# ESTIMATION AND FORECASTING OF A LAG 2 DYNAMIC MODEL FOR INFECTIOUS DISEASES

CHEN ZHANG

# Estimation and Forecasting of a Lag 2 Dynamic Model for Infectious Diseases

by

© *Chen Zhang*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of *Science*

Department of *Mathematics & Statistics*

Memorial University of Newfoundland

*August, 2012*

St. John's Newfoundland

## Abstract

When a common infectious disease is first detected in a community, it may quickly spread out through air, water, public facilities and personal contacts. At a given time point, each infected individual may or may not infect other individuals in the community. Meanwhile, it is also possible that some individuals who carry the same disease travel into the community. In the present work, we discuss estimation and forecasting of an extension to the lag 1 longitudinal dynamic model for correlated data used by Oyet & Sutradhar (2011) for modelling the spread of infectious disease. The lag 1 model only allow individuals with infection at time point $t-1$ to cause new infections at time point $t$. Clearly, if at time point $t-2$, there is an individual who is still infected by the disease, it is also possible for this individual to infect others at time point $t$. The present model discussed in this work allows for such a possibility. During the modelling, we consider stationary and nonstationary covariates. We also extend the model to situations where unobservable community effect and the latent community effect is present. The regression parameter $\beta$ and the parameter of latent community effect $\sigma_\gamma^2$ are estimated by generalized quasi-likelihood (GQL) approach. The correlation parameters $\rho_1$ and $\rho_2$ are estimated by using method of moments. In each of the cases, we examined the accuracy of the estimates and forecasts through simulation studies.

## Acknowledgements

I want to express my deepest gratitude to my supervisor, Dr. Alwell Oyet in paticular for his guidance, inspirations and encouragement. The thesis would be impossible without his intellectual, practical and financial support.

My sincere thanks also goes to Dr. Brajendra C. Sutradhar for providing me the very useful background used in the thesis. I have learned a lot from his several courses.

It is my pleasure to thank Dr. Yuan Yuan, Dr. Zhaozhi Fan, Dr. Jie Xiao, Dr. Xiaoqiang Zhao, Dr. Chunhua Ou, and Haiyan Yang for their advice and encouragement. I have really enjoyed my time with them.

I would like to extend my thanks to all my friends, especially Zhen Wang, for her constant encouragement and support during my masters study.

I wish to acknowledge with thanks the administrative staff and computer technicians in the Department of Mathematics and Statistics for making their time and help available to me.

Words fail me to express my appreciation to my parents whose dedication, love and persistent confidence in me, have taken the load off my shoulder.

The financial support of the Memorial University of Newfoundland is gratefully acknowledged.

Finally, I would like to thank everybody who was important to the successful realization of thesis, as well as expressing my apology that I could not mention personally one by one.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In recent years, statistical models have shown to be of great value in the modelling of infectious disease. In particular, Oyet & Sutradhar (2011) proposed a branching process with immigration type longitudinal count data model for modelling the number of infections in each community. In their paper, they assumed that one infected person may possibly infect none, one or more individuals in a small time interval. Some immigrants with the same disease may enter the community during the same time interval which will increase the number of infected individuals in the community. Let $y_{it}$ be the number of infected individuals at time $t$ $(t = 2, 3, ..., T)$ in the community $i$ $(i = 1, 2, ..., K)$, Oyet & Sutradhar (2011) model is given by

$$y_{it} = \sum_{j=1}^{y_{i,t-1}} B_j(n_t, \rho) + d_{it} \tag{1.1}$$

with assumptions:

Assumption 1. $y_{i1} \sim Poi(\mu_{i1} = exp(x'_{i1}\beta))$.

Assumption 2. $d_{it} \sim Poi(\mu_{it} - \rho_1 \mu_{i,t-1})$ for $t = 2, ..., T$ with $\mu_{it} = exp(x'_{it}\beta)$, for all $t = 1, ..., T$.

Assumption 3. $d_{it}$ and $y_{i,t-1}$ are independent for $t = 2, ..., T$.

For model (1.1), they found that for all t = 1,2,...,T and k = 1,2,...,T-1

$$\begin{aligned} E(Y_{it}) &= \mu_{it}, \\ Var(Y_{it}) &= \mu_{it} - n_t\rho^2\mu_{i,t-1} + n_t^2\rho^2 var(Y_{i,t-1}) = \sigma_{itt}, \\ Cov(Y_{it}, Y_{i,t-k}) &= \left(\prod_{l=0}^{k-l} n_{t-l}\right)\rho^k \sigma_{i,t-k,t-k}, \\ Corr(Y_{it}, Y_{i,t-k}) &= \left(\prod_{l=0}^{k-l} n_{t-l}\right)\rho^k \sqrt{\frac{\sigma_{i,t-k,t-k}}{\sigma_{itt}}}. \end{aligned}$$

Note that when $n_t = 1$, for all $t = 1, 2, .., T$, the binomial random variable $B_j(n_t, \rho)$ will become a binary variable with the probability of infection $\rho$. The model (1.1) will then reduce to an autoregressive, of order 1, (AR(1)) type Poisson process. This reduced model implies that the infectious individuals at time $t - 1$ can only infect none or one individual at time point $t$.

## 1.1 Poisson AR(1) Process

In the previous section, we had noted that when $n_t = 1$, for all $t = 1, 2, ..., T$, model (1.1) becomes an AR(1) type Poisson process. In this case, the assumptions for model (1.1) becomes $y_{i1} \sim Poi(\mu_{i1} = exp(x'_{i1}\beta))$ and $d_{it} \sim Poi(\mu_{it} - \rho\mu_{i,t-1})$ and

the process can be written as

$$y_{it} = \sum_{j=1}^{y_{i,t-1}} B_j(\rho) + d_{it}. \tag{1.2}$$

Sutradhar (2011) (see also McKenzie (1988)) used (1.2) for $t = 1, 2, ..., T$ to model longitudinal count data over time. The basic properties of the model are

$$\begin{aligned} E(Y_{it}) &= \mu_{it},\ Var(Y_{it}) = \mu_{it}, \\ Cov(Y_{it}, Y_{i,t-k}) &= \rho^k \mu_{it} \text{ and } Corr(Y_{it}, Y_{i,t-k}) = \rho^k \sqrt{\frac{\mu_{it}}{\mu_{ik}}}. \end{aligned}$$

Mckenzie (1988) showed that the distribution of the process (1.2) is Poisson with mean $\mu_{it} = exp(x_{i1}'\beta)$ by using alternate probability generating function (a.p.g.f.s). We can consider this model as a model for spread of disease for only some limited cases because it only allows each of the $y_{i,t-1}$ infected individuals at time $t-1$ to infect at most one individual. When $y_{i,t-1}$ is considered to be an offspring variable at time $t-1$ and $d_{it}$ is the immigration variable, the model (1.2) represents a branching process with immigration for $K = 1$ and large $T$. This model was recently considered by Sutradhar, Oyet and Gadag (2010) as a special case of a negative binomial branching process with immigration.

## 1.2 Generalized Quasi-Likelihood (GQL)

For the longitudinal regression setup, interest may be focused on the regression parameters for the marginal expectations of the longitudinal responses and the longitudinal correlation parameters. For the regression parameters, there exists a "work-

ing" correlation matrix based generalized estimating equation (GEE) approach for the estimation of the regression parameters and generalized quasi-likelihood (GQL) estimation approach. The GEE approach was proposed by Liang & Zeger (1986). It has been used extensively in recent years in estimation for longitudinal count response. However, as demonstrated by Crowder (1995), because of the uncertainty in the definition of the working correlation matrix, the GEE approach may in some cases lead to a complete breakdown of the estimation of the regression parameters. Furthermore, Sutradhar and Das (1999) have demonstrated that even though the GEE approach in many situations yields consistent estimators for the regression parameters, the GEE approach may, however, produce less efficient estimates than the independence assumption based quasi-likelihood (QL) or moment estimates. Sutradhar ((2011) p.4) suggests that based on studies by Crowder (1995), Sutradhar and Das (1999), Sutradhar (2003), and Sutradhar (2010), the GEE approach cannot be trusted for regression estimation in discrete models such as longitudinal binary or count data. Sutradhar (2003, Section 3) therefore suggested an efficient GQL approach for time independent covariates which is an extension of the QL approach (or weighted least squares approach) for the independent data introduced by Wedderburn (1974). Sutradhar (2010) introduced nonstationary autocorrelation structures for the cases when covariates are time dependent, and applied the GQL approach for consistent and efficient estimation of the regression effects.

## 1.3 Motivation

From the model (1.1), (1.2) and (1.3), we can see that these lag 1 models only allow individuals with infection at time point $t-1$ to cause new infections at time point $t$. Clearly, if at time point $t-2$, there is an individual who is still infected by the disease, it is possible for this individual to infect others at time $t$ as well. We develop a model which include infections from time $t-2$. For simplicity, we first extend the Poisson AR(1) process to a lag 2 model. Since the number of infections in each community may also be affected by unobserved community effects such as environmental pollution, we also extend the lag 2 model a little further by introducing a random variable to represent the latent community effect. However, these AR(1) type extended models will have similar limitations as a Poisson AR(1) process when used to model the spread of infectious disease. That is, the infected individuals at time $t-1$ or $t-2$ can only infect at most one individual at time $t$. Consequently, we also consider an extension to Oyet & Sutradhar's (2011) lag 1 model for infectious disease which would be more appropriate to model the number of infections in reality.

# Chapter 2

# Lag 2 Dynamic Binary Sum Infectious Disease Model

## 2.1 Preliminaries

In this chapter, we begin to construct new models which will include the information from previous two stages. Suppose that we have K communities. First, we begin with a simple case, which deals with an infectious individual who can only infect at most one person at time point $t$. In section 2, we consider both stationary and nonstationary covariates assuming that no community effect is present. That is, all communities are assumed to be independent of each other, but they are time-wise correlated with themselves. We discuss the structures and assumptions of the model and derive some basic properties of the model, such as the mean, variance, covariance and correlation. Finally, we estimate the parameters and forecast the future number of infections with both stationary and nonstationary covariates based on the

simulated data. In section 3, we assume there is an unobservable community effect which can affect our responses. For example, this unobservable community effect $\gamma_i$ could be wealth difference or education levels. The mean function depends on the regression effect and community effect. We discuss the structures and assumption for this dynamic mixed model and obtain properties of the mixed model. Finally, we estimate the parameters involved in this mixed model with nonstationary covariates.

The main goal of this chapter is to estimate the parameters involved in the model and to forecast the number of infections in community at time point t. We have found out that the Generalized Quasi-likelihood (GQL) method of Sutradhar (2003) for estimating the regression parameters of longitudinal response works very well for estimating the regression effects of this model. In the same year, Sutradhar and Jowaheer (2003) have also used the GQL method to estimate the variance parameter of the community random effect $\gamma_i$. We will use this GQL approach for the estimation of our model parameters. We estimate the longitudinal correlation parameters by using the Method of Moments (MM). Once all parameters are properly estimated, we then use the information from $t$ and $t-1$ to examine the forecast performance of this model.

## 2.2 Lag 2 Fixed Binary Sum Infectious Disease Model

We begin our modelling by considering the simple case where the offspring random variable is binary with correlation index parameters $\rho_1$ and $\rho_2$ for two consecutive gen-

erations. We assume that K independent communities are at the risk of an infectious disease. Suppose that at initial time point, $t = 1$, $y_{i1}$ individuals in the $ith$ community developed the disease where $y_{i1}$ is assumed to follow a Poisson distribution with mean parameter $\mu_{i1} = exp(x'_{i1}\beta)$. Because we assume an infected individual can only effect none or one individual for two time intervals, and also because there may be other infected individuals arriving from other communities, we shall model the number of infected people in $i$th community at time $t$ as

$$\begin{aligned} Y_{i1} &\sim Poi(\mu_{i1}), where\ \mu_{i1} = e^{x'_{i1}\beta} \\ Y_{i2} &= \sum_{j=1}^{Y_{i1}} b_{1j}(\rho_1) + d_{i2} \\ Y_{it} &= \sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1) + \sum_{j=1}^{Y_{i,t-2}} b_{2j}(\rho_2) + d_{it},\ for\ t = 3, 4, ..., T, \end{aligned} \tag{2.1}$$

where $b_{1j} \sim Bin(\rho_1), b_{2j} \sim Bin(\rho_2)$. This is an extension of the AR(1) model used by Staudenmayer and Buonaccorsi (2005) and Sutradhar (2003). In model (2.1), we make the following assumptions:

(1) $d_{i2} \sim Poi(\mu_{i2} - \rho_1\mu_{i1})$.

(2) $d_{it} \sim Poi(\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2})$.

(3) $Y_{i,t-1}$& $d_{it}$ *are independent for* $t = 2, 3, ..., T$.

(4) $Y_{i,t-2}$& $d_{it}$ *are independent for* $t = 3, ..., T$.

From the model (2.1), it is clear that $E[Y_{i1}] = Var(Y_{i1}) = \mu_{i1}$. Using assumption (1) and (3), it can be shown that $E[Y_{i2}] = Var(Y_{i2}) = \mu_{i2}$, $Cov(Y_{i1}, Y_{i2}) = \rho_1\mu_{i1}$ and $Corr(Y_{i1}, Y_{i2}) = \rho_1\sqrt{\frac{\mu_{i1}}{\mu_{i2}}}$. We note that in assumption (1), the Poisson mean

parameter must satisfy $\mu_{i2} - \rho_1\mu_{i1} \geq 0$ , yielding $\rho_1 \leq \frac{\mu_{i2}}{\mu_{i1}}$. Similarly, for $\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} \geq 0$ to be satisfied, we need $\rho_1 \leq \frac{\mu_{it}-\rho_2\mu_{i,t-2}}{\mu_{i1}}$. Therefore, the range of $\rho_1$ is

$$0 \leq \rho_1 \leq min\left(\frac{\mu_{i2}}{\mu_{i1}}, \frac{\mu_{it} - \rho_2\mu_{i,t-2}}{\mu_{i,t-1}}, 1\right), \ for\ t \geq 3\ and\ fixed\ \rho_2.$$

For stationary case, if we let $\mu_{i1} = \mu_{i2} = \cdots = \mu_{iT} = \mu_i$, then the range of $\rho_1$ simplifies to

$$0 \leq \rho_1 \leq (1 - \rho_2).$$

### 2.2.1 Basic Properties of the Lag 2 Fixed Binary Sum Infectious Disease Model

#### 2.2.1.1 The Mean

Based on the previous discussion, we know that $E[Y_{i1}] = \mu_{i1}$ and $E[Y_{i2}] = \mu_{i2}$. Then, it follows that for $t = 3, 4, ...T$,

$$\begin{aligned}
E[Y_{it}] &= E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1) + \sum_{j=1}^{Y_{i,t-2}} b_{2j}(\rho_2) + d_{it}\right] \\
&= E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1)\right] + E\left[\sum_{j=1}^{Y_{i,t-2}} b_{2j}(\rho_2)\right] + E[d_{it}] \\
&= E_{y_{i,t-1}}E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1)\Big|_{y_{i,t-1}}\right] + E_{y_{i,t-2}}E\left[\sum_{j=1}^{Y_{i,t-2}} b_{2j}(\rho_2)\Big|_{y_{i,t-2}}\right] + E[d_{it}] \\
&= \rho_1 E[Y_{i,t-1}] + \rho_2 E[Y_{i,t-2}] + \mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2}. \qquad (2.2)
\end{aligned}$$

Next, we consider some specific cases:

For $t = 3$,

$$\begin{aligned} E[Y_{i3}] &= \rho_1 E[Y_{i2}] + \rho_2 E[Y_{i1}] + \mu_{i3} - \rho_1 \mu_{i2} - \rho_2 \mu_{i1} \\ &= \rho_1 \mu_{i2} + \rho_2 \mu_{i1} + \mu_{i3} - \rho_1 \mu_{i2} - \rho_2 \mu_{i1} \\ &= \mu_{i3}. \end{aligned}$$

For $t = 4$,

$$\begin{aligned} E[Y_{i4}] &= \rho_1 E[Y_{i3}] + \rho_2 E[Y_{i2}] + \mu_{i4} - \rho_1 \mu_{i3} - \rho_2 \mu_{i2} \\ &= \rho_1 \mu_{i3} + \rho_2 \mu_{i2} + \mu_{i4} - \rho_1 \mu_{i3} - \rho_2 \mu_{i2} \\ &= \mu_{i4}. \end{aligned}$$

By mathematical induction, if we have $E[Y_{i,t-1}] = \mu_{i,t-1}$ and $E[Y_{i,t-2}] = \mu_{i,t-2}$, then:

$$\begin{aligned} E[Y_{it}] &= \rho_1 \mu_{i,t-1} + \rho_2 \mu_{i,t-2} + \mu_{it} - \rho_1 \mu_{i,t-1} - \rho_2 \mu_{i,t-2} \\ &= \mu_{it} \\ &= exp(x'_{it}\beta). \end{aligned} \tag{2.3}$$

So, $E[Y_{it}] = \mu_{it} = exp(x'_{it}\beta)$, for all $t = 1, 2, ..., T$.

