.0112	0.0115	0.0163	0.0160	0.0111	0.0115	0.0160	0.0158
ESE	0.0106	0.0110	0.0167	0.0164	0.0109	0.0114	0.0168	0.0166
CPr	0.946	0.002	0.956	0.954	0.948	0.002	0.954	0.952
SM(β_3)	0.9997	0.9537	0.9994	0.9993	0.9996	0.9533	0.9992	0.9991
SSE	0.0073	0.00787	0.0135	0.0124	0.0071	0.0078	0.0134	0.0124
ESE	0.0072	0.0076	0.0126	0.0120	0.0072	0.0080	0.0132	0.0125
CPr	0.946	0.000	0.930	0.940	0.952	0.000	0.942	0.953
SM(γ)	0.2998	0.2870	0.2996	0.2995	0.2998	0.2895	0.2996	0.2996
SES	0.0060	0.0065	0.0069	0.0069	0.0055	0.0062	0.0066	0.0066
ESE	0.0058	0.0063	0.0069	0.0069	0.0057	0.0062	0.0067	0.0067
CPr	0.953	0.460	0.946	0.946	0.952	0.594	0.952	0.952
SM(α_1)			0.2980	0.2979			0.2980	0.2980
SSE			0.0202	0.0195			0.0201	0.0192
ESE			0.0194	0.0186			0.0196	0.0189
CPr			0.956	0.946			0.952	0.949
SM(α_2)			-0.5508	-0.5485			-0.5492	-0.5469
SSE			0.2325	0.2250			0.2218	0.2135
ESE			0.2185	0.2119			0.2142	0.2077
CPr			0.959	0.960			0.960	0.954
SM(π^{+})			0.8510				0.8516	
SSE			0.0937				0.0929	
ESE			0.0838				0.0945	
CPr			0.928				0.960	

in estimation due to application of a false covariance structure. Therefore, if the true covariance matrix is available, we prefer to use the GQL approach in estimating model effects. As mentioned previously, the estimation of the error related effects α and π^+ under the corrected approaches is of significant interest in evaluating the accuracy of the data collected through a series of specific procedures.

Chapter 6

Discussion and Future Studies

6.1 Some Remarks

Due to the widely existing measurement errors in data from epidemiologic studies, it is of great interest to examine the adverse effect of measurement errors on statistical inference. The lack of explicit models to describe measurement errors in discrete data leads to complexity and difficulty in the analysis of measurement errors, especially for count data. To our best knowledge, there are still not any models proposed to accommodate both overnumerated and undernumerated data. There are few methods available in existing literatures for modeling count errors. In Chapter 2 of this thesis, we proposed a generalized thinning operation by extending the binomial thinning operation [Steutel and Harn (1979)] and multinomial thinning operation [McKenzie (1991) (2000)]. The aim of this operation is to model different types of transitions between discrete variables. Based on the generalized thinning operation, we gave an explicit misclassification model (2.7) for categorical data and multinomial data, which clearly describes the relationship between the latent variable and its observed but error-prone surrogate. This model can be used to describe both balanced and unbalanced misclassification. It can also be used to model dynamic categorical data from some longitudinal processes. More importantly, we initially build two measurement error models (2.15) and (2.17) to characterize the miscounts in aggregated data. These two models can effectively accommodate overcounted data as well as undercounted data with imperfect sensitivity and specificity of a data collecting procedure. In addition, the binomial version of the misclassification model (2.7) can be used to characterize the reported disease cases in an area with known population size in censal years. The binomial count error model (2.15) can be used to describe the reported disease cases in an area with an unknown but partially informative population size. The corrected additive error model (2.17) can be used to model the reported count of disease cases in an area with almost non-informative unknown population size.