#### 2.2.1.2 The Variance

The variance of this model can be derived by finding the conditional and unconditional variances. By assumptions (2) and (3), we have

$$Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}, \sum_{j=1}^{Y_{i,t-2}} b_{2j} \middle| Y_{i,t-1}, Y_{i,t-2}\right) = 0$$

and

$$Cov(Y_{i,t-1}, d_{it}) = 0 \ , \ and \ Cov(Y_{i,t-2}, d_{it}) = 0.$$

Then

$$\begin{aligned} Var(Y_{it}|Y_{i,t-1}, Y_{i,t-2}) &= Y_{i,t-1}Var(b_{1j}) + Y_{i,t-2}Var(b_{2j}) + Var(d_{it}) \\ &= Y_{i,t-1}\rho_1(1-\rho_1) + Y_{i,t-2}\rho_2(1-\rho_2) \\ &\quad +\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2}. \end{aligned} \tag{2.4}$$

Letting $\sigma_{i,t-1,t-1}$ represent the variance of $Y_{i,t-1}$, then

$$\begin{aligned} Var(Y_{it}|Y_{i,t-2}) &= E_{Y_{i,t-1}}[Var(Y_{it}|Y_{i,t-1}, Y_{i,t-2})] + Var_{Y_{i,t-1}}(E[Y_{it}|Y_{i,t-1}, Y_{i,t-2}]) \\ &= E_{Y_{i,t-1}}[Y_{i,t-1}\rho_1(1-\rho_1) + Y_{i,t-2}\rho_2(1-\rho_2) \\ &\quad +(\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2})] \\ &\quad +Var_{Y_{i,t-1}}(Y_{i,t-1}\rho_1 + Y_{i,t-2}\rho_2 + (\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2})) \\ &= \mu_{i,t-1}\rho_1(1-\rho_1) + Y_{i,t-2}\rho_2(1-\rho_2) + \rho_1^2\sigma_{i,t-1,t-1} \\ &\quad +\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2}. \end{aligned} \tag{2.5}$$

Similarly, letting $\sigma_{i,t-2,t-2}$ represent the variance of $Y_{i,t-2}$, we have

$$\begin{aligned}
Var(Y_{it}) &= E_{Y_{i,t-2}}[Var(Y_{it}|Y_{i,t-2})] + Var_{Y_{i,t-2}}(E[Y_{it}|Y_{i,t-2}]) \\
&= E_{Y_{i,t-2}}[\mu_{i,t-1}\rho_1(1-\rho_1) + Y_{i,t-2}\rho_2(1-\rho_2) \\
&\quad +\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} + \rho_1^2\sigma_{i,t-1,t-1}] \\
&\quad +Var_{Y_{i,t-2}}(E_{Y_{i,t-1}}E[Y_{it}|Y_{i,t-1}, Y_{i,t-2}]) \\
&= \mu_{i,t-1}\rho_1(1-\rho_1) + \mu_{i,t-2}\rho_2(1-\rho_2) \\
&\quad +\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} + \rho_1^2\sigma_{i,t-1,t-1} \\
&\quad +Var(\mu_{i,t-1}\rho_1 + Y_{i,t-2}\rho_2 + \mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2}) \\
&= \mu_{i,t-1}\rho_1(1-\rho_1) + \rho_1^2\sigma_{i,t-1,t-1} + \rho_2^2\sigma_{i,t-2,t-2} \\
&\quad +\mu_{i,t-2}\rho_2(1-\rho_2) + \mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} \\
&= \mu_{it} - \mu_{i,t-1}\rho_1^2 - \mu_{i,t-2}\rho_2^2 + \rho_1^2\sigma_{i,t-1,t-1} + \rho_2^2\sigma_{i,t-2,t-2} \\
&= \mu_{it} - (\mu_{i,t-1} - \sigma_{i,t-1,t-1})\rho_1^2 - (\mu_{i,t-2} - \sigma_{i,t-2,t-2})\rho_2^2. \qquad (2.6)
\end{aligned}$$

From this formula, we can see that the variance of $Y_{it}$ has a recursive relationship with the variance of $Y_{i,t-1}$ and the variance of $Y_{i,t-2}$. We know that $Var(Y_{i1}) = \mu_{i1}$

and $Var(Y_{i2}) = \mu_{i2}$ from our assumptions. It turns out that when $t = 3$,

$$\begin{aligned} Var(Y_{i3}) &= \mu_{i2}\rho_1(1-\rho_1) + \rho_1^2\sigma_{i22} + \rho_2^2\sigma_{i11} \\ &\quad +\mu_{i1}\rho_2(1-\rho_2) + \mu_{i3} - \rho_1\mu_{i2} - \rho_2\mu_{i1} \\ &= \mu_{i2}\rho_1(1-\rho_1) + \rho_1^2\mu_{i2} + \rho_2^2\mu_{i1} \\ &\quad +\mu_{i1}\rho_2(1-\rho_2) + \mu_{i3} - \rho_1\mu_{i2} - \rho_2\mu_{i1} \\ &= \mu_{i3}. \end{aligned}$$

When $t = 4$,

$$\begin{aligned} Var(Y_{i4}) &= \mu_{i2}\rho_1(1-\rho_1) + \rho_1^2\sigma_{i33} + \rho_2^2\sigma_{i22} \\ &\quad +\mu_{i1}\rho_2(1-\rho_2) + \mu_{i4} - \rho_1\mu_{i3} - \rho_2\mu_{i2} \\ &= \mu_{i4}. \end{aligned}$$

If we continue doing the calculation, by mathematical induction if we assume that $\sigma_{i,t-1,i,t-1} = \mu_{i,t-1}$ and $\sigma_{i,t-2,i,t-2} = \mu_{i,t-2}$, then using the formula of the variance, for $t = 3, 4, ..., T$, we have:

$$
\begin{aligned}
Var(Y_{it}) &= \mu_{i,t-1}\rho_1(1-\rho_1) + \rho_1^2\sigma_{i,t-1,t-1} + \rho_2^2\sigma_{i,t-2,t-2} \\
&\quad +\mu_{i,t-2}\rho_2(1-\rho_2) + \mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} \\
&= \mu_{i,t-1}\rho_1(1-\rho_1) + \rho_1^2\mu_{i,t-1} + \rho_2^2\mu_{i,t-2} \\
&\quad +\mu_{i,t-2}\rho_2(1-\rho_2) + \mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} \\
&= \mu_{it}. \qquad (2.7)
\end{aligned}
$$

So, $Var(Y_{it}) = \mu_{it} = exp(x'_{it}\beta)$ for all $t = 1, 2, ..., T$.

#### 2.2.1.3 The Covariance

The lag k covariance between $Y_{it}$ and $Y_{i,t-k}$ will also have a recursive relationship in terms of covariance between $Y_{i,t-1}$ and $Y_{i,t-k}$ & $Y_{i,t-2}$ and $Y_{i,t-k}$. By assumption (2), we have

$$
\begin{aligned}
Cov(Y_{it}, Y_{i,t-k}) &= Cov(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1), Y_{i,t-k}) + Cov(\sum_{j=1}^{Y_{i,t-2}} b_{2j}(\rho_2), Y_{i,t-k}) \\
&\quad +Cov(d_{it}, Y_{i,t-k}) \\
&= Cov(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1), Y_{i,t-k}) + Cov(\sum_{j=1}^{Y_{i,t-2}} b_{2j}(\rho_2), Y_{i,t-k}).
\end{aligned}
$$

Consider only the first term of the equation,

$$\begin{aligned}
Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1), Y_{i,t-k}\right) &= E\left[Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1), Y_{i,t-k}\Big|_{Y_{i,t-1},Y_{i,t-k}}\right)\right] \\
&\quad + Cov\left(E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1)\Big|_{Y_{i,t-1},Y_{i,t-k}}\right], E(Y_{i,t-k}|_{Y_{i,t-1},Y_{i,t-k}})\right) \\
&= Cov_{Y_{i,t-1},Y_{i,t-k}}\left(E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1)\Big|_{Y_{i,t-1}}\right], E(Y_{i,t-k}|_{Y_{i,t-k}})\right) \\
&= Cov_{Y_{i,t-1},Y_{i,t-k}}\left(Y_{i,t-1}\rho_1, Y_{i,t-k}\right) \\
&= \rho_1 Cov(Y_{i,t-1}, Y_{i,t-k}) \\
&= \rho_1 \sigma_{i,t-1,t-k}.
\end{aligned}$$

Similarly, we can show that:

$$Cov\left(\sum_{j=2}^{Y_{i,t-2}} b_{2j}(\rho_2), Y_{i,t-k}\right) = \rho_2 \sigma_{i,t-2,t-k}.$$

Therefore, we have:

$$Cov(Y_{it}, Y_{i,t-k}) = \rho_1 \sigma_{i,t-1,t-k} + \rho_2 \sigma_{i,t-2,t-k}. \tag{2.8}$$

We can use this formula to calculate the covariances for some specific cases.

For $t = 2$, by using the properties of AR(1) Poisson process, we have

$$Cov(Y_{i1}, Y_{i2}) = \rho_1 \mu_{i1}.$$

For $t = 3$,

$$\begin{aligned} Cov(Y_{i3}, Y_{i2}) &= \rho_1\sigma_{i22} + \rho_2\sigma_{i12} = \rho_1\mu_{i2} + \rho_2\rho_1\mu_{i1}. \\ Cov(Y_{i3}, Y_{i1}) &= \rho_1\sigma_{i21} + \rho_2\sigma_{i11} = \rho_1^2\mu_{i1} + \rho_2\mu_{i1}. \end{aligned}$$

For $t = 4$,

$$\begin{aligned} Cov(Y_{i4}, Y_{i3}) &= \rho_1\sigma_{i33} + \rho_2\sigma_{i23} = \rho_1\mu_{i3} + \rho_2\rho_1\mu_{i2} + \rho_2^2\rho_1\mu_{i1}. \\ Cov(Y_{i4}, Y_{i2}) &= \rho_1\sigma_{i32} + \rho_2\sigma_{i22} = (\rho_1^2 + \rho_2)\mu_{i2} + \rho_1^2\rho_2\mu_{i1}. \\ Cov(Y_{i4}, Y_{i1}) &= \rho_1\sigma_{i31} + \rho_2\sigma_{i21} = (\rho_1^3 + 2\rho_1\rho_2)\mu_{i1}. \end{aligned}$$

By summarizing lots of the specific covariances, We have found a general formula for covariance of $Y_{it}$ and $Y_{i,t-k}$ for all $t = 2, ..., T$ and $k = 1, 2, ..., T-1$ to be

$$Cov(Y_{it}, Y_{i,t-k}) = a_{ik} \sum_{j=1}^{t-k} \rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{i,k-1}\rho_2\mu_{i,t-k}, \tag{2.9}$$

where $a_{i0} = 0$, $a_{i1} = 1$, $a_{ik} = \rho_1 a_{i,k-1} + \rho_2 a_{i,k-2}$, for $k = 2, 3, ..., T-1$.

#### 2.2.1.4 The Correlation

Once we found out the covariance between $Y_{it}$ and $Y_{i,t-k}$, the lag k correlation between $Y_{it}$ and $Y_{i,t-k}$ will simply be

$$\begin{aligned} Corr(Y_{it}, Y_{i,t-k}) &= \frac{a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{ik-1}\rho_2\mu_{i,t-k}}{\sqrt{var(Y_{it})var(Y_{i,t-k})}} \\ &= \frac{a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{ik-1}\rho_2\mu_{i,t-k}}{\sqrt{\mu_{it}\mu_{i,t-k}}}. \end{aligned} \tag{2.10}$$

We again consider some specific cases such as lag 1 and lag 2 correlations. These correlation formulas will be needed in the estimation section.

$$Corr(Y_{it}, Y_{i,t+1}) = \frac{\sum_{j=1}^{t} \rho_1 \rho_2^{j-1} \mu_{i,t+1-j}}{\sqrt{\mu_{it}\mu_{i,t+1}}} \tag{2.11}$$

$$Corr(Y_{it}, Y_{i,t+2}) = \frac{\rho_1 \sum_{j=1}^{t} \rho_1 \rho_2^{j-1} \mu_{i,t+1-j} + \rho_2 \mu_{it}}{\sqrt{\mu_{it}\mu_{i,t+1}}}. \tag{2.12}$$

Note that when $\rho_2 = 0$, the lag k covariance and correlation will reduce to

$$Cov(Y_{it}, Y_{i,t-k}) = \rho_1^k \mu_{i,t-k} \; and \; Corr(Y_{it}, Y_{t-k}) = \rho_1^k \sqrt{\frac{\mu_{i,t-k}}{\mu_i t}},$$

which is the same for AR(1) based count data model considered by Sutradhar (2010, eqns, (15)-(16), p.178). Thus, this model is an extension to the AR(1) based count data model.

## 2.2.2 Estimation of Parameters of the Lag 2 Fixed Binary Sum Infectious Disease Model

### 2.2.2.1 GQL Estimation of $\beta$

Following Sutradhar (2011, sec.6.4.2), let $\mu_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{it}, \cdots, \mu_{iT})'$ be the $T \times 1$ dimensional mean vector of $y_i = (y_{i1}, y_{i2}, \cdots, y_{it}, \cdots, y_{iT})'$. If we assume $\rho_1$, and $\rho_2$ are known, a consistent and efficient estimate of $\beta$ can be obtained by solving

the so-called generalized quasi-likelihood (GQL) estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i) = 0 \tag{2.13}$$

where, $\Sigma_i(\rho) = Cov(Y_i) = A_i^{1/2} C_i(\rho) A_i^{1/2}$, with $A_i = diag(\sigma_{i1}, ..., \sigma_{it}, ...\sigma_{iT})$ and $C_i(\rho)$ as the true correlation structure

$$C_i(\rho) = \begin{pmatrix} 1 & \rho_{i12} & \rho_{i13} & \cdots & \rho_{i1T} \\ & 1 & \rho_{i23} & \cdots & \rho_{i2T} \\ & & \cdots & \cdots & \cdots \\ & & & 1 & \rho_{i,t-1,T} \\ & & & & 1 \end{pmatrix}$$

with $\rho_{i,t-k,t} = Corr(Y_{i,t-k}, Y_{it})$ for $t = 2, ..., T$ and $k = 1, ..., T-1$. It is clear that $E\left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i)\right] = 0$, hence the GQL estimate will be a consistent estimate. This GQL estimating equation (2.13) can be solved iteratively by using the Newton Raphson iterative equation

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho) \frac{\partial \mu_i}{\partial \beta'}\right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i)\right] \Bigg|_{\beta = \hat{\beta}_{(r)}} \tag{2.14}$$

where $\hat{\beta}_{(r)}$ is the value of $\beta$ at $r$th iteration.

#### 2.2.2.2 MM Estimation of $\rho_1$ and $\rho_2$

The GQL estimating equation (2.13) may be solved for $\beta$ when the correlation structure is known. Thus, we need to estimate the parameters $\rho_1$ and $\rho_2$ in order to

obtain a good estimate for $\beta$. These two parameters can be consistently estimated by using the method of moments. Let $S_{itt}$, $S_{it,t+1}$, and $S_{it,t+2}$ be the standardized sample variance, the standardized lag 1 sample autocovariance and the standardized lag 2 sample autocovariance, respectively, defined as

$$\begin{aligned} S_{itt} &= \sum_{i=1}^{K}\sum_{t=1}^{T}\left(\frac{y_{it}-\mu_{it}}{\sigma_{it}}\right)^2/KT \\ S_{it,t+1} &= \sum_{i=1}^{K}\sum_{t=1}^{T-1}\left(\frac{y_{it}-\mu_{it}}{\sigma_{it}}\right)\left(\frac{y_{i,t+1}-\mu_{i,t+1}}{\sigma_{i,t+1}}\right)/K(T-1) \\ S_{it,t+2} &= \sum_{i=1}^{K}\sum_{t=1}^{T-2}\left(\frac{y_{it}-\mu_{it}}{\sigma_{it}}\right)\left(\frac{y_{i,t+2}-\mu_{i,t+2}}{\sigma_{i,t+2}}\right)/K(T-2), \end{aligned}$$

where $\sigma_{it}=\sqrt{\sigma_{itt}}$, $\sigma_{it+1}=\sqrt{\sigma_{it+1,t+1}}$ ,and $\sigma_{it+2}=\sqrt{\sigma_{it+2,t+2}}$. Then

$$\begin{aligned} E[S_{itt}] &= 1 \\ E[S_{it,t+1}] &= \sum_{i=1}^{K}\sum_{t=1}^{T-1} Corr(y_{it},y_{i,t+1})/K(T-1) \\ E[S_{it,t+2}] &= \sum_{i=1}^{K}\sum_{t=1}^{T-2} Corr(y_{it},y_{i,t+2})/K(T-2). \end{aligned}$$

Using first order approximation of the expectation of the ratio of two sample variance, we will have moment equations

$$\frac{S_{it,t+1}}{S_{itt}} = E\left[\frac{S_{it,t+1}}{S_{itt}}\right] \approx \frac{E[S_{it,t+1}]}{E[S_{itt}]} = E[S_{it,t+1}]$$

$$\frac{S_{it,t+2}}{S_{itt}} = E\left[\frac{S_{it,t+2}}{S_{itt}}\right] \approx \frac{E[S_{it,t+2}]}{E[S_{itt}]} = E[S_{it,t+2}].$$

Then, one may obtain the estimates for $\rho_1$ and $\rho_2$ by solving the marginal moment equations

$$\frac{S_{it,t+1}}{S_{itt}} - E[S_{it,t+1}] = 0 \tag{2.15}$$

$$\frac{S_{it,t+2}}{S_{itt}} - E[S_{it,t+2}] = 0 \tag{2.16}$$

Due to the nonlinearity of the estimating equations (2.15) and (2.16), the solutions can be obtained by using Newton's iteration method.

$$\hat{\rho}_{1(r+1)} = \hat{\rho}_{1(r)} + \left[\frac{\partial E[S_{it,t+1}]}{\partial \rho_1}\right]^{-1} \left[\frac{S_{it,t+1}}{S_{itt}} - E[S_{it,t+1}]\right] \Bigg|_{\rho_1=\hat{\rho}_{1(r)},\rho_2=\hat{\rho}_{2(r)}} \tag{2.17}$$

$$\hat{\rho}_{2(r+1)} = \hat{\rho}_{2(r)} + \left[\frac{\partial E[S_{it,t+2}]}{\partial \rho_2}\right]^{-1} \left[\frac{S_{it,t+2}}{S_{itt}} - E[S_{it,t+2}]\right] \Bigg|_{\rho_1=\hat{\rho}_{1(r)},\rho_2=\hat{\rho}_{2(r)}} \tag{2.18}$$

where $\hat{\rho}_{1(r)}$ and $\hat{\rho}_{2(r)}$ are the values of $\rho_1$ and $\rho_2$ at $r$th iteration respectively.

### 2.2.3 Forecasting Performance

Once all parameters of the model (2.1) have been estimated, we can carry out a one-step forecast for the purpose of planning and control. From model (2.1), it is clear that the conditional mean of $Y_{it}$ given $Y_{i,t-1}$ and $Y_{i,t-2}$ will have the formula

$$E(Y_{it}|y_{i,t-1}, y_{i,t-2}) = \mu_{it} + \rho_1(y_{i,t-1} - \mu_{i,t-1}) + \rho_2(y_{i,t-2} - \mu_{i,t-2}). \tag{2.19}$$

Next, if we define an $l$-step ahead forecasting function of $y_{i,t+l}$ as $y_{it}(l) = \hat{y}_{i,t+l} = E(Y_{i,t+l}|y_{i,t+l-1}, y_{i,t+l-2})$, then, from (2.17), the one step ahead forecasting function is given by

$$\begin{aligned} y_{it}(1) &= E(Y_{i,t+1}|y_{it}, y_{i,t-1}) \\ &= \mu_{i,t+1} + \rho_1(y_{it} - \mu_{it}) + \rho_2(y_{i,t-1} - \mu_{i,t-1}), \end{aligned} \tag{2.20}$$

with $y_{it}(0) = y_{it}$. Once we have a one step ahead forecast, we can calculate the forecast error $e_{it}(1)$ by using

$$\begin{aligned} e_{it}(1) &= Y_{i,t+1} - Y_{it}(1) \\ &= (y_{i,t+1} - \mu_{i,t+1}) - \rho_1(y_{it} - \mu_{it}) - \rho_2(y_{i,t-1} - \mu_{i,t-1}). \end{aligned} \tag{2.21}$$

From the above equation, we noticed that the conditional mean

$$E(e_{it}(1)|y_{it}, y_{i,t-1}) = E(Y_{i,t+1}|y_{it}, y_{i,t-1}) - E(E(Y_{i,t+1}|y_{it}, y_{i,t-1})|y_{it}, y_{i,t-1}) = 0$$

and the mean of $e_{it}(1)$ is

$$E(e_{it}(1)) = E(E(e_{it}(1)|y_{it}, y_{i,t-1})) = 0.$$

The conditional variance of $e_{it}(1)|y_{it}, y_{i,t-1}$ is given by

$$\begin{aligned} Var(e_{it}(1)|y_{it}, y_{i,t-1}) &= Var(Y_{i,t+1} - Y_{it}(1)|y_{it}, y_{i,t-1}) \\ &= Var(Y_{i,t+1}|y_{it}, y_{i,t-1}) \\ &= \mu_{i,t+1} - \rho_1\mu_{it} - \rho_2\mu_{i,t-1} + y_{it}\rho_1(1-\rho_1) + y_{i,t-1}\rho_1(1-\rho_2). \end{aligned}$$

Then, the variance of $e_{it}(1)$ follows the formula

$$\begin{aligned} Var(e_{it}(1)) &= E(Var(e_{it}(1)|y_{it}, y_{i,t-1})) + Var(E(e_{it}(1)|y_{it}, y_{i,t-1})) \\ &= E(Var(e_{it}(1)|y_{it}, y_{i,t-1})) \\ &= E(\mu_{i,t+1} - \rho_1\mu_{it} - \rho_2\mu_{i,t-1} + y_{it}\rho_1(1-\rho_1) + y_{i,t-1}\rho_1(1-\rho_2)) \\ &= \mu_{i,t+1} - \rho_1^2\mu_{it} - \rho_2^2\mu_{i,t-1}. \end{aligned} \tag{2.22}$$

### 2.2.4 Simulation Study

In this section, we perform a simulation study. We consider the case of $K = 100$ communities and $T = 5$ time points. We will use $Y_{it}$, $t = 1, 2, 3, 4$ for the purpose of estimation and try to forecast the number of infections at $t = 5$ in each community. First, we consider a time independent covariate vector $x'_{it} = (x_{it1}, x_{it2})$ for the stationary case, where $x_{it1}$ and $x_{it2}$ are generated as follows:

$$x_{it1} = \begin{cases} -0.5, & t = 1, 2, 3, 4, 5; i = 1, 2, ..., \frac{K}{2} \\ 0.5, & t = 1, 2, 3, 4, 5; i = \frac{K}{2} + 1, 2, ..., K \end{cases} \tag{2.23}$$

and

$$x_{it2} = \begin{cases} 0, & t = 1,2,3,4,5; i = 1,2,...,\frac{K}{2} \\ 1, & t = 1,2,3,4,5; i = \frac{K}{2}+1,...,K. \end{cases} \tag{2.24}$$

Then we consider a time dependent covariate vector $x'_{it} = (x_{it1}, x_{it2})$ for studying the nonstationary case, where $x_{it1}$ and $x_{it2}$ are generated as follows:

$$x_{it1} = \begin{cases} -1, & t = 1,2; i = 1,2,...,\frac{K}{2} \\ 1, & t = 3,4,5; i = 1,2,...,\frac{K}{2} \\ 0, & t = 1; i = \frac{K}{2}+1,...,K \\ 0.5, & t = 2,3; i = \frac{K}{2}+1,...,K \\ 1, & t = 4,5; i = \frac{K}{2}+1,...,K \end{cases} \tag{2.25}$$

and

$$x_{it2} = \begin{cases} \frac{t}{T}, & t = 1,2,3,4,5; i = 1,2,...,\frac{K}{4} \\ -1, & t = 1; i = \frac{K}{4}+1,...,\frac{3K}{4} \\ 0, & t = 2.3; i = \frac{K}{4}+1,...,\frac{3K}{4} \\ 0.5, & t = 4,5; i = \frac{K}{4}+1,...,\frac{3K}{4} \\ \frac{0.5+(t-1)0.5}{T}, & t = 1,2,3,4,5; i = \frac{3K}{4}+1,...,K. \end{cases} \tag{2.26}$$

Even though we have only two covariates in simulation, however, in practical cases, these covariates can represent more factors. These covariates can be time independent

factors, such as geographic locations and policy restrictions, or, they can also be time dependent factors, such as economic situations and age. From the assumptions, we know that $0 \leq \rho_1 \leq min\left(\frac{\mu_{i2}}{\mu_{i1}}, \frac{\mu_{it}-\rho_2\mu_{i,t-2}}{\mu_{i,t-1}}, 1\right)$, $for\ t \geq 3$. Since $\rho_2$ is the lag 2 correlation, we can naturally assume that $\rho_2 < \rho_1$. So, we choose a small $\rho_2$, then compute the upper bond $\rho_1^* = min\left(\frac{\mu_{i2}}{\mu_{i1}}, \frac{\mu_{it}-\rho_2\mu_{i,t-2}}{\mu_{i,t-1}}, 1\right)$. Then, we use this $\rho_2$ and $\rho_1 = \rho_1^* - 0.1$ or $\rho_1 = \rho_1^* - 0.2$ as the true values of $\rho_1$ and $\rho_2$ for the simulation. Using suitable initial values of $\beta$, $\rho_1$ and $\rho_2$, we solve the marginal estimating equation for $\beta$ by using Newton Raphson algorithm. Then by using the initial values of $\rho_1$ & $\rho_2$ and the estimate of $\beta$ obtained from previous step, we obtain estimates for $\rho_1$ & $\rho_2$ by using moment estimating equations. We use estimated $\rho_1$ ,$\rho_2$ to estimate $\beta$ again, then use this new $\beta$ and repeat the above steps to get the improved estimates of $\rho_1$ and $\rho_2$. This iterative step continues until convergence. The table below shows estimated $\beta$ and $\rho_1, \rho_2$ from 1000 simulations.

Table 2.1: Stationary Model Parameters Estimation Results.

| | | | Parameter Estimation | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\beta$ | $\rho_1$ | $\rho_2$ | $\hat{\beta}$ | $SE_{\hat{\beta}}$ | $\hat{\rho}_1$ | $SE_{\hat{\rho}_1}$ | $\hat{\rho}_2$ | $SE_{\hat{\rho}_2}$ |
| (0.5, 1.0) | 0.40 | 0.10 | (0.508, 0.994) | (0.224, 0.123) | 0.388 | 0.060 | 0.106 | 0.059 |
| (1.0, 1.0) | 0.40 | 0.10 | (1.000, 1.000) | (0.260, 0.136) | 0.388 | 0.060 | 0.102 | 0.057 |
| (0.5, 1.0) | 0.35 | 0.20 | (0.517, 0.992) | (0.224, 0.128) | 0.345 | 0.055 | 0.178 | 0.064 |
| (1.0, 1.0) | 0.35 | 0.20 | (1.018, 0.990) | (0.257, 0.138) | 0.350 | 0.057 | 0.176 | 0.064 |
| (0.5, 1.0) | 0.60 | 0.20 | (0.522, 0.987) | (0.283, 0.155) | 0.571 | 0.057 | 0.177 | 0.070 |
| (1.0, 1.0) | 0.60 | 0.20 | (1.052, 0.974) | (0.304, 0.162) | 0.575 | 0.059 | 0.170 | 0.067 |
| (0.5, 1.0) | 0.75 | 0.20 | (0.524, 0.991) | (0.320, 0.178) | 0.705 | 0.051 | 0.163 | 0.040 |
| (1.0, 1.0) | 0.75 | 0.20 | (1.037, 0.982) | (0.351, 0.187) | 0.707 | 0.066 | 0.159 | 0.066 |
| (0.5, 1.0) | 0.60 | 0.30 | (0.510,0.992) | (0.287, 0.161) | 0.572 | 0.056 | 0.248 | 0.066 |
| (1.0, 1.0) | 0.60 | 0.30 | (1.038, 0.980) | (0.330, 0.175) | 0.571 | 0.057 | 0.247 | 0.067 |
| (0.5, 1.0) | 0.45 | 0.40 | (0.517, 0.990) | (0.268, 0.146) | 0.453 | 0.047 | 0.328 | 0.054 |
| (1.0, 1.0) | 0.45 | 0.40 | (1.037, 0.981) | (0.307, 0.163) | 0.430 | 0.059 | 0.349 | 0.064 |

Table 2.2: Nonstationary Model Parameters Estimation Results.

| | | | Parameter Estimation | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\beta$ | $\rho_1$ | $\rho_2$ | $\hat{\beta}$ | $SE_{\hat{\beta}}$ | $\hat{\rho}_1$ | $SE_{\hat{\rho}_1}$ | $\hat{\rho}_2$ | $SE_{\hat{\rho}_2}$ |
| (0.5, 1.0) | 0.40 | 0.10 | (0.500, 0.997) | (0.070, 0.131) | 0.393 | 0.074 | 0.128 | 0.080 |
| (1.0,1.0) | 0.40 | 0.10 | (1.003, 0.995) | (0.070, 0.123) | 0.393 | 0.084 | 0.145 | 0.091 |
| (0.5, 1.0) | 0.35 | 0.20 | (0.502, 0.996) | (0.072, 0.135) | 0.362 | 0.073 | 0.174 | 0.084 |
| (1.0,1.0) | 0.35 | 0.20 | (1.002, 0.992) | (0.073, 0.128) | 0.366 | 0.075 | 0.175 | 0.093 |
| (0.5, 1.0) | 0.60 | 0.20 | (0.497, 1.000) | (0.069, 0.131) | 0.587 | 0.078 | 0.197 | 0.104 |
| (1.0,1.0) | 0.60 | 0.20 | (1.002, 0.995) | (0.067, 0.119) | 0.586 | 0.081 | 0.210 | 0.119 |
| (0.5, 1.0) | 0.75 | 0.20 | (0.500, 0.994) | (0.064, 0.124) | 0.725 | 0.072 | 0.189 | 0.102 |
| (1.0,1.0) | 0.75 | 0.20 | (1.002, 0.996) | (0.067, 0.117) | 0.723 | 0.078 | 0.209 | 0.129 |
| (0.5, 1.0) | 0.60 | 0.30 | (0.495, 1.004) | (0.070, 0.140) | 0.593 | 0.075 | 0.264 | 0.107 |
| (1.0,1.0) | 0.60 | 0.30 | (1.002, 0.994) | (0.071, 0.125) | 0.596 | 0.082 | 0.266 | 0.120 |
| (0.5, 1.0) | 0.45 | 0.40 | (0.498, 0.998) | (0.071, 0.142) | 0.473 | 0.067 | 0.307 | 0.091 |
| (1.0, 1.0) | 0.45 | 0.40 | (0.996, 1.003) | (0.072, 0.131) | 0.475 | 0.076 | 0.288 | 0.105 |

From Table 2.1 and Table 2.2, we can see that the estimates for $\beta$ are very close to the true value of $\beta$ irrespective of the combinations of parameters. However, for some combinations of $\rho_1$ and $\rho_2$, the estimates for $\rho_1$ and $\rho_2$ may not be as accurate as the others, for instance, under stationary case, when true $\rho_2 = 0.20$, according to the $rho_1$ restriction from our model assumption, $rho_1$ should be less the $1 - \rho_2 = 0.8$. If the true combination is $\rho_1 = 0.75$ and $\rho_1 = 0.20$ , the estimates are less close to the true values compare to other combinations. Similar result happens when true $\rho_1 = 0.45$ or $0.55$, $\rho_2 = 0.40$. The upper boundary for $\rho_2$ is 0.5 because we have assumed that $\rho_1 \geq \rho_2$. These estimates are less close to the true values than others because $\rho_1$ or $\rho_2$ are close to its boundary.

For the purpose of examining the forecast performance of the model (2.1) in forecasting the future infections, we use the parameter estimates obtained by using only the first four observations, $Y_{i1}, Y_{i2}, Y_{i3}$ and $Y_{i4}$ for $i = 1, 2, ..., 100$ and the forecasting

function in Section 2.2.3 to compute a one-step ahead forecast of the fifth observation. The sum of squares of the forecast error as well as the variance of the forecast error for these 100 communities were calculated for each simulation run. We denote the average sum of squares of the forecast errors and the average variance of the forecast error by ASS and AV respectively. The results.summarized from 1000 simulations, are reported in Table 2.3 and Table 2.4.