Most of the existing literatures about measurement error problems are focusing on the mis-measured covariates for different types of responses. In recent years, the analysis of longitudinal responses subject to measurement errors has attracted more and more attentions. This thesis mainly focuses on the analysis of mis-measured longitudinal data. To do so, we firstly introduced some longitudinal models in Chapter 3 to fit the categorical and count data. In the first part of Chapter 3, we built a transition model based on the generalized thinning operation for dynamic categorical data. This thinning-operation-based transition model can flexibly accommodate some ordinary linear or non-linear transition models. In the second part of Chapter 3, we developed two dynamic models for longitudinal count data. One is the thinning-operation-based linear transition model, that is, the non-stationary AR(1) model

(3.45), the other is the ordinary linear transition model (3.47). These two models can be used to model dynamic population, incidence counts, or prevalence counts in public health studies. The simulation studies based on the two models showed that the GQL approach produced highly efficient estimates of model parameters which were competitive with the OGQL approach under the NS-AR(1) model, and the ML approach under the LT model. This implies potentially wide applications of the GQL approach because its simplicity and less dependence on model assumptions.

We modeled the misclassified longitudinal categorical responses in Chapter 4. For this purpose, we combined the longitudinal models in Chapter 3 with the explicit misclassification models in Chapter 2 to clearly describe the relationship between the latent and the observed longitudinal responses. The analysis of misclassified dynamic binary responses showed that ignoring classification error can lead to biased naive estimates of parameters of interest, hence result in poor performance of confidence intervals. To obtain more reliable statistical inference, we developed the corrected GQL, OGQL and ML approaches taking into account the measurement errors which are described by EMC model (4.8). All the three approaches produced approximately unbiased estimates of model parameters and satisfactory confidence intervals. Especially, the OGQL approach which includes the second order responses into the estimating procedure exhibited almost identical estimates to those from the ML approach. The EM algorithm was applied in the ML approach due to the unobservable latent response. We reached the same conclusion as that in the case where data are free of errors. In the EM algorithm, we proposed a new estimator of the Fisher information matrix which is demonstrated to be consistent by the simulations. In addition, we also considered an interesting situation under which the data involve unbalanced misclassification. This approach can be used to deal with a special type of missing information which is MAR caused by "unsure" responses. It showed that ignoring the missing values due to "unsure" responses leads to loss of efficiency, but still produces approximately unbiased estimates of model parameters. On the other hand, ignoring classification errors definitely results in biased estimates and poor confidence intervals of model parameters.

As far as the error-contaminated aggregated data are concerned, we discussed two kinds of mis-measured count responses in Chapter 5. In the first case, the dynamic population sizes of an area are assumed to follow the LT model. The error-prone count response is described by the binomial count error model. Simulation study demonstrated that ignoring measurement errors in count responses leads to biased estimates, for both the GEE and GQL approaches, of the effects of covariates associated with true disease rate. However, measurement errors do not affect the estimates of parameters in the model defined for population sizes. This is because the data of population size keep the same no matter we use the true counts or the error-prone counts of disease cases. Our analysis also showed that the corrected GEE and GQL approaches can consistently estimate all parameters of interest. In the second case, the population size was only assumed to follow a Poisson distribution of which we do not have any further knowledge. The true count response was assumed to follow the LT model, and the observed response was characterized by the corrected additive error model. Analysis showed that ignoring measurement errors leads to biased estimates of all parameters in the model of true response. It is interesting to see that, under the corrected GEE and GQL approaches, besides the unbiased estimates in the LT model, we have obtained satisfactory estimates of the error-related parameters, including the parameters in the model of additive errors and the sensitivity, which describe the type I and type II errors of a surveillance program or a registration system. This analysis implies that the effective estimates of these error-related parameters can be used to access the severity of the measurement errors and to evaluate the quality of the data collection procedure.

6.2 Future studies

Our research on mis-measured longitudinal discrete data presents numerous additional research opportunities.