Table 2.3: Stationary Model Forecasting Error.

| $\beta$ | $\rho_1$ | $\rho_2$ | $ASS$ | $AV$ |
|---|---|---|---|---|
| (0.5, 1.0) | 0.40 | 0.10 | 1.658 | 1.743 |
| (1.0, 1.0) | 0.40 | 0.10 | 2.153 | 2.122 |
| (0.5, 1.0) | 0.35 | 0.20 | 1.802 | 1.798 |
| (1.0, 1.0) | 0.35 | 0.20 | 2.159 | 2.134 |
| (0.5, 1.0) | 0.60 | 0.20 | 1.298 | 1.353 |
| (1.0, 1.0) | 0.60 | 0.20 | 1.561 | 1.605 |
| (0.5, 1.0) | 0.75 | 0.20 | 0.872 | 1.007 |
| (1.0, 1.0) | 0.75 | 0.20 | 1.036 | 1.189 |
| (0.5, 1.0) | 0.60 | 0.30 | 1.190 | 1.288 |
| (1.0, 1.0) | 0.60 | 0.30 | 1.431 | 1.543 |
| (0.5, 1.0) | 0.45 | 0.40 | 1.381 | 1.456 |
| (1.0, 1.0) | 0.45 | 0.40 | 1.658 | 1.743 |

Table 2.4: Nonstationary Model Forecasting Error.

| $\beta$ | $\rho_1$ | $\rho_2$ | $ASS$ | $AV$ |
|---|---|---|---|---|
| (0.5, 1.0) | 0.40 | 0.10 | 2.726 | 2.654 |
| (1.0, 1.0) | 0.40 | 0.10 | 4.539 | 4.360 |
| (0.5, 1.0) | 0.35 | 0.20 | 2.775 | 2.695 |
| (1.0, 1.0) | 0.35 | 0.20 | 4.623 | 4.422 |
| (0.5, 1.0) | 0.60 | 0.20 | 2.103 | 2.047 |
| (1.0, 1.0) | 0.60 | 0.20 | 3.477 | 3.368 |
| (0.5, 1.0) | 0.75 | 0.20 | 1.490 | 1.534 |
| (1.0, 1.0) | 0.75 | 0.20 | 2.518 | 2.513 |
| (0.5, 1.0) | 0.60 | 0.30 | 1.993 | 1.971 |
| (1.0, 1.0) | 0.60 | 0.30 | 3.329 | 3.326 |
| (0.5, 1.0) | 0.45 | 0.40 | 2.329 | 2.298 |
| (1.0, 1.0) | 0.45 | 0.40 | 3.906 | 3.831 |

From Table (2.3) and Table (2.4), we see that the value of average sum of squares and the average variance of the forecast errors are very close to each other for all different combinations of parameters. This indicates that the average sum of squares of the forecast errors can closely estimate the average variance of the forecast errors and a satisfactory performance of the estimation of the parameters of the model. It could also be seen that the average variance and sum of squares of forecast errors for the nonstationary model are generally larger than the stationary case which means we have more accurate estimates for the parameters in stationary case. This makes sense because in stationary case, the covariates does not change with respect to time $t$. Hence less variation was introduced into the modelling. Note that for $\beta = (0.5, 1)$, the average variance of forecasting error is smaller than that of $\beta = (1, 1)$. This is because the mean function contains $\beta$ and the forecasting variance is a function of means. In our particular setup, $\beta = (0.5, 1)$ will lead to a smaller sum of means than

$\beta = (1, 1)$. For example, in stationary case, we have $x'_{i,t+1} = x'_{it} = x'_{i,t-1} = x'_i$ and $\mu_{i,t+1} = \mu_{it} = \mu_{i,t-1} = \mu_i$. If we let $\mu_{ia}$, $\mu_{ib}$ and represent the means when $\beta = (0.5, 1)$ and $\beta = (1, 1)$ respectively. $Var(e_{ia}(1))$ and $Var(e_{ib})$ are defined by (2.22) using $\mu_{ia}$ and $\mu_{ib}$ respectively. We also let $AV_a$ and $AV_b$ be the average variance of forecast errors when $\beta = (0.5, 1)$ and $\beta = (1, 1)$ respectively, Then

$$AV_a = \frac{1}{K}\sum_{i=1}^{K} Var(e_{ia}(1)) = (1 - \rho_1^2 - \rho_2^2)\frac{1}{K}\sum_{i=1}^{K}\mu_{ia},$$

$$AV_b = \frac{1}{K}\sum_{i=1}^{K} Var(e_{ib}(1)) = (1 - \rho_1^2 - \rho_2^2)\frac{1}{K}\sum_{i=1}^{K}\mu_{ib}.$$

Considering the stationary covariate structure as (2.21) and (2.22), we find out that $\sum_{i=1}^{K}\mu_{ia} \leq \sum_{i=1}^{K}\mu_{ib}$. By definition of a poisson distribution, we have $\mu_i - \rho_1\mu_i - \rho_2\mu_i \geq 0$, and we know that $\rho_1^2 \leq \rho_1$ and $\rho_2^2 \leq \rho_2$ since our $\rho_1$ and $\rho_2$ are numbers within interval 0 to 1. So $1 - \rho_1^2 - \rho_2^2 \geq 0$. Therefore,

$$AV_a - AV_b = (1 - \rho_1^2 - \rho_2^2)\frac{1}{K}\sum_{i=1}^{K}(\mu_{ia} - \mu_{ib}) \leq 0$$

Therefore, under stationary case, we expect to see that the average variance of forecasting error is smaller when $\beta = (0.5, 1)$ than when $\beta = (1, 1)$, regardless the combination of $\rho_1$ and $\rho_2$. This situation still holds for nonstationary case in our particular setup.

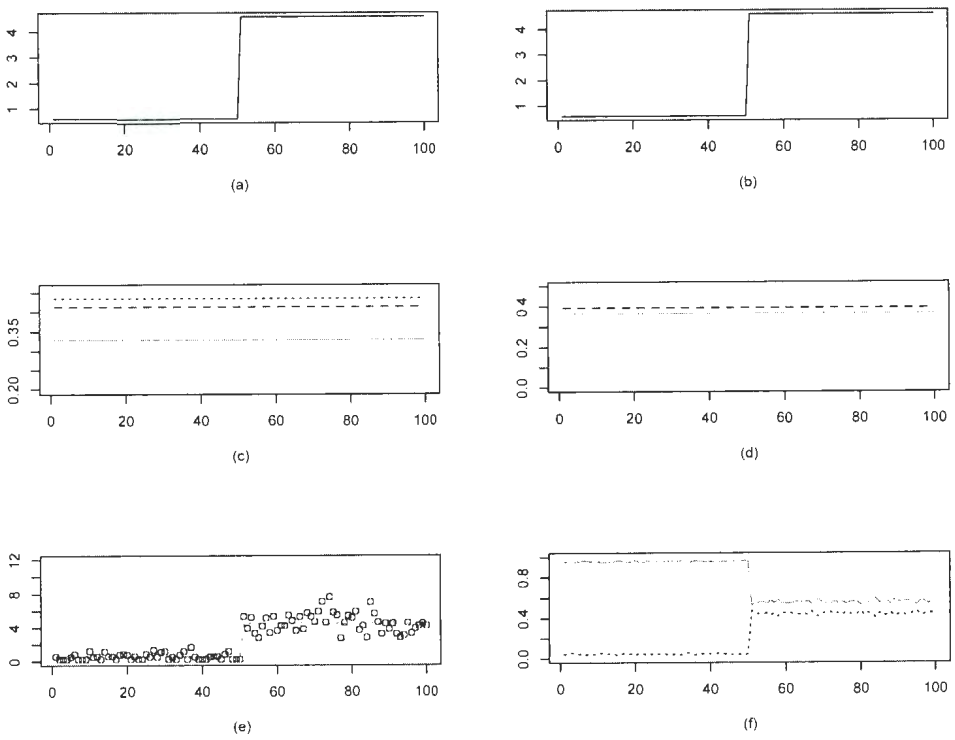


Figure 2.1: A plot of (a) values of stationary mean for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (b) values of stationary variance for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (c) values of stationary lag 1 correlation for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line); (d) values of stationary lag 2 correlation for $t = 1$ (solid line), $t = 2$ (dashed line); (e) Average forecast overlaid on average of longitudinal data; and (f) proportion of absolute values of forecast error that are 0 or 1 (solid line) and $> 1$ (dotted line); by communities obtained from 1000 simulations with $\rho_1 = 0.35$, $\rho_2$=0.20, $\beta$=(1,1), stationary covariates (2.23)-(2.24)

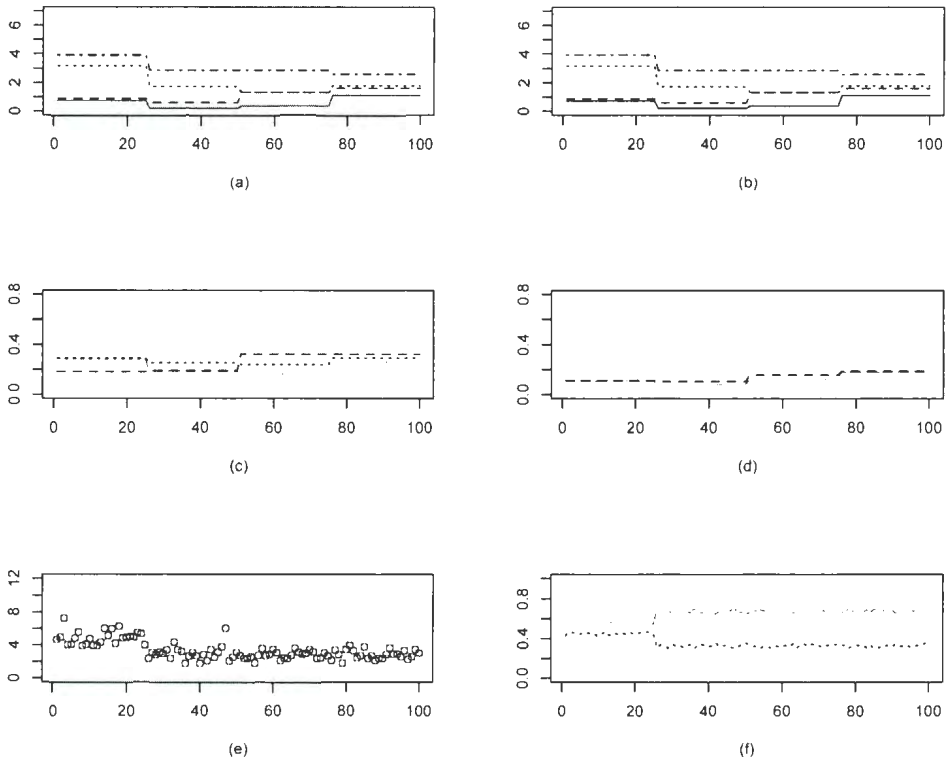


Figure 2.2: A plot of (a) values of nonstationary mean for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (b) values of nonstationary variance for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (c) values of nonstationary lag 1 correlation for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line); (d) values of nonstationary lag 2 correlation for $t = 1$ (solid line), $t = 2$ (dashed line); (e) Average forecast overlaid on average of longitudinal data; and (f) proportion of absolute values of forecast error that are 0 or 1 (solid line) and $> 1$ (dotted line); by communities obtained from 1000 simulations with $\rho_1 = 0.35$, $\rho_2$=0.20, $\beta$=(1,1), nonstationary covariates (2.25)-(2.26)

Figure 2.1(a),(b),(c) and (d) show the stationary patterns in the mean $\mu_{it}$, variance $\sigma_{itt}$, lag 1 correlation $\rho_{i,t-1,t}$, and lag 2 correlation $\rho_{i,t-2,t}$. In Figure 2.1(e), we have overlaid a graph of the average of the forecast in 1000 simulations over a scatterplot of the average of the observations $y_{i5}$. The plot shows that the average forecast follows

the general pattern of the infections at the fifth time point. In order to assess the accuracy of out forecasts, we have also displayed a graph showing the average of the proportions of the forecast error $e_{it}$ with absolute deviations 0,1 and greater than 1. Figure 2.1(f) shows that the deviations of magnitude 0 and 1 appear to be over 90% for the first 50 communities and around 60% for the remaining 50 communities. The deviations of magnitude for 0 & 1 is about 60% for the last 50 communities is caused by the large variation in the number of infections for these communities as seen in Figure 2.1 (e). For the purpose of comparing the difference between the stationary case and nonstationary case, we constructed similar plots in Figure 2.2 for a nonstationary case obtained from covariates generated by using (2.25) and (2.26).

## 2.3 Lag 2 Mixed Binary Sum Infectious Disease Mixed Model

In Section 2.2 we have discussed the model under the assumption that there is no community effect. In this section, we will discuss the model with an unobservable community effect. Suppose that for the $i$th community, there exists a community effect $\gamma_i$ and $\gamma_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Conditional on this $i$th community effect $\gamma_i$, a dynamic

mixed model for the number of infections at time t can be written as

$$\begin{aligned} Y_{i1}|_{\gamma_i} &\sim Poi(\mu^*_{i1}), where\ \mu^*_{i1} = e^{x'_{i1}\beta+\gamma_i} \\ Y_{i2}|_{\gamma_i} &= \sum_{j=1}^{Y_{i1}} b_{1j}(\rho_1)|_{\gamma_i} + d_{i2}|_{\gamma_i} \\ Y_{it}|_{\gamma_i} &= \sum_{j=1}^{Y_{i,t-1}} b_{1j}(\rho_1)|_{\gamma_i} + \sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t,\rho_2)|_{\gamma_i} + d_{it}|_{\gamma_i},\ for\ t = 3,4,...,T, \end{aligned} \quad (2.27)$$

where $b_{1j} \sim Bin(\rho_1), b_{2j} \sim Bin(\rho_2)$, with the following assumptions:

(1) $d_{i2} \sim Poi(\mu^*_{i2} - \rho_1 n_2 \mu^*_{i1})$, where $\mu^*_{i2} = e^{x'_{i2}\beta+\gamma_i}$.

(2) $d_{it} \sim Poi(\mu^*_{it} - \rho_1 n_t \mu^*_{i,t-1} - \rho_2 n_t \mu^*_{i,t-2})$.where $\mu^*_{it} = e^{x'_{it}\beta+\gamma_i}$,

(3) $Y_{i,t-1}|_{\gamma_i}$& $d_{it}|_{\gamma_i}$ are independent for $t = 2,3,...,T$.

(4) $Y_{i,t-2}|_{\gamma_i}$& $d_{it}|_{\gamma_i}$ are independent for $t = 3,4,...,T$.

From the model (2.21), it is clear that $E[Y_{i1}|_{\gamma_i}] = Var(Y_{i1}|_{\gamma_i}) = \mu^*_{i1}$, $E[Y_{i2}|_{\gamma_i}] = Var(Y_{i2}|_{\gamma_i}) = \mu^*_{i2}$, $Cov(Y_{i1}, Y_{i2}|_{\gamma_i}) = \rho_1\mu^*_{i1}$ and $Corr(Y_{i1}, Y_{i2}|_{\gamma_i}) = \rho_1\sqrt{\frac{\mu^*_{i1}}{\mu^*_{i2}}}$. For the model to be well-defined, we require that $\mu^*_{i2} - \rho_1\mu^*_{i1} \geq 0$, yielding $\rho_1 \leq \frac{\mu^*_{i2}}{\mu^*_{i1}}$. Similarly, the condition $\mu_{it} - \rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} \geq 0$ leads to $\rho_1 \leq \frac{\mu_{it}-\rho_2\mu_{i,t-2}}{\mu_{i1}}$. Therefore, the range of $\rho_1$ is

$$0 \leq \rho_1 \leq min(\frac{\mu^*_{i2}}{\mu^*_{i1}}, \frac{\mu^*_{it} - \rho_2\mu^*_{i,t-2}}{\mu^*_{i,t-1}}, 1),\ for\ t \geq 3.$$

In the stationary case, $\mu^*_{i1} = \mu^*_{i2} = \cdots = \mu^*_{iT} = \mu^*_i$, and the range of $\rho_1$ can be simplified to

$$0 \leq \rho_1 \leq (1 - \rho_2).$$

### 2.3.1 Basic Properties of the Proposed Mixed Model

#### 2.3.1.1 Conditional Properties

This model is a generalization of model (2.1). Conditional on $\gamma_i$, the model become exactly the same as model (2.1) with a different mean parameter $\mu_{it}^* = exp(x_{it}'\beta + \gamma_i)$. Therefore, all the conditional properties are the same as the properties in Section 2.3.1. That is

$$E[Y_{it}|_{\gamma_i}] = \mu_{it}^* \tag{2.28}$$

$$Var(Y_{it}|_{\gamma_i}) = \mu_{it}^* \tag{2.29}$$

$$Cov(Y_{it}, Y_{i,t-k}|_{\gamma_i}) = a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k}^* + a_{i,k-1}\rho_2\mu_{i,t-k}^* \tag{2.30}$$

$$Corr(Y_{it}, Y_{i,t-k}|_{\gamma_i}) = \frac{a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k}^* + a_{i,k-1}\rho_2\mu_{i,t-k}^*}{\sqrt{\mu_{it}^*\mu_{i,t-k}^*}}, \tag{2.31}$$

where $a_{i0} = 0$, $a_{i1} = 1$, $a_{ik} = \rho_1 a_{i,k-1} + \rho_2 a_{i,k-2}$, for $k = 2, 3, ..., T-1$.

#### 2.3.1.2 Unconditional Properties

In order to proceed to the estimation and the forecasting, we need to find the unconditional properties of the mixed model. By using the assumption that $\gamma_i \overset{iid}{\sim} N(0, \sigma^2)$, we can use moment generating function or direct integration to easily find out that

$$E[e^{\gamma_i}] = e^{\sigma^2/2}, E[e^{2\gamma_i}] = e^{2\sigma^2}, E[e^{3\gamma_i}] = e^{9\sigma^2/2} \text{ and } E[e^{4\gamma_i}] = e^{8\sigma^2}. \tag{2.32}$$

Then, we can find out the unconditional properties by taking expectation over $\gamma_i$. The unconditional mean of $Y_{it}$ is

$$\begin{aligned} E[Y_{it}] &= E_{\gamma_i} E[Y_{it}|_{\gamma_i}] \\ &= E[\mu^*_{it}] \\ &= E[exp(x'_{it}\beta + \gamma_i)] \\ &= exp(x'_{it}\beta) E(e^{\gamma_i}) \\ &= exp(x'_{it}\beta) exp(\sigma^2/2) \\ &= exp(x'_{it}\beta + \sigma^2/2) \\ &= \mu_{it}. \end{aligned} \tag{2.33}$$

The variance of $Y_{it}$ can be calculated in a similar way by conditioning on $\gamma_i$

$$\begin{aligned} Var(Y_{it}) &= E_{\gamma_i} Var(Y_{it}|_{\gamma_i}) + Var_{\gamma_i} E(Y_{it}|_{\gamma_i}) \\ &= E[\mu^*_{it}] + Var[\mu^*_{it}] \\ &= exp(x'_{it}\beta + \sigma^2/2) + E[(\mu^*_{it})^2] - (E[\mu^*_{it}])^2 \\ &= \mu_{it} + exp(2x'_{it}\beta) E(exp(2\gamma_i)) + \mu^2_{it} \\ &= \mu_{it} + exp(2x'_{it}\beta + 2\sigma^2) + \mu^2_{it} \\ &= \mu_{it} + \mu^2_{it}(exp(\sigma^2) - 1). \end{aligned} \tag{2.34}$$

We note that as opposed to the fixed model (2.1), the unconditional mean and variance of the responses are not the same. The unconditional covariance will be

$$\begin{aligned}
Cov(Y_{it}, Y_{i,t-k}) &= E_{\gamma_i} Cov(Y_{it}, Y_{i,t-k}|_{\gamma_i}) + Cov_{\gamma_i}(E(Y_{it}|_{\gamma_i}), E(Y_{i,t-k}|_{\gamma_i})) \\
&= E\left(a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu^*_{i,t-j+1-k} + a_{i,k-1}\rho_2\mu^*_{i,t-k}\right) + Cov(\mu^*_{it}, \mu^*_{i,t-k}) \\
&= a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{i,k-1}\rho_2\mu_{i,t-k} \\
&\quad + E(\mu^*_{it}\mu^*_{i,t-k}) - E(\mu^*_{it})E(\mu^*_{i,t-k}) \\
&= a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{i,k-1}\rho_2\mu_{i,t-k} \\
&\quad + E[e^{x'_{it}\beta}e^{x'_{i,t-k}\beta}e^{2\gamma_i}] - \mu_{it}\mu_{i,t-k} \\
&= a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{i,k-1}\rho_2\mu_{i,t-k} + \mu_{it}\mu_{i,t-k}(exp(\sigma^2) - 1)
\end{aligned} \tag{2.35}$$

Finally, the lag k correlation is

$$Corr(Y_{it}, Y_{i,t-k}) = \frac{a_{ik}\sum_{j=1}^{t-k}\rho_1\rho_2^{j-1}\mu_{i,t-j+1-k} + a_{i,k-1}\rho_2\mu_{i,t-k} + \mu_{it}\mu_{i,t-k}(e^{\sigma^2} - 1)}{\sqrt{(\mu_{it} + \mu_{it}^2(e^{\sigma^2} - 1))(\mu_{i,t-k} + \mu_{i,t-k}^2(e^{\sigma^2} - 1))}} \tag{2.36}$$

where $a_{i0} = 0$, $a_{i1} = 1$, $a_{ik} = \rho_1 a_{i,k-1} + \rho_2 a_{i,k-2}$, for $k = 2, 3, ..., T-1$.

### 2.3.2 Estimation of Parameters

The dynamic mixed model (2.23) contains four unknown parameters, $\beta$, $\rho_1$, $\rho_2$ and $\sigma^2$. $\beta$ is a regression parameters involved in the mean function of $y_{it}$ which measures

the effect of the covariates, so we can use first order responses to estimate $\beta$. However, $\rho_1$, $\rho_2$ and $\sigma^2$ are involved in the variance and lag k autocovariance, we need to use all second order response to estimate those parameters. Sutradhar (2003, Section 3) and Sutradhar (2004) show that for correlated responses, one may use GQL to estimate $\beta$ and $\sigma^2$ and MM to estimate the correlation parameter.

#### 2.3.2.1 Estimation of $\beta$

For the present mixed model with the unconditional mean vector $\mu_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{it}, \cdots, \mu_{iT})'$ and covariance matrix $\Sigma_i$, respectively, following Sutradhar (2011, sec.6.4.2), for fixed $\sigma^2$, the marginal GQL estimating equation for $\beta$, is given by

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i) = 0 \tag{2.37}$$

where $\Sigma_i(\rho) = Cov(Y_i) = A_i^{1/2} C_i(\rho) A_i^{1/2}$, with $A_i = diag(\sigma_{i1}, ..., \sigma_{it}, ...\sigma_{iT})$ and $C_i(\rho)$ as the true correlation structure

$$C_i(\rho) = \begin{pmatrix} 1 & \rho_{i12} & \rho_{i13} & \cdots & \rho_{i1T} \\ & 1 & \rho_{i23} & \cdots & \rho_{i2T} \\ & & \cdots & \cdots & \cdots \\ & & & 1 & \rho_{i,t-1,T} \\ & & & & 1 \end{pmatrix}$$

with $\rho_{i,t-k,t} = Corr(Y_{i,t-k}, Y_{it})$ for $t = 2, ..., T$ and $k = 1, ..., T-1$. This GQL estimating equation (2.37) can be solved iteratively by using the Newton Raphson

algorithm

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho) \frac{\partial \mu_i}{\partial \beta'}\right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i)\right] \Bigg|_{\beta=\hat{\beta}_{(r)}} \tag{2.38}$$

where $\hat{\beta}_{(r)}$ is the value of $\beta$ at $r$th iteration.

#### 2.3.2.2 Estimation of $\rho_1$ and $\rho_2$

Similar to that of Section 2.2.2.2, let $S_{itt}$, $S_{itt+1}$, and $S_{itt+2}$ be the standardized sample variance, the standardized lag 1 sample autocovariance and the standardized lag 2 sample autocovariance, respectively, defined as

$$\begin{aligned}
S_{itt} &= \sum_{i=1}^{K}\sum_{t=1}^{T} \left(\frac{y_{it} - \mu_{it}}{\sigma_{it}}\right)^2 / KT \\
S_{it,t+1} &= \sum_{i=1}^{K}\sum_{t=1}^{T-1} \left(\frac{y_{it} - \mu_{it}}{\sigma_{it}}\right)\left(\frac{y_{i,t+1} - \mu_{i,t+1}}{\sigma_{i,t+1}}\right) / K(T-1) \\
S_{it,t+2} &= \sum_{i=1}^{K}\sum_{t=1}^{T-2} \left(\frac{y_{it} - \mu_{it}}{\sigma_{it}}\right)\left(\frac{y_{i,t+2} - \mu_{i,t+2}}{\sigma_{i,t+2}}\right) / K(T-2),
\end{aligned}$$

where $\sigma_{it} = \sqrt{\sigma_{itt}}$, $\sigma_{it+1} = \sqrt{\sigma_{it+1,t+1}}$ ,and $\sigma_{it+2} = \sqrt{\sigma_{it+2,t+2}}$. Then

$$\begin{aligned}
E[S_{itt}] &= 1 \\
E[S_{it,t+1}] &= \sum_{i=1}^{K}\sum_{t=1}^{T-1} Corr(y_{it}, y_{i,t+1}) / K(T-1) \\
E[S_{it,t+2}] &= \sum_{i=1}^{K}\sum_{t=1}^{T-2} Corr(y_{it}, y_{i,t+2}) / K(T-2).
\end{aligned}$$

Using first order approximation of the expectation of the ratio of two sample variance, we will have moment equations

$$\frac{S_{it,t+1}}{S_{itt}} = E\left[\frac{S_{it,t+1}}{S_{itt}}\right] \approx \frac{E[S_{it,t+1}]}{E[S_{itt}]} = E[S_{it,t+1}]$$

$$\frac{S_{it,t+2}}{S_{itt}} = E\left[\frac{S_{it,t+2}}{S_{itt}}\right] \approx \frac{E[S_{it,t+2}]}{E[S_{itt}]} = E[S_{it,t+2}].$$

Then, one may obtain the estimates for $\rho_1$ and $\rho_2$ by solving the marginal moment equations

$$\frac{S_{it,t+1}}{S_{itt}} - E[S_{it,t+1}] = 0 \tag{2.39}$$

$$\frac{S_{it,t+2}}{S_{itt}} - E[S_{it,t+2}] = 0 \tag{2.40}$$

Due to the nonlinearity of the estimating equations (2.39) and (2.40), the solutions can be otained by using Newton iteration method.

$$\hat{\rho}_{1(r+1)} = \hat{\rho}_{1(r)} + \left[\frac{\partial E[S_{it,t+1}]}{\rho_1}\right]^{-1} \left[\frac{S_{it,t+1}}{S_{itt}} - E[S_{it,t+1}]\right] \Bigg|_{\rho_1=\hat{\rho}_{1(r)},\rho_2=\hat{\rho}_{2(r)}} \tag{2.41}$$

$$\hat{\rho}_{2(r+1)} = \hat{\rho}_{2(r)} + \left[\frac{\partial E[S_{it,t+2}]}{\rho_2}\right]^{-1} \left[\frac{S_{it,t+2}}{S_{itt}} - E[S_{it,t+2}]\right] \Bigg|_{\rho_1=\hat{\rho}_{1(r)},\rho_2=\hat{\rho}_{2(r)}} \tag{2.42}$$

where $\hat{\rho}_{1(r)}$ and $\hat{\rho}_{2(r)}$ are the values of $\rho_1$ and $\rho_2$ at $r$th iteration respectively.

#### 2.3.2.3 Esitmation of $\sigma^2$

Sutradhar(2011), Section 4.2.6.2 has shown how the marginal GQL estimation can be done for $\sigma^2$. For $t = 1, 2, ..., T$ and $k = 1, 2, ..., T-1$, let

$$u_i = (u'_{i1}, u'_{i2})'$$

be the vector of all second-order responses under the $ith$ community, where

$$u_{i1} = (y_{i1}^2, \cdots, y_{it}^2, \cdots, y_{iT}^2) : T \times 1$$

$$u_{i2} = (y_{i1}y_{i2}, \cdots, y_{ik}y_{it}, \cdots, y_{i,t-1}y_{iT}), k < t : \frac{T(T-1)}{2} \times 1.$$

Furthermore, let

$$\lambda_i = E[U_i] = (\lambda_{i11}, \cdots, \lambda_{itt}, \cdots, \lambda_{iTT}, \lambda_{i12}, \cdots, \lambda_{ijk}, \cdots, \lambda_{i(T-1)T})'$$

where

$$\lambda_{ijj} = E[Y_{it}^2] = \sigma_{itt} + \mu_{it}^2 = \mu_{it} + exp(\sigma^2)\mu_{it}^2 \tag{2.43}$$

$$\lambda_{ijk} = E[Y_{it}Y_{ik}] = \sigma_{itk} + \mu_{it}\mu_{ik}. \tag{2.44}$$

Also, Let

$$\Omega_i = Cov(U_i) = \begin{pmatrix} Cov(U_{i1}) & Cov(U_{i1}, U'_{i2}) \\ & Cov(U_{i2}) \end{pmatrix}.$$

In a similar fashion, the marginal GQL estimation equation for $\sigma^2$ is:

$$\sum_{i=1}^{K} \frac{\partial \lambda_i'}{\partial \sigma^2} \Omega_i^{-1}(u_i - \lambda_i) = 0, \tag{2.45}$$

where the elements of the vector $\frac{\partial \lambda_i'}{\partial \sigma^2}$ are given by:

$$\begin{aligned} \frac{\partial \lambda_{itt}'}{\partial \sigma^2} &= \frac{1}{2}\mu_{it} + 2\mu_{it}^2 exp(\sigma^2) \\ \frac{\partial \lambda_{itk}'}{\partial \sigma^2} &= \frac{1}{2}\sigma_{itk} + \frac{3}{2}\mu_{it}\mu_{ik} exp(\sigma^2). \end{aligned}$$

Clearly, computing the matrix $\Omega_i$ will require exact second order, third order and fourth order joint moments of $y_{it}$. However, computing third order and fourth order joint moments will require further distributional assumptions, which may not be practical. Now, since $\Omega_i$ will not affect the consistent estimation of $\sigma^2$, we shall use the assumption of conditional independence ($\rho = 0$) to obtain a 'working' $\Omega_i$. To begin the computation of the components of $\Omega_i$, we use the assumption that $\gamma_i \sim N(0, \sigma^2)$, Oyet & Sutradhar (2011) have shown that by taking expectation over $\gamma_i$ and using (2.28), we will have

$$E_{\gamma_i}[\mu_{ij}^{*2}] = \mu_{ij}^2 exp(\sigma^2), E_{\gamma_i}[\mu_{ij}^{*3}] = \mu_{ij}^3 exp(3\sigma^2) \; and \; E_{\gamma_i}[\mu_{ij}^{*4}] = \mu_{ij}^4 exp(6\sigma^2).$$

By Mckenzie (1988), for a Poisson AR(1) model, if $X_{t-1}$ is Poisson($\theta$), then, by using alternate probability generating function (a.p.g.f), it is easy to verify that $\alpha * X_{t-1}$ is Poisson($\alpha\theta$), where " * " denote a binomial thinning operation, that is: $\alpha * X = \sum_{k=1}^{x} B_k(\alpha)$, where $B_k(\alpha)$ is a sequence of independent identically distributed binary

random variables with $P[B(\alpha) = 1] = \alpha = 1 - P[B(\alpha) = 0]$. In our case, we know that $Y_{i1}|_{\gamma_i} \sim Poi(\mu_{i1})$. Given this community effect $\gamma_i$, individuals within the $i$th community at time point 1 are independent and will cause a new infection with probability $\rho_1$. Therefore, $\sum_{j=1}^{Y_{i1}} b_{1j}(\rho_1)|_{\gamma_i} \sim Poi(\rho_1\mu_{i1}^*)$. We also know that $d_{i2}|_{\gamma_i} \sim Poi(\mu_{i2}^* - \rho_1\mu_{i1}^*)$ and it is independent with $Y_{i1}|_{\gamma_i}$. By using the properties of Poisson distribution,

$$Y_{i2}|_{\gamma_i} = \sum_{j=1}^{Y_{i1}} b_{1j}(\rho_1)|_{\gamma_i} + d_{i2}|_{\gamma_i} \sim Poi(\rho_1\mu_{i1}^* + \mu_{i2}^* - \rho_1\mu_{i1}^*) = Poi(\mu_{i2}^*).$$

Under the 'working' conditional independence ($\rho = 0$)case, $Y_{i1}|_{\gamma_i}, Y_{i2}|_{\gamma_i}, ..., Y_{it}|_{\gamma_i}$ are independent. Since $Y_{i1}|_{\gamma_i} \sim Poi(\mu_{i1}^*)$, $Y_{i2}|_{\gamma_i} \sim Poi(\mu_{i2}^*)$ and $d_{i3}|_{\gamma_i} \sim Poi(\mu_{i3}^* - \rho_1\mu_{i2}^* - \rho_2\mu_{i1}^*)$, in addition, all of them are conditional independent, then

$$\sum_{j=1}^{Y_{i2}} b_{1j}(\rho_1)|_{\gamma_i} \sim Poi(\rho_1\mu_{i2}^*), \sum_{j=1}^{Y_{i1}} b_{2j}(\rho_2)|_{\gamma_i} \sim Poi(\rho_2\mu_{i1}^*) \text{ and } Y_{i3}|_{\gamma_i,\rho=0} \sim Poi(\mu_{i3}^*).$$