1. We previously assumed that all the misclassification probabilities in the EMC models and the multinomial count error models are known or their estimates can be obtained from prior knowledge. However, these misclassification probabilities may be unknown, and even their estimates are unavailable in practice. In this case, the estimates of these probabilities are of considerable importance because of their scientific interests and their influence on the estimates of the parameters in the model defined on the true responses. A popular solution to this problem is to estimate these misclassification probabilities through independent validation studies. It therefore follows that an appropriate validation study design based on the original sampling scheme deserve to be investigated, especially when there are some other factors need to be taken into consideration, for example, expense of the validation study. Another intuitive way is to simultaneously estimate these probabilities together with other interested parameters, like the analysis of the miscounted data under the corrected additive error model in

Section 5.2. The second way may involve many extra parameters in estimation, which could lead to loss of efficiency as showed in the simulation study in Section 5.2.6.3.

2. Another interesting problem related to these misclassification probabilities is that they may not be constant over time and subjects. The MC matrix in the EMC models (4.1), the sensitivity and specificity in the binomial count error model (5.4), and the sensitivity in the corrected additive error model (5.23) may be associated with some time dependent covariates. Or they may vary over different subpopulations. For example, in the children asthma studies, the asthma statuses of children are reported by parents up to the age 9, and self reported after that [Speizer (1990)]. The sensitivity and the specificity between the parental-report group and the self-report group are demonstrated to be different [Jenkins (1996)]. It would be very interesting to develop the analysis of mis-measured longitudinal data with time-varying, or group-specific, or covariates-dependent MC matrices.

3. In the corrected additive measurement error model, the assumption that the population size follows a Poisson distribution may be violated. This may lead to violation of the independence between $\pi_1 * T$ and e and the assumption about the Poisson distribution of e. For example, we are interested in the lung cancer incidence count T among people who are older than 50 in an area. In a censal year the total population size K is known, but the size of this age-specific subpopulation N may be unknown. In this example, the size of the specific group N can be assumed to follow a binomial distribution, that is $N \sim binomial(K, p)$, where p is the proportion of the people older than 50 among the population in this area. Under this assumption, the size of the infected group T and the size of the healthy group T^0 are not any longer independent Poisson variables. Hence assumptions about the corrected additive model should be adjusted. Therefore, this problem deserves further investigations.

4. Recall that in Chapter 4, we discussed the misclassified categorical data with a special type of missing information due to "unsure" responses which can be accommodated by the unbalanced misclassification. This reminds us that it would be an interesting topic in epidemiologic studies to analyze data suffering from both measurement errors and missing values. In this context, the mechanism of missing value may follow the missing completely at random (MCAR) or missing at random (MAR) mechanism. Recently, some related literatures have appeared [Yi (2008); Wang et al. (2008); Liu (2006); Nicoletti, Peracchi and Foliano (2009)]. However, all of these discussions are focusing on the measurement errors in covariates. Taking this into consideration, the joint modeling of measurement errors and non-ignorable missing values on the responses is of great interest.

5. Furthermore, it can be reasonably assumed that measurement errors in health and population data are not likely to be spatially independent. Therefore, takeing the spatial effects into consideration when modeling measurement errors or missing values is very promising in gaining extra efficiency in the statistical inference.

6. In practice, there are two important sources of measurement errors on count data,

the first type is the misclassification, the other type is the random error caused by unknown and unpredictable changes. In this thesis, we focused on the measurement errors on count data due to misclassification. As far as the random errors are concerned, they tend to yield the same expectations of the observed variable Y and the true variable T. It is known that the measurement error model for continuous data, takes the form as

$$Y = T + e, \tag{6.1}$$

where e is the random error which follows a distribution with zero mean. For normally distributed measurement, e is often assumed to follow $N(0, \sigma^2)$. For count measurements, one may attempt to build a count error model with a similar form to (6.1) in which the random error follows a distribution with zero mean, unique mode and symmetric probability mass function about 0. The question is how to construct such a distribution for integer-valued variables. Here we introduce two types of discrete normal distributions with these properties which deserve further insight. The first is Dasgupta's (1993) "discrete version of the normal distribution" of which the probability of mass function is given by

$$P(x) = ce^{-\theta x^2}, \text{ for some } \theta > 0, \tag{6.2}$$

on integer support $(-\infty, +\infty)$, where c is such that the total probability mass is one, that is, $c^{-1} = \sum_{x=-\infty}^{+\infty} exp(-\theta x^2)$. The other is a modified version of Roy's (2003) discrete normal distribution by latticing a normal distribution, therefore, we call it the "latticed normal distribution". A latticed normal (Lnormal) variate, LX, can be viewed as the discrete concentration of the normal variate