By using the mathematical induction, we can conclude that

$$Y_{it}|_{\gamma_i,\rho=0} \sim Poi(\mu_{it}^*). \tag{2.46}$$

For $Y_{it}|_{\gamma_i,\rho=0} \sim Poi(\mu^*_{it})$, the raw moments can be calculated by using the moment generating function $exp[\mu^*_{it}(e^{\mu^*_{it}} - 1)]$. We have

$$\begin{aligned}
E(Y_{it}|\gamma_i, \rho = 0) &= \mu^*_{it} \\
E(Y^2_{it}|\gamma_i, \rho = 0) &= \mu^*_{it} + \mu^{*\,2}_{it} \\
E(Y^3_{it}|\gamma_i, \rho = 0) &= \mu^*_{it} + 3\mu^{*\,2}_{it} + \mu^{*\,3}_{it} \\
E(Y^4_{it}|\gamma_i, \rho = 0) &= \mu^*_{it} + 7\mu^{*\,2}_{it} + 6\mu^{*\,3}_{it} + \mu^{*\,4}_{it}.
\end{aligned}$$

After we find out the conditional distribution and its moments, the elements in $\Omega_i$ can be easily computed. By Oyet & Sutradhar (2011), eqn. 4.13 (see also Sutradhar & Bari (2007)), these conditional moments can be calculated as

$$\begin{aligned}
E(Y^2_{it}|\rho = 0) &= \mu_{it} + \mu^2_{it} exp(\sigma^2) \\
E(Y_{iu}Y_{it}|\rho = 0) &= \mu_{iu}\mu_{it} exp(\sigma^2) \\
E(Y^4_{it}|\rho = 0) &= \mu_{it} + 7\mu^2_{it} exp(\sigma^2) + 6\mu^3_{it} exp(3\sigma^2) + \mu^4_{it} exp(6\sigma^2) \\
E(Y^2_{iu}Y^2_{it}|\rho = 0) &= [1 + \{\mu_{iu} + \mu_{it}\} exp(2\sigma^2) + \mu_{iu}\mu_{it} exp(5\sigma^2)]\mu_{iu}\mu_{it} exp(\sigma^2) \\
E(Y^3_{iu}Y_{it}|\rho = 0) &= [1 + 3\mu_{iu} exp(2\sigma^2) + \mu^2_{iu} exp(5\sigma^2)]\mu_{iu}\mu_{it} exp(\sigma^2) \qquad (2.47) \\
E(Y^2_{iu}Y_{iv}Y_{it}|\rho = 0) &= [1 + \mu_{iu} exp(3\sigma^2)]\mu_{iu}\mu_{iv}\mu_{it} exp(3\sigma^2) \\
E(Y_{iu}Y_{iv}Y_{is}Y_{it}|\rho = 0) &= \mu_{iu}\mu_{iv}\mu_{is}\mu_{it} exp(6\sigma^2).
\end{aligned}$$

The conditional moments have been used for computing the elements of $\Omega_i$. For instance,

$$\begin{aligned} Cov(Y_{it}^2, Y_{iu}Y_{is}|\rho=0) &= E[Y_{it}^2 Y_{iu}Y_{is}|\rho=0] - E[Y_{it}^2]E[Y_{iu}Y_{is}|\rho=0] \\ Cov(Y_{it}^2, Y_{it}Y_{iu}|\rho=0) &= E[Y_{it}^3 Y_{iu}|\rho=0] - E[Y_{it}^2]E[Y_{it}Y_{iu}|\rho=0] \\ Cov(Y_{it}Y_{iu}, Y_{iu}Y_{is}|\rho=0) &= E[Y_{iu}^2 Y_{it}Y_{is}|\rho=0] - E[Y_{it}Y_{iu}]E[Y_{iu}Y_{is}|\rho=0]. \end{aligned}$$

The solution of the estimating equation takes the form (2.39). Its solution can be obtained by using Gauss-Newton iterative equation:

$$\hat{\sigma}_\gamma^2(r+1) = \hat{\sigma}_\gamma^2(r) + [\sum_{i=1}^{K} \frac{\partial \lambda_i'}{\partial \sigma^2}\Omega_i^{-1}\frac{\partial \lambda_i}{\partial \sigma^2}]_r^{-1}[\sum_{i=1}^{K} \frac{\partial \lambda_i'}{\partial \sigma^2}\Omega_i^{-1}(\mu_i - \lambda_i)]\Big|_{\sigma^2=\sigma_{\gamma(r)}^2}. \quad (2.48)$$

### 2.3.3 Simulation Study

In this section, we consider the nonstationary covariates (2.25) and (2.26). By choosing suitable initial values of $\beta$, $\sigma^2$ , $\rho_1$ and $\rho_2$, we numerically solve the marginal estimating equation for $\beta$ by using Newton Raphson algorithm. Next, using estimates of $\beta$ from previous step and initial $\sigma^2$, we obtain estimates for $\rho_1$ & $\rho_2$ by using moment estimating equations. Then, using the estimates of $\rho_1$ & $\rho_2$ and the estimate of $\beta$ obtained from previous steps, we solve the marginal estimating equation for $\sigma^2$ by using the Newton Raphson iterative procedure. We then use the estimates of $\rho_1$, $\rho_2$ and $\sigma^2$ to estimate $\beta$ again, then use this $\beta$ and repeat the above steps until convergence.

Table 2.5: Non-stationary $\beta$ Estimation for fixed $\rho_1$, $\rho_2$ and $\sigma^2$

| | | $\sigma^2 = 0.25$ | $\sigma^2 = 0.75$ |
|---|---|---|---|
| | | $\rho_1 = 0.40, \rho_2 = 0.15$ , $\rho_1 = 0.50, \rho_2 = 0.20$ | $\rho_1 = 0.40, \rho_2 = 0.15$ , $\rho_1 = 0.50, \rho_2 = 0.20$ |
| (0.5,1.0) | SM | (0.503,0.987) , (0.501,0.998) | (0.499,1.003) , (0.501,0.999) |
| | SSE | (0.072,0.143) , (0.069,0.137) | (0.063,0.136) , (0.062,0.130) |
| (1.0,1.0) | SM | (0.998,1.002) , (0.996,1.006) | (0.996,1.004) , (1.001,0.999) |
| | SSE | (0.074,0.136) , (0.074,0.128) | (0.072,0.128) , (0.071,0.124) |

Table 2.6: Nonstationary $\rho_1$ and $\rho_2$ Estimation for fixed $\beta$ and $\sigma^2$

| | | $\sigma^2 = 0.25$ | $\sigma^2 = 0.75$ |
|---|---|---|---|
| | | $\beta = (0.50, 1.0), \beta = (1.0, 1.0)$ | $\beta = (0.50, 1.0), \beta = (1.0, 1.0)$ |
| $\rho_1 = 0.40, \rho_2 = 0.15$ | SM | (0.411,0.156) , (0.434,0.180) | (0.456,0.173) , (0.507,0.218) |
| | SSE | (0.108,0.101) , (0.125,0.118) | (0.177,0.124) , (0.186,0.152) |
| $\rho_1 = 0.50, \rho_2 = 0.20$ | SM | (0.488,0.190) , (0.512,0.217) | (0.509,0.204) , (0.541,0.224) |
| | SSE | (0.109,0.109) , (0.126,0.129) | (0.180,0.137) , (0.193,0.157) |

Table 2.7: Non-stationary $\sigma^2$ Estimation for fixed $\beta$, $\rho_1$ and $\rho_2$

| | | $\beta = (0.50, 1.0)$ | $\beta = (1.0, 1.0)$ |
|---|---|---|---|
| | | $\rho_1 = 0.40, \rho_2 = 0.15$ , $\rho_1 = 0.50, \rho_2 = 0.20$ | $\rho_1 = 0.40, \rho_2 = 0.15$ , $\rho_1 = 0.50, \rho_2 = 0.20$ |
| $\sigma^2$=0.25 | SM | 0.237 , 0.244 | 0.239 , 0.242 |
| | SSE | 0.075 , 0.062 | 0.079 , 0.065 |
| $\sigma^2$=0.75 | SM | 0.731 , 0.748 | 0.739 , 0.730 |
| | SSE | 0.176 , 0.202 | 0.204 , 0.197 |

Table 2.5, Table 2.6 and Table 2.7 are obtained by estimating a single parameter with other parameters fixed. Table 2.5 and Table 2.7 suggest that the GQL approach works very well for estimating the covariate effect $\beta$ and the variance component in the latent community effect $\sigma^2$. However, Table 2.6 shows that the moment estimates of the correlation parameters $\rho_1$ and $\rho_2$ are less accurate especially when $\sigma^2$ or $\beta$ get larger.

Table 2.8: Non-stationary Parameters Estimation

| $\beta$ | $\rho_1$ | $\rho_2$ | $\sigma^2$ | $\hat{\beta}$ | $SE_{\hat{\beta}}$ | $\hat{\rho}_1$ | $SE_{\hat{\rho}_1}$ | $\hat{\rho}_2$ | $SE_{\hat{\rho}_2}$ | $\hat{\sigma}^2_\gamma$ | $SE_{\hat{\sigma}^2_\gamma}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (0.5,1.0) | 0.40 | 0.15 | 0.25 | (0.500,0.995) | (0.072,0.152) | 0.427 | 0.123 | 0.227 | 0.107 | 0.248 | 0.096 |
| (0.5,1.0) | 0.40 | 0.15 | 0.75 | (0.494,0.995) | (0.064,0.133) | 0.479 | 0.148 | 0.239 | 0.111 | 0.732 | 0.102 |
| (0.5,1.0) | 0.50 | 0.20 | 0.25 | (0.499,0.997) | (0.067,0.142) | 0.496 | 0.120 | 0.240 | 0.105 | 0.246 | 0.096 |
| (0.5,1.0) | 0.50 | 0.20 | 0.75 | (0.495,0.990) | (0.060,0.128) | 0.506 | 0.140 | 0.242 | 0.111 | 0.728 | 0.098 |
| (1.0,1.0) | 0.40 | 0.15 | 0.25 | (0.993,0.993) | (0.077,0.135) | 0.450 | 0.128 | 0.268 | 0.121 | 0.261 | 0.081 |
| (1.0,1.0) | 0.40 | 0.15 | 0.75 | (0.982,0.986) | (0.071,0.125) | 0.456 | 0.144 | 0.231 | 0.122 | 0.763 | 0.076 |
| (1.0,1.0) | 0.50 | 0.20 | 0.25 | (0.995,0.987) | (0.075,0.131) | 0.501 | 0.124 | 0.271 | 0.114 | 0.261 | 0.082 |
| (1.0,1.0) | 0.50 | 0.20 | 0.75 | (0.984,0.993) | (0.071,0.121) | 0.484 | 0.138 | 0.239 | 0.126 | 0.762 | 0.071 |

In Table 2.8, we report the estimates of all parameters with their standard errors. These results follow the general conclusion that we have made from Table 2.5 , 2.6 and 2.7. During the iteration process, $\rho_1$ and $\rho_2$ could sometimes fall outside the range of restrictions especially when $\rho_1$ and $\rho_2$ are close to their lower or upper bound. In this case, new observations are generated. Therefore, the overall mean of $\rho_1$ and $\rho_2$ estimates from 1000 simulations will be affected by using only $\rho_1$ and $\rho_2$ estimates that satisfy the conditions, especially for $\rho_1$ or $\rho_2$ close to the boundary.

# Chapter 3

# Lag 2 Dynamic Binomial Sum Infectious Disease Model

## 3.1 Preliminaries

In Chapter 2, we have assumed that each infected individual can only infect none or one individual each time. However, the more common case is that the infected individual could infect up to more than one individuals at a time. In this chapter, we extend model (2.1) in order to consider this more practical situation. Instead of using the binary sum, we use binomial sum in our model. For simplicity, we only discuss the binomial sum model without considering the community effect. Therefore, our mean function will depend on the covariate effects only. Similar to what we have done in Chapter 2, we will discuss the structures and assumptions of the model. Then, we will find some basic properties, such as the mean, variance, covariance and correlation for this model. Since the properties of a binomial distribution is different from that

of a binary distribution, we expect to have some slightly different properties. Finally, we will estimate the parameters and obtain forecasts to check the performance of the model. The main goal of this chapter again is to estimate the parameters involved in the model and forecast the number of infections in each community at time point $t+1$. We use GQL approach to estimate the regression parameters. We estimate the longitudinal correlation parameters by using the Method of Moments (MM). Once all parameters are properly estimated, we can check the forecast performance of this model.

## 3.2 Lag 2 Binomial Sum Infectious Disease Model

For modelling without community effect, we can assume that K independent communities are at risk of an infectious disease. At initial time point, $t = 1$, we assume that $y_{i1}$ individuals in the $i$th community have developed the disease where $y_{i1}$ follows a Poisson distribution with mean parameter $\mu_{i1} = exp(x'_{i1}\beta)$. Because we assume an infected individual can affect up to more than one individual over two time intervals, and also because there may be other infected individuals arriving from other communities, we shall model the number of infected people in the $i$th community at time $t$ as

$$\begin{aligned} Y_{i1} &\sim Poi(\mu_{i1}), where\ \mu_{i1} = e^{x'_{i1}\beta} \\ Y_{i2} &= \sum_{j=1}^{Y_{i1}} b_{1j}(n_t, \rho_1) + d_{i2} \\ Y_{it} &= \sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1) + \sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2) + d_{it},\ for\ t = 3, 4, ..., T, \end{aligned} \tag{3.1}$$

where $b_{1j} \sim Bin(n_t, \rho_1), b_{2j} \sim Bin(n_t, \rho_2)$. Then, we make the following assumptions about model (3.1)

(1) $d_{i2} \sim Poi(\mu_{i2} - n_t\rho_1\mu_{i1})$.

(2) $d_{it} \sim Poi(\mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2})$.

(3) $Y_{i,t-1}$& $d_{it}$ *are independent for* $t = 2, 3, ..., T$.

(4) $Y_{i,t-2}$& $d_{it}$ *are independent for* $t = 3, ..., T$.

By model (3.1), $E[Y_{i1}] = Var(Y_{i1}) = \mu_{i1}$, $E[Y_{i2}] = \mu_{i2}$, $Var(Y_{i2}) = \mu_{i2} + \rho_1^2 n_2(n_2 - 1)\mu_{i1}$, $Cov(Y_{i1}, Y_{i2}) = \rho_1 n_1 \mu_{i1}$ and $Corr(Y_{i1}, Y_{i2}) = n_1\rho_1\sqrt{\frac{\mu_{i1}}{\mu_{i2}+\rho_1^2 n_2(n_2-1)\mu_{i1}}}$. Since the mean of the Poisson r.v.s $Y_{i1}$ and $d_{it}$ have to be positive, we need to have $\rho_1 \leq \frac{\mu_{i2}}{n_t\mu_{i1}}$ and $\rho_1 \leq \frac{\mu_{it}-\rho_2 n_t\mu_{i,t-2}}{n_t\mu_{i,t-1}}$. Therefore, the range of $\rho_1$ is

$$0 \leq \rho_1 \leq min\left(\frac{\mu_{i2}}{n_t\mu_{i1}}, \frac{\mu_{it} - \rho_2 n_t\mu_{i,t-2}}{n_t\mu_{i,t-1}}, 1\right), \ for\ t \geq 3.$$

For stationary case, if we let $\mu_{i1} = \mu_{i2} = \cdots = \mu_{iT} = \mu_i$, then the range of $\rho_1$ and $\rho_2$ will simplify to

$$0 \leq \rho_1 \leq \frac{1 - \rho_2 n_t}{n_t}$$

and

$$0 \leq \rho_2 \leq \frac{1}{n_t}.$$

Note that if we assume there is a community effect $\gamma_i$ present and $\gamma_i \overset{\text{iid}}{\sim} N(0, \sigma_\gamma^2)$, then, conditional on this $i$th community effect $\gamma_i$, we may construct our model in a similar structure as model (2.25). We are using a binomial sum instead of a binary

sum. A dynamic mixed model for the number of infections at time t can then be written as

$$\begin{aligned} Y_{i1}|_{\gamma_i} &\sim Poi(\mu^*_{i1}), where\ \mu^*_{i1} = e^{x'_{i1}\beta+\gamma_i} \\ Y_{i2}|_{\gamma_i} &= \sum_{j=1}^{Y_{i1}} b_{1j}(n_t,\rho_1)|_{\gamma_i} + d_{i2}|_{\gamma_i} \\ Y_{it}|_{\gamma_i} &= \sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t,\rho_1)|_{\gamma_i} + \sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t,\rho_2)|_{\gamma_i} + d_{it}|_{\gamma_i},\ for\ t = 3,4,...,T, \end{aligned} \quad (3.2)$$

where $b_{1j} \sim Bin(n_t,\rho_1), b_{2j} \sim Bin(n_t,\rho_2)$, with the following assumptions:

(1) $d_{i2} \sim Poi(\mu^*_{i2} - \rho_1 n_2 \mu^*_{i1})$, where $\mu^*_{i2} = e^{x'_{i2}\beta+\gamma_i}$.