X following $N(\mu, \sigma^2)$. The corresponding probability mass function of LX can be written as

$$P(x) = \Phi(\frac{x + 0.5 - \mu}{\sigma}) - \Phi(\frac{x - 0.5 - \mu}{\sigma}), \tag{6.3}$$

where $\Phi(x)$ represents the cumulative distribution function of the standard normal random variable Z. It is easy to see that both Dasgupta's (1993) discrete normal distribution and our latticed normal distribution have similar properties as the standard normal distribution such as, symmetry, zero mean and the unique mode at 0 and so on. Studies about other properties and the specific applications of these two models are very promising.

Besides the problems mentioned above, the more relevant topics are to be investigated.

Bibliography

[1] Amemiya, T. (1985) *Advanced Econometrics.* Harvard University Press, Cambridge, Massachusetts.

[2] Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal of Statistical Society: Series B*, **36**, 192-236.

[3] Biswas, A., Datta, S., Fine, J.P. and Segal, M. R. (2008) *Statistical Advances in the Biomedical Sciences: Clinical Trials Epidemiology, Survival Analysis, and Bioinformatics.* John Wiley & Sons Inc., Hoboken, New Jersey.

[4] Blundell, R., Griffith, R., and Windmeijer, F. (2002) Individual effects and dynamics in count data models. *Journal of Econometrics*, **108**, 113-31.

[5] Brandi, A.G., Young, D.M. and Stamey, J.D. (2009) Bayesian inference for comparing two Poisson rates using data subject to underreporting and validation data. *Statistical Methodology*, **7**, 98-108.

[6] Bratcher, T.L. and Stamey, J.D. (2002) Estimation of Poisson rates with misclassified counts. *Biometrical Journal*, **44**, 946-56.

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

[8] Cameron, A.C. and Trivedi, P.K. (1998) *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

[9] Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, M. (2006) *Measurement Error in Non Linear Models: A Modern Perspective*. Chapman & Hall, New York.

[10] Chen, T.T. (1989) A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, **9**, 1095-106; discussion 1107-8.

[11] Colby, T.V., Tazelaar, H.D., Travis, W.D. , Bergstralh, E.J. and Jett, J.R. (2002) Pathologic review of the Mayo Lung Project cancers [corrected]. Is there a case for misdiagnosis or overdiagnosis of lung carcinoma in the screened group? *Cancer* , **95**, 2361-65.

[12] Cooper, G.S., Yuan, Z., Stange, K.C., Dennis, L.K., Amini, S.B., and Rimm, A.A. (1999) The sensitivity of Medicare claims data for case ascertainment of six common cancers. *Medical Care*, **37**, 436-44.

[13] Cui, Y. and Lund, R. (2009) A new look at time series of counts. *Biometrika*, **96**, 781-92.

[14] Dasgupta, R. (1993) Cauchy equation on discrete domain and some characterization theorems. *Theory of Probability and Its Applications*, **38**, 520-24.

[15] Diggle, P.J., Heagerty P., Liang, K.-Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

[16] Dundas, I. and Mckenzie, S. (2006) Spirometry in the diagnosis of asthma in children. *Current Opinion in Pulmonary Medicine*, **12**, 28-33.

[17] Farrell, P.J. and Sutradhar, B.C. (2006) A non-linear conditional probability model for generating correlated binary data. *Statistics and probability Letters*, **76**, 353-61.

[18] Ferris, B.G., Ware, J.M., Berkey, C.S. Dockery, D.W., Spiro, III A. and Speizer, F.E. (1985) Effects of passive smoking on health of children. *Environmental Health Perspectives*, **62**, 289-95.