(2) $d_{it} \sim Poi(\mu^*_{it} - \rho_1 n_t \mu^*_{i,t-1} - \rho_2 n_t \mu^*_{i,t-2})$, where $\mu^*_{it} = e^{x'_{it}\beta+\gamma_i}$.

(3) $Y_{i,t-1}|_{\gamma_i}$& $d_{it}|_{\gamma_i}$ are independent for $t = 2,3,...,T$.

(4) $Y_{i,t-2}|_{\gamma_i}$& $d_{it}|_{\gamma_i}$ are independent for $t = 3,4,...,T$.

Similar to the relationship between model (2.1) and model (2.25), this mixed model (3.2) is an extension to model (3.1). All the conditional properties should match with the properties of model (3.1). The unconditional properties can then be found by taking expectations over $\gamma_i$. We can use similar approaches as in Section 2.3 to estimate the parameters once all conditional and unconditional moments are obtained.

### 3.2.1 Moments of Lag 2 Binary Sum Infectious Disease Model

#### 3.2.1.1 The Mean

Based on our previous discussion, we know that $E[Y_{i1}] = \mu_{i1}$ & $E[Y_{i2}] = \mu_{i2}$. Then, it follows that for $t = 3, 4, ...T$,

$$\begin{aligned}
E[Y_{it}] &= E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1) + \sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2) + d_{it}\right] \\
&= E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1)\right] + E\left[\sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2)\right] + E[d_{it}] \\
&= E_{y_{i,t-1}} E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1)\Big|_{y_{i,t-1}}\right] + E_{y_{i,t-2}} E\left[\sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2)\Big|_{y_{i,t-2}}\right] + E[d_{it}] \\
&= n_t\rho_1 E[Y_{i,t-1}] + n_t\rho_2 E[Y_{i,t-2}] + \mu_{it} - \rho_1 n_t \mu_{i,t-1} - \rho_2 n_t \mu_{i,t-2} \qquad (3.3)
\end{aligned}$$

Next, we consider some specific cases:

For $t = 3$,

$$\begin{aligned}
E[Y_{i3}] &= n_3\rho_1 E[Y_{i2}] + n_3\rho_2 E[Y_{i1}] + \mu_{i3} - n_3\rho_1\mu_{i2} - n_3\rho_2\mu_{i1} \\
&= n_3\rho_1\mu_{i2} + n_3\rho_2\mu_{i1} + \mu_{i3} - n_3\rho_1\mu_{i2} - n_3\rho_2\mu_{i1} \\
&= \mu_{i3}.
\end{aligned}$$

For $t = 4$,

$$\begin{aligned} E[Y_{i4}] &= n_4\rho_1 E[Y_{i3}] + n_4\rho_2 E[Y_{i2}] + \mu_{i4} - n_4\rho_1\mu_{i3} - n_4\rho_2\mu_{i2} \\ &= n_4\rho_1\mu_{i3} + n_4\rho_2\mu_{i2} + \mu_{i4} - n_4\rho_1\mu_{i3} - n_4\rho_2\mu_{i2} \\ &= \mu_{i4}. \end{aligned}$$

By mathematical induction, if we have $E[Y_{i,t-1}] = \mu_{i,t-1}$ and $E[Y_{i,t-2}] = \mu_{i,t-2}$, then:

$$\begin{aligned} E[Y_{it}] &= n_t\rho_1\mu_{i,t-1} + n_t\rho_2\mu_{i,t-2} + \mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2} \\ &= \mu_{it} \\ &= exp(x'_{it}\beta). \end{aligned} \tag{3.4}$$

So, $E[Y_{it}] = \mu_{it} = exp(x'_{it}\beta)$, for all $t = 1, 2, ..., T$.

#### 3.2.1.2 The Variance

The variance of this model can be derived by finding the conditional and unconditional variances. By assumption (3), (4) and the property that covariance of two constants is zero, we have

$$Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}, d_{it}\Big|_{Y_{i,t-1},Y_{i,t-2}}\right) = 0 \ , \ Cov\left(\sum_{j=1}^{Y_{i,t-2}} b_{2j}, d_{it}\Big|_{Y_{i,t-1},Y_{i,t-2}}\right) = 0$$

and

$$Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}, \sum_{j=1}^{Y_{i,t-2}} b_{2j}\Big|_{Y_{i,t-1},Y_{i,t-2}}\right) = 0.$$

Then

$$\begin{aligned} Var(Y_{it}|Y_{i,t-1}, Y_{i,t-2}) &= Y_{i,t-1}Var(b_{1j}) + Y_{i,t-2}Var(b_{2j}) + Var(d_{it}) \\ &= Y_{i,t-1}n_t\rho_1(1-\rho_1) + Y_{i,t-2}n_t\rho_2(1-\rho_2) \\ &\quad +\mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2}. \end{aligned} \tag{3.5}$$

Letting $\sigma_{i,t-1,t-1}$ represent the variance of $Y_{i,t-1}$, then

$$\begin{aligned} Var(Y_{it}|Y_{i,t-2}) &= E_{Y_{i,t-1}}[Var(Y_{it}|Y_{i,t-1}, Y_{i,t-2})] + Var_{Y_{i,t-1}}(E[Y_{it}|Y_{i,t-1}, Y_{i,t-2}]) \\ &= E_{Y_{i,t-1}}[Y_{i,t-1}n_t\rho_1(1-\rho_1) + Y_{i,t-2}n_t\rho_2(1-\rho_2) \\ &\quad +(\mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2})] \\ &\quad +Var_{Y_{i,t-1}}(Y_{i,t-1}n_t\rho_1 + Y_{i,t-2}n_t\rho_2 + (\mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2})) \\ &= \mu_{i,t-1}n_t\rho_1(1-\rho_1) + Y_{i,t-2}n_t\rho_2(1-\rho_2) + n_t^2\rho_1^2\sigma_{i,t-1,t-1} \\ &\quad +\mu_{it} - n_t\rho_1\mu_{i,t-1} - \rho_2 n_t\mu_{i,t-2}. \end{aligned} \tag{3.6}$$

Similarly, letting $\sigma_{i,t-2,t-2}$ represent the variance of $Y_{i,t-2}$, we have

$$\begin{aligned}
Var(Y_{it}) &= E_{Y_{i,t-2}}[Var(Y_{it}|Y_{i,t-2})] + Var_{Y_{i,t-2}}(E[Y_{it}|Y_{i,t-2}]) \\
&= E_{Y_{i,t-2}}[\mu_{i,t-1}n_t\rho_1(1-\rho_1) + Y_{i,t-2}n_t\rho_2(1-\rho_2) \\
&\quad +\mu_{it} - n_t\rho_1\mu_{i,t-1} - \rho_2\mu_{i,t-2} + n_t^2\rho_1^2\sigma_{i,t-1,t-1}] \\
&\quad +Var_{Y_{i,t-2}}(E_{Y_{i,t-1}}E[Y_{it}|Y_{i,t-1}, Y_{i,t-2}]) \\
&= \mu_{i,t-1}n_t\rho_1(1-\rho_1) + \mu_{i,t-2}n_t\rho_2(1-\rho_2) \\
&\quad +\mu_{it} - n_t\rho_1\mu_{i,t-1} - \rho_2 n_t\mu_{i,t-2} + n_t^2\rho_1^2\sigma_{i,t-1,t-1} \\
&\quad +Var(\mu_{i,t-1}n_t\rho_1 + Y_{i,t-2}n_t\rho_2 + \mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2}) \\
&= \mu_{i,t-1}n_t\rho_1(1-\rho_1) + \mu_{i,t-2}n_t\rho_2(1-\rho_2) + n_t^2\rho_1^2\sigma_{i,t-1,t-1} + n_t^2\rho_2^2\sigma_{i,t-2,t-2} \\
&\quad +\mu_{it} - n_t\rho_1\mu_{i,t-1} - n_t\rho_2\mu_{i,t-2} \\
&= \mu_{it} - \mu_{i,t-1}n_t\rho_1^2 - \mu_{i,t-2}n_t\rho_2^2 + n_t^2\rho_1^2\sigma_{i,t-1,t-1} + n_t^2\rho_2^2\sigma_{i,t-2,t-2} \\
&= \mu_{it} - (\mu_{i,t-1} - n_t\sigma_{i,t-1,t-1})n_t\rho_1^2 - (\mu_{i,t-2} - n_t\sigma_{i,t-2,t-2})n_t\rho_2^2. \qquad (3.7)
\end{aligned}$$

From (3.7), we can see that the variance of $Y_{it}$ has a recursive relationship with the variance of $Y_{i,t-1}$ and the variance of $Y_{i,t-2}$. It is difficult to find a closed expression for each individual variance. So we list some specific examples. We know from our assumptions that $Var(Y_{i1}) = \mu_{i1}$ and $Var(Y_{i2}) = \mu_{i2} + \rho_1^2 n_2(n_2-1)\mu_{i1}$ and, we can

find that when $t = 3$,

$$\begin{aligned} Var(Y_{i3}) &= n_3\mu_{i2}\rho_1(1-\rho_1) + n_3^2\rho_1^2\sigma_{i22} + n_3^2\rho_2^2\sigma_{i11} \\ &\quad +\mu_{i1}n_3\rho_2(1-\rho_2) + \mu_{i3} - n_3\rho_1\mu_{i2} - n_3\rho_2\mu_{i1} \\ &= \mu_{i2}n_3\rho_1(1-\rho_1) + n_3^2\rho_1^2(\mu_{i2} + \rho_1^2 n_2(n_2-1)\mu_{i1}) + n_3^2\rho_2^2\mu_{i1} \\ &\quad +\mu_{i1}n_3\rho_2(1-\rho_2) + \mu_{i3} - n_3\rho_1\mu_{i2} - n_3\rho_2\mu_{i1} \\ &= \mu_{i3} + \rho_1^2 n_3(n_3-1)\mu_{i2} + \rho_2^2 n_3(n_3-1)\mu_{i1} + \rho_1^4 n_3^2 n_2(n_2-1)\mu_{i1}. \quad (3.8) \end{aligned}$$

When $t = 4$,

$$\begin{aligned} Var(Y_{i4}) &= \mu_{i2}n_4\rho_1(1-\rho_1) + n_4^2\rho_1^2\sigma_{i33} + n_4^2\rho_2^2\sigma_{i22} + \mu_{i1}n_4\rho_2(1-\rho_2) \\ &\quad +\mu_{i4} - n_4\rho_1\mu_{i3} - n_4\rho_2\mu_{i2} \\ &= \mu_{i4} + \rho_1^2 n_4(n_4-1)\mu_{i3} + \rho_2^2 n_4(n_4-1)\mu_{i2} + \rho_1^4 n_4^2 n_3(n_3-1)\mu_{i2} \\ &\quad +\rho_1^2\rho_2^2 n_4^2 n_2(n_2-1)\mu_{i1} + \rho_1^2\rho_2^2 n_4^2 n_3(n_3-1)\mu_{i1} \\ &\quad +\rho_1^6 n_4^2 n_3^2 n_2(n_2-1)\mu_{i1}. \quad (3.9) \end{aligned}$$

#### 3.2.1.3 The Covariance

The lag k covariance between $Y_{it}$ and $Y_{i,t-k}$ will also have a recursive relationship in terms of covariance between $Y_{i,t-1}$ and $Y_{i,t-k}$ & $Y_{i,t-2}$ and $Y_{i,t-k}$. By assumption

(3) and (4), we have

$$\begin{aligned} Cov(Y_{it}, Y_{i,t-k}) &= Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1), Y_{i,t-k}\right) + Cov\left(\sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2), Y_{i,t-k}\right) \\ &\quad + Cov(d_{it}, Y_{i,t-k}) \\ &= Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1), Y_{i,t-k}\right) + Cov\left(\sum_{j=1}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2), Y_{i,t-k}\right). \end{aligned}$$

Considering only the first term of the equation,

$$\begin{aligned} Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1), Y_{i,t-k}\right) &= E\left[Cov\left(\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1), Y_{i,t-k}\Big|_{Y_{i,t-1}, Y_{i,t-k}}\right)\right] \\ &\quad + Cov\left(E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1)\Big|_{Y_{i,t-1}, Y_{i,t-k}}\right], E(Y_{i,t-k}|_{Y_{i,t-1}, Y_{i,t-k}})\right) \\ &= Cov_{Y_{i,t-1}, Y_{i,t-k}}\left(E\left[\sum_{j=1}^{Y_{i,t-1}} b_{1j}(n_t, \rho_1)\Big|_{Y_{i,t-1}}\right], E(Y_{i,t-k}|_{Y_{i,t-k}})\right) \\ &= Cov_{Y_{i,t-1}, Y_{i,t-k}}(Y_{i,t-1} n_t \rho_1, Y_{i,t-k}) \\ &= n_t \rho_1 Cov(Y_{i,t-1}, Y_{i,t-k}) \\ &= n_t \rho_1 \sigma_{i,t-1,t-k}. \end{aligned}$$

Similarly, we can show that:

$$Cov\left(\sum_{j=2}^{Y_{i,t-2}} b_{2j}(n_t, \rho_2), Y_{i,t-k}\right) = n_t \rho_2 \sigma_{i,t-2,t-k}.$$

Therefore, we have:

$$Cov(Y_{it}, Y_{i,t-k}) = n_t \rho_1 \sigma_{i,t-1,t-k} + n_t \rho_2 \sigma_{i,t-2,t-k} \tag{3.10}$$

We can use this formula to obtain the covariances for some specific cases:

For $t = 2$, Oyet & Sutradhar (2011) have shown that

$$Cov(Y_{i1}, Y_{i2}) = n_2\rho_1\mu_{i1}.$$

For $t = 3$,

$$\begin{aligned} Cov(Y_{i3}, Y_{i2}) &= n_3\rho_1\sigma_{i22} + n_3\rho_2\sigma_{i12} = \rho_1 n_3\mu_{i2} + \rho_1^3 n_3 n_2(n_2-1)\mu_{i1} + \rho_1\rho_2 n_2\mu_{i1}. \\ Cov(Y_{i3}, Y_{i1}) &= n_3\rho_1\sigma_{i21} + n_3\rho_2\sigma_{i11} = \rho_1^2 n_3 n_2\mu_{i1} + \rho_2 n_3\mu_{i1}. \end{aligned}$$

For $t = 4$,

$$\begin{aligned} Cov(Y_{i4}, Y_{i3}) &= \rho_1\sigma_{i33} + \rho_2\sigma_{i23} \\ &= \rho_1 n_4\mu_{i3} + \rho_1^3 n_4 n_3(n_3-1)\mu_{i2} + \rho_1\rho_2^2 n_4 n_3(n_3-1)\mu_{i1} + \rho_1^5 n_4 n_3^2(n_2-1)\mu_{i1} \\ &\quad +\rho_1\rho_2 n_4 n_3\mu_{i2} + \rho_1^3\rho_2 n_4 n_3 n_2(n_2-1)\mu_{i1} + \rho_1\rho_2^2 n_4 n_2\mu_{i1}. \\ Cov(Y_{i4}, Y_{i2}) &= \rho_1\sigma_{i32} + \rho_2\sigma_{i22} \\ &= \rho_1^2 n_4 n_3\mu_{i2} + \rho_1^4 n_4 n_3 n_2(n_2-1)\mu_{i1} + \rho_1^2\rho_2 n_4 n_2\mu_{i1} \\ &\quad +\rho_2 n_4\mu_{i2} + \rho_1^2\rho_2 n_4 n_2(n_2-1)\mu_{i1}. \\ Cov(Y_{i4}, Y_{i1}) &= \rho_1\sigma_{i31} + \rho_2\sigma_{i21} \\ &= \rho_1^3 n_4 n_3 n_2\mu_{i1} + \rho_1\rho_2 n_4 n_3\mu_{i1} + \rho_1\rho_2 n_4 n_2\mu_{i1}. \end{aligned}$$

#### 3.2.1.4 The Correlation

Once we have found the covariance between $Y_{it}$ and $Y_{i,t-k}$, the lag k correlation between $Y_{it}$ and $Y_{i,t-k}$ will simply be

$$Corr(Y_{it}, Y_{i,t-k}) = \frac{Cov(Y_{it}, Y_{i,t-k})}{\sqrt{var(Y_{it})var(Y_{i,t-k})}}. \tag{3.11}$$

Note that when $n_t = 1, t = 1, 2, ..., T$, the lag k covariance and correlation will reduce to the covariance and correlation formulas (2.9) and (2.10) for the binary sum infectious disease model considered in Chapter 2. Thus, this model can be considered a generalization of the binary sum infectious disease model. When $\rho_2 = 0$, the lag k covariance and correlation will reduce to

$$Cov(Y_{it}, Y_{i,t-k}) = \left( \prod_{l=0}^{k-l} n_{t-l} \right) \rho_1^k \sigma_{i,t-k,t-k}$$

and

$$Corr(Y_{it}, Y_{t-k}) = \left( \prod_{l=0}^{k-l} n_{t-l} \right) \rho_1^k \sqrt{\frac{\sigma_{i,t-k,t-k}}{\sigma_{itt}}},$$

which are the same expressions for lag 1 infectious disease model considered by Oyet & Sutradhar (2011, eqns, (2.6) (2.7)). Thus, this model is also an extension to the lag 1 infectious disease model.