[19] Fisher E.S., Whaley F.S., Krushat W.M., Malenka, D.J., Fleming, C., Baron, J.A. and Hsia, D.C. (1992) The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. *American Journal of Public Health*, **82**, 243-48.

[20] Fitzmaurice, G.M. and Laird, N.M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141-51.

[21] Friebele, E. (1996) The attack of asthma. *Environmental Health Perspectives*, **104**, 22- 25.

[22] Fuhlbrigge, A.L., Kitch, B.T., Paltiel, A.D., Kuntz, K.M., Neumann, P.J., Dockery, D.W. and Weiss, S.T. (2001) FEV1 is associated with risk of asthma attacks in a pediatric population. *Journal of Allergy Clinical Immunology*, **107**, 61-67.

[23] Fuller, W. (1987) *Measurement Error Models.* Wiley, New York.

[24] Furlow, B, (2007) Accuracy of US cancer surveillance under threat. *Lancet Oncology,* **8**, 762-63.

[25] Giercksky, K.E. (1997) Misdiagnosis of cancer due to multiple glove powder granulomas. *European Journal of Surgery, Supplement,* **579**, 11-14.

[26] Gilliand, F., Li, Y-F., Peters J. (2001) Effect of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children. *American Journal of Respiratory and Critical Care Medicine,* **163**, 429-36.

[27] Gustafson, P. (2007) Measurement error modeling with an approximate instrumental variable. *Journal of the Royal Statistical Society, Series B,* **69**, 797-815.

[28] Gustafson P. (2003) M*easurement Error and Misclassification in Statistics and : Impacts and Bayesian Adjustments.* Chapman & Hall/CRC, Boca Raton.

[29] Hardin, J. and Hilbe, J. (2003) *Generalized Estimating Equations.* Chapman and Hall/CRC, London.

[30] Hossain, S. and Gustafson, P. (2009) Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Statistics in Medicine,* **28**, 1580-600.

[31] Jenkins, M.A., Clarke, J.R., Carlin, J.B. , Robertson, C.F., Hopper, J.L., Dalton, M.F., Holst, D.P., Choi, K. and Giles, G.G. (1996) Validation of questionnaire and bronchial hyperresponsiveness against respiratory physician assessment in the diagnosis of asthma. *International Journal of Epidemiology,* **25**, 609-16.

[32] Kanter, M. (1975) Autoregression for discrete processes mod 2. *Journal of Applied Probability*, **12**, 371-75.

[33] Kim, J.H. (2009) Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, **53**(11), 3735-45

[34] Kipnis, V., Carroll, R.J., Freedman, L.S. and Li, L. (1999) 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.

[35] Kipnis, V., Midthune, D., Freedman, L.S., Bingham, S., Day, N.E., Riboli, E. and Carroll, R.J.,(2003) Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition*, **5**, 915-23.

[36] Korn, E.L. and Whittemore, A.S. (1979) Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, **35**, 795-802.

[37] Küchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006) A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*, **62**, 85-96.

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

[39] Liu, W. (2006) The theory and Methods for measurement errors and missing data problems in semiparametric nonlinear mixed-effect models. *Doctorial Thesis*, The University of British Columbia, Vaccover, British Columbia, Canada.

[40] Louis, T. A. (1982) Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society: Series B*, **44**, 226-33.

[41] Mallick, B.K., and Gelfand, A.E. (1996) Semiparametric error-in-variables models: A Bayesian approach. *Journal of Statistical Planning and Inference*, **52**, 307-21.

[42] Mallick T.S. (2009) Conditional weighted generalized wuasi-likelihood inferences in incomplete longitudinal models for binary and count data. *Doctorial Thesis*, Memorial University of Newfoundland, St. John's, Newfoundland, Canada.

[43] Manski, C.F. (1987) Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, **55**, 357-62.

[44] Marshall, R.J. (1990) Validation study methods of estimating proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, **43**, 941-47.

[45] McGlothlin A., Stamey, J.D. and Seaman, J.W. (2008) Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach. *Biometrical Journal*, **50**, 123-34.

[46] McKenzie, E. (1985) Some simple models for discrete variate time series. *Water Resources Bulletin*, **21**, 645-50.

[47] McKenzie, E. (1986) Autoregressive-moving average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability* **18**, 679-705.

[48] McKenzie, E. (1988) Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, **20**, 822-35.

[49] McKenzie, E. (2003) Discrete variate time series. *In Handbook of Statistics*, Rao, C.R. and Shanbhag, D., Eds., Elsevier Science, Amsterdam, 573-606.

[50] Motoo, Y., Watanabe, H. and Sawabu, N. (1996) Sensitivity and specificity of tumor markers in cancer diagnosis. *Nippon Rinsho*, **54**, 1587-91.

[51] Neuhaus, J.M. (1999) Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86**, 843-55.

[52] Neuhaus, J.M. (2002) Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, **58**, 675-83.

[53] Newey, W.K. and McFadden, D. (1993) Estimation in large samples. *In Handbook of Economics*, McFadden, D. and Engler, R. eds., North Holland, Amsterdam.

[54] Nicoletti, C., Peracchi, F. and Foliano, F. (2009) Estimating income poverty in the presence of missing data and measurement error. *German Socio-Economic Panel Study*, paper No. 252.

[55] Pattenden, S., Antova, T., Neuberger, M., Nikiforov, B., De Sario, M., Grize, L., Heinrich, J., Hruba, F., Janssen, N., Luttmann-Gibson, H., Privalova, L., Rudnai, P., Splichalova, A., Zlotkowska, R. and Fletcher T. (2006) Parental smoking and children's respiratory health: independent effects of prenatal and postnatal exposure. *Tobacco Control*, **15**(4), 294-301.

[56] Qaqish, B.F. (2003) A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, **90**, 455-63.

[57] Rosychuk, R.J. (1999) Accounting for misclassification in binary longitudinal data. *Doctorial Thesis*, Waterloo University, Waterloo, Ontario, Canada.

[58] Rosychuk, R.J., Thompson, M.E. (2001) A semi-Markov model for binary longitudinal responses subject to misclassification. *The Canadian Journal of Statistics*, **29**, 395-404.

[59] Rosychuka, R.J., and Islam, S. (2009) Parameter estimation in a model for misclassified Markov data - a Bayesian approach. *Computational Statistics and Data Analysis*, **53**, 3805-16.

[60] Roy, D. (2003) The discrete normal distribution. *Communications in Statistics: Theory and Methods*, **32**(10), 1871-83.

[61] Roy, S., Banerjee, T. and Maiti, T. (2005) Measurement error model for misclassified binary responses. *Statistics in Medicine*, **24**, 269-83.

[62] Roy, S., Banerjee, T. (2009) Analysis of misclassified correlated binary data using a multivariate probit model when covariates are subject to measurement error. *Biometrical Journal*, **51**, 420-32.

[63] Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2003) Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, **3**, 215-32.

[64] Schafer, D.W. (1987) Covariate measurement error in generalized linear models. *Biometrika*, **74**, 385-91.

[65] Speizer, F.E. (1990) Asthma and persistent in Harvard Six Cities Study. *Chest*, **98**, 191S-95S.

[66] Spiegelman, D., Rosner, D.L. and Logan, R. (2000) Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study design. *Journal of American Statistical Association*, **95**(449), 51-61.

[67] Stamey, D., Seaman, J.W. and Young, D.M. (2005) Bayesian analysis of complementary Poisson rate parameters with data subject to misclassification. *Journal of Statistical Planning and Inference*, **134**, 36-48.

[68] Stefanski, L.A. and Cook, J. (1995) Simulation extrapolation: The measurement error jackknife. *Journal of the American Statistics Association*, **90**, 1247-56.

[69] Stefanski, L.A. and Carroll, R.J. (1985) Covariate measurement error in logistic regression. *Annual of Statististics*, **13**, 1335-51.

[70] Stefanski, L.A. (1987) The effect of measurement error in parameter estimation. *Biometrika*, **72**, 385-89.

[71] Steutel, F.W. and Harn K-van (1979) Discrete analogues of self-decomposability and stability. *Annual of Probability*, **7**, 893-99.

[72] Steutel, F.W., Vervaat, W. and Wolfe, S.J. (1983) Integer valued branching processes with immigration. *Advances in Applied Probability*, **15**, 713-25.

[73] Sutradhar, B.C. (2003) An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**, 377-93.

[74] Sutradhar, B.C. (2008) Inferences in familial Poisson mixed models for survey data. *Sankhya*, **70**, 18-33.

[75] Sutradhar, B.C. and Farrell, P.J. (2007) On optimal lag 1 dependence estimation for dynamic binary models with application to Asthma data. *Sankhya*, **69**, 448- 67.

[76] Sutradhar, B.C. and Das, K. (1999) On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, **86**, 459-65.

[77] Sutradhar, B.C., Jowaheer, V. and Sneddon, G. (2008) On a unified generalized Quasi-likelihood approach for familial-longitudinal non-stationary count data. *Scandinavian Journal of Statistics*, **35**, 597-612.

[78] Tong, H. (1990) *Nonlinear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford.

[79] Wang, C.Y. (2008) Nonparametric maximum likelihood estimation for Cox regression with subject-specific Measurement Error. *Scandinavian Journal of Statistics*, **35**, 613-28.

[80] Wang, C.Y., Huang, Y., Chao, E.C., and Jeffcoat, M.K. (2008) Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorably missing data. *Biometrics*, **64**, 85-95.

[81] Wang, P.S., Walker, A.M., Tsuang M.T., Orav, E.J., Levin, R. and Avorn, J. (2001) Finding incident breast cancer cases through US claims data and a state cancer registry. *Cancer Causes Control*, **12**, 257-65

[82] Wang, P.S., Walker, A., Tsuang, M., Orav, E.J., Levin, R. and Avorn, J. (2000) Strategies for improving comorbidity measures based on Medicare and Medicaid claims data. *Journal of Clinical Epidemiology,* **53**, 571-78.

[83] Ware, J.H., Dockery, D.W., Spiro, A. III, Speizer, F.E. and Ferris B.G. (1984) Passive smoking, gas cooking and respiratory health of children living in six cities. *American Reviews of Respiratory Disease,* **129**, 366-74.

[84] Wedderburn, R.W. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika,* **61**, 439-47.

[85] Whittemore, A.S. and Gong, G. (1991) Poisson regression with misclassified counts: application to cervical cancer. *Journal of Royal Statistics Society, Series C,* **40**, 81-93.

[86] Wong, W.H. (1986) Theory of partial likelihood. *Annals of Statistics,* **14**, 88-123.

[87] Yang, C.-Y., Tien, Y.-C., Hsieh, H.-J., Kao, W.-Y. and Lin, M.-C. (1998) Indoor environmental risk factors and child asthma: a case-control study in a subtropical area. *Pediatric Pulmonology,* **26**, 120-24.

[88] Yerushalmy, J. (1947) Statistical problems in assessing methods of medical diagnosis with special reference to X-ray technique. *Public Health Respective,* **62**, 1432-49.

[89] Yi, G.Y. (2008) A simulation-based marginal method for longitudinal data with drop-out and mismeasured covariates. *Biostatistics,* **9**, 501-12.

[90] Youden, W.J. (1950) Index for rating diagnostic tests. *Cancer,* **3**, 32-35.

[91] Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988) Models for longitudinal data: A gneralized estimating equations approach. *Biometrics*, **44**, 1049-60.

[92] Zeger SL and Liang K.-Y. (1986) The analysis of discrete and continuous longitudinal data. *Biometrics*, **42**, 121-30.

[93] Zeger, S.L., Liang, K.-Y. and Self, S.G. (1985) The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, **72**, 31-8.

[94] Zeger, S.L. and Qaqish, B. (1988) Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, **44**, 1019-31.

[95] Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642-48.

[96] Zucker D.M. and Spiegelman D. (2008) Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Mededicine*, **27**(11), 1911-33.