### 3.2.2 Estimation of the Parameters of the Lag 2 Binary Sum Infectious Disease Model

Similar to what we have done in Section 2.2.2, we use GQL approach to estimate the covariate effects and use MM approach to estimate $\rho_1$ and $\rho_2$.

#### 3.2.2.1 GQL Estimation of $\beta$

Following Sutradhar (2011, sec.6.4.2), let $\mu_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{it}, \cdots, \mu_{iT})'$ be the $T \times 1$ dimensional mean vector of $y_i = (y_{i1}, y_{i2}, \cdots, y_{it}, \cdots, y_{iT})'$. If we assume $\rho_1$, and $\rho_2$ are known, a consistent and efficient estimate of $\beta$ can be obtained by solving the so-called generalized quasi-likelihood (GQL) estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i) = 0 \tag{3.12}$$

where $\Sigma_i(\rho) = Cov(Y_i) = A_i^{1/2} C_i(\rho) A_i^{1/2}$, with $A_i = diag(\sigma_{i1}, ..., \sigma_{it}, ...\sigma_{iT})$ and $C_i(\rho)$ as the true correlation structure

$$C_i(\rho) = \begin{pmatrix} 1 & \rho_{i12} & \rho_{i13} & \cdots & \rho_{i1T} \\ & 1 & \rho_{i23} & \cdots & \rho_{i2T} \\ & & \cdots & \cdots & \cdots \\ & & & 1 & \rho_{i,t-1,T} \\ & & & & 1 \end{pmatrix}$$

with $\rho_{i,t-k,t} = Corr(Y_{i,t-k}, Y_{it})$, for $t = 2, ..., T$ and $k = 1, ..., T-1$. This GQL estimating equation (3.12) can be solved iteratively by using the Newton Raphson

algorithm

$$\hat{\beta}_{(r+1)} = \hat{\beta}_{(r)} + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho) \frac{\partial \mu_i}{\partial \beta'}\right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1}(\rho)(y_i - \mu_i)\right] \Bigg|_{\beta=\hat{\beta}_{(r)}} \tag{3.13}$$

where $\hat{\beta}_{(r)}$ is the value of $\beta$ at $r$th iteration.

#### 3.2.2.2 MM Estimation of $\rho_1$ and $\rho_2$

The GQL estimating equation (3.12) can be solved for $\beta$ when the correlation structure is known. Thus, we need to estimate the parameters $\rho_1$ and $\rho_2$ in order to obtain a good estimate for $\beta$. These two parameters can be consistently estimated by using the method of moments. Let $S_{itt}$, $S_{itt+1}$ and $S_{itt+2}$ be the standardized sample variance, the standardized lag 1 sample autocovariance and the standardized lag 2 sample autocovariance, respectively, defined as

$$\begin{aligned}
S_{itt} &= \sum_{i=1}^{K}\sum_{t=1}^{T} \left(\frac{y_{it} - \mu_{it}}{\sigma_{it}}\right)^2 / KT \\
S_{it,t+1} &= \sum_{i=1}^{K}\sum_{t=1}^{T-1} \left(\frac{y_{it} - \mu_{it}}{\sigma_{it}}\right)\left(\frac{y_{i,t+1} - \mu_{i,t+1}}{\sigma_{i,t+1}}\right) / K(T-1) \\
S_{it,t+2} &= \sum_{i=1}^{K}\sum_{t=1}^{T-2} \left(\frac{y_{it} - \mu_{it}}{\sigma_{it}}\right)\left(\frac{y_{i,t+2} - \mu_{i,t+2}}{\sigma_{i,t+2}}\right) / K(T-2),
\end{aligned}$$

where $\sigma_{it} = \sqrt{\sigma_{itt}}$, $\sigma_{it+1} = \sqrt{\sigma_{it+1,t+1}}$ ,and $\sigma_{it+2} = \sqrt{\sigma_{it+2,t+2}}$. Then

$$\begin{aligned} E[S_{itt}] &= 1 \\ E[S_{it,t+1}] &= \sum_{i=1}^{K}\sum_{t=1}^{T-1} Corr(y_{it}, y_{i,t+1})/K(T-1) \\ E[S_{it,t+2}] &= \sum_{i=1}^{K}\sum_{t=1}^{T-2} Corr(y_{it}, y_{i,t+2})/K(T-2). \end{aligned}$$

Using first order approximation of the expectation of the ratio of two sample variance, we will have moment equations

$$\frac{S_{it,t+1}}{S_{itt}} = E\left[\frac{S_{it,t+1}}{S_{itt}}\right] \approx \frac{E[S_{it,t+1}]}{E[S_{itt}]} = E[S_{it,t+1}]$$

$$\frac{S_{it,t+2}}{S_{itt}} = E\left[\frac{S_{it,t+2}}{S_{itt}}\right] \approx \frac{E[S_{it,t+2}]}{E[S_{itt}]} = E[S_{it,t+2}].$$

Then, one may obtain the estimates for $\rho_1$ and $\rho_2$ by solving the marginal moment equations

$$\frac{S_{it,t+1}}{S_{itt}} - E[S_{it,t+1}] = 0 \tag{3.14}$$

$$\frac{S_{it,t+2}}{S_{itt}} - E[S_{it,t+2}] = 0. \tag{3.15}$$

Due to the nonlinearity of the estimating equations (3.14) and (3.15), the solutions can be obtained by using iteration method. However, the variances contains $\rho_1$ and $\rho_2$. The practical derivatives with respect to $\rho_1$ and $\rho_2$ would be a little complicated to find if we want to use Newton's iteration. Instead, we simply iterate to convergence

the following equations:

$$\hat{\rho}_{1(r+1)} = \left(\frac{S_{it,t+1}}{S_{itt}}\right)\left[\frac{1}{\hat{\rho}_{1(r)}}\sum_{i=1}^{K}\sum_{t=1}^{T-1} Corr(y_{it}, y_{i,t+1})/K(T-1)\right]^{-1}\Bigg|_{\rho_1=\hat{\rho}_{1(r)},\rho_2=\hat{\rho}_{2(r)}} \tag{3.16}$$

$$\hat{\rho}_{2(r+1)} = \left(\frac{S_{it,t+2}}{S_{itt}}\right)\left[\frac{1}{\hat{\rho}_{2(r)}}\sum_{i=1}^{K}\sum_{t=1}^{T-2} Corr(y_{it}, y_{i,t+2})/K(T-2)\right]^{-1}\Bigg|_{\rho_1=\hat{\rho}_{1(r)},\rho_2=\hat{\rho}_{2(r)}} \tag{3.17}$$

where $\hat{\rho}_{1(r)}$ and $\hat{\rho}_{2(r)}$ are the values of $\rho_1$ and $\rho_2$ at $r$th iteration respectively.

### 3.2.3 Forecasting Performance

Once all parameters of the model (3.1) are estimated, we can obtain a one-step forecast for the purpose of planning and control. From model (3.1), it is clear that the conditional mean of $Y_{it}$ given $Y_{i,t-1}$ and $Y_{i,t-2}$ has the formula

$$E(Y_{it}|y_{i,t-1}, y_{i,t-2}) = \mu_{it} + n_t\rho_1(y_{i,t-1} - \mu_{i,t-1}) + n_t\rho_2(y_{i,t-2} - \mu_{i,t-2}). \tag{3.18}$$

Next, if we define an $l$-step ahead forecasting function of $y_{i,t+l}$ as $y_{it}(l) = \hat{y}_{i,t+l} = E(Y_{i,t+l}|\ y_{i,t+l-1}, y_{i,t+l-2})$, then, from (3.18), the one step ahead forecasting function is given by

$$\begin{aligned} y_{it}(1) &= E(Y_{i,t+1}|y_{it}, y_{i,t-1}) \\ &= \mu_{i,t+1} + n_{t+1}\rho_1(y_{it} - \mu_{it}) + n_{t+1}\rho_2(y_{i,t-1} - \mu_{i,t-1}), \end{aligned} \tag{3.19}$$

with $y_{it}(0) = y_{it}$. Once we have a one-step ahead forecast, we can calculate the forecast error $e_{it}(1)$ by using

$$\begin{aligned} e_{it}(1) &= Y_{i,t+1} - Y_{it}(1) \\ &= (y_{i,t+1} - \mu_{i,t+1}) - n_{t+1}\rho_1(y_{it} - \mu_{it}) - n_{t+1}\rho_2(y_{i,t-1} - \mu_{i,t-1}). \quad (3.20) \end{aligned}$$

From the above equation, we have

$$E(e_{it}(1)|y_{it}, y_{i,t-1}) = E(Y_{i,t+1}|y_{it}, y_{i,t-1}) - E(E(Y_{i,t+1}|y_{it}, y_{i,t-1})|y_{it}, y_{i,t-1}) = 0,$$

and the mean of $e_{it}(1)$ is

$$E(e_{it}(1)) = E(E(e_{it}(1)|y_{it}, y_{i,t-1})) = 0.$$

The conditional variance of $e_{it}(1)|y_{it}, y_{i,t-1}$ is given by

$$\begin{aligned} Var(e_{it}(1)|y_{it}, y_{i,t-1}) &= Var(Y_{i,t+1}|y_{it}, y_{i,t-1}) \\ &= \mu_{i,t+1} - \rho_1 n_{t+1}\mu_{it} - \rho_2 n_{t+1}\mu_{i,t-1} + y_{it}n_{t+1}\rho_1(1-\rho_1) \\ &\quad + y_{i,t-1}n_{t+1}\rho_1(1-\rho_2). \end{aligned}$$

Then, the variance of $e_{it}(1)$ is

$$\begin{aligned} Var(e_{it}(1)) &= E(Var(e_{it}(1)|y_{it}, y_{i,t-1})) + Var(E(e_{it}(1)|y_{it}, y_{i,t-1})) \\ &= E(Var(e_{it}(1)|y_{it}, y_{i,t-1})) \\ &= E(\mu_{i,t+1} - n_{t+1}\rho_1\mu_{it} - n_{t+1}\rho_2\mu_{i,t-1} + y_{it}n_{t+1}\rho_1(1-\rho_1) \\ &\quad + y_{i,t-1}n_{t+1}\rho_1(1-\rho_2)) \\ &= \mu_{i,t+1} - n_{t+1}\rho_1^2\mu_{it} - n_{t+1}\rho_2^2\mu_{i,t-1}. \end{aligned} \tag{3.21}$$

### 3.2.4 Simulation Study

Similar to Section 2.2.4, we conduct simulation studies with $K = 100$ communities and $T = 5$ time intervals. We will estimate the model parameters based on the counts from $t = 1, 2, 3, 4$ time points and forecast the number of infections at $t = 5$. The covariate vector $x'_{it} = (x_{it1}, x_{it2})$ and nonstationary covariate vector $x'_{it} = (x_{it1}, x_{it2})$ are the same as in (2.21) & (2.22) and (2.23) & (2.24), respectively. From the assumptions, we know that $0 \leq \rho_1 \leq min\left(\frac{\mu_{i2}}{\mu_{i1}}, \frac{\mu_{it}-\rho_2 n_t \mu_{i,t-2}}{n_t \mu_{i,t-1}}, 1\right)$, $for\ t \geq 3$. We again assume that $\rho_2 < \rho_1$. So we choose a small $\rho_2$, then compute the upper bound $\rho_1^* = min\left(\frac{\mu_{i2}}{\mu_{i1}}, \frac{\mu_{it}-\rho_2 n_t \mu_{i,t-2}}{n_t \mu_{i,t-1}}, 1\right)$. Then, we use this $\rho_2$ and $\rho_1 = \rho_1^* - 0.1$ or $\rho_1 = \rho_1^* - 0.2$ as the true values of $\rho_1$ and $\rho_2$ for the simulation. We selected the same $n_t$ values as in Oyet & Sutradhar (2011), that is, $n' = (n_1, n_2, n_3, n_4, n_5) \equiv (1,2,2,2,2), (1,2,2,3,2), (1,2,3,4,2), (1,2,2,2,3)$ , $(1,2,2,3,3), (1,2,3,4,3)$.

Using suitable initial values of $\beta$, $\rho_1$ and $\rho_2$, we solve the marginal estimating equation for $\beta$ by using Newton Raphson algorithm. Then, by using the initial values of $\rho_1$ & $\rho_2$ and the estimate of $\beta$ obtained from previous step, we obtain estimates for

$\rho_1$ & $\rho_2$ by using moment estimating equations. We use estimated $\rho_1$, $\rho_2$ to estimate $\beta$ again, then use this new $\beta$ and repeat the above steps until convergence. Table (3.1) and Table (3.2) below report the estimated $\beta$, $\rho_1$ and $\rho_2$ from 1000 simulations.

Table 3.1: Stationary Model Parameters Estimation Results.

| | | | | Parameter Estimation | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Nt$ | $\beta$ | $\rho_1$ | $\rho_2$ | $\hat{\beta}$ | $SE_{\hat{\beta}}$ | $\hat{\rho_1}$ | $SE_{\hat{\rho_1}}$ | $\hat{\rho_2}$ | $SE_{\hat{\rho_2}}$ |
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.20 | 0.05 | (0.518,0.990) | (0.230,0.126) | 0.198 | 0.032 | 0.045 | 0.032 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.20 | 0.05 | (1.031,0.984) | (0.266,0.139) | 0.197 | 0.033 | 0.048 | 0.035 |
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.15 | 0.10 | (0.505,1.000) | (0.223,0.124) | 0.156 | 0.029 | 0.084 | 0.033 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.15 | 0.10 | (1.024,0.988) | (0.254,0.135) | 0.157 | 0.031 | 0.085 | 0.033 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | (0.526,0.983) | (0.243,0.134) | 0.147 | 0.023 | 0.028 | 0.020 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | (1.028,0.985) | (0.265,0.140) | 0.146 | 0.024 | 0.029 | 0.021 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | (0.507,0.998) | (0.217,0.120) | 0.102 | 0.021 | 0.045 | 0.020 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.10 | 0.05 | (1.020,0.990) | (0.249,0.133) | 0.101 | 0.022 | 0.044 | 0.020 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | (0.506,0.995) | (0.236,0.131) | 0.148 | 0.021 | 0.030 | 0.021 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | (1.008,0.996) | (0.266,0.142) | 0.148 | 0.024 | 0.029 | 0.021 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | (0.511,0.993) | (0.216,0.118) | 0.101 | 0.022 | 0.045 | 0.020 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.10 | 0.05 | (1.024,0.990) | (0.247,0.131) | 0.101 | 0.022 | 0.044 | 0.020 |
| N2=N3=N5=2,N4 =3 | (0.5, 1.0) | 0.20 | 0.05 | (0.517,0.990) | (0.247,0.137) | 0.196 | 0.029 | 0.045 | 0.028 |
| N2=N3=N5=2,N4 =3 | (1.0, 1.0) | 0.20 | 0.05 | (1.002,0.998) | (0.278,0.150) | 0.196 | 0.029 | 0.046 | 0.029 |
| N2=N3=N5=2,N4 =3 | (0.5, 1.0) | 0.15 | 0.10 | (0.515,0.988) | (0.234,0.127) | 0.153 | 0.027 | 0.086 | 0.027 |
| N2=N3=N5=2,N4 =3 | (1.0, 1.0) | 0.15 | 0.10 | (1.024,0.988) | (0.275,0.146) | 0.153 | 0.027 | 0.089 | 0.027 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | (0.510,0.995) | (0.220,0.122) | 0.157 | 0.029 | 0.085 | 0.033 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | (1.035,0.981) | (0.257,0.135) | 0.157 | 0.029 | 0.082 | 0.033 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | (0.491,1.001) | (0.196,0.108) | 0.107 | 0.028 | 0.039 | 0.028 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | (1.012,0.994) | (0.223,0.121) | 0.104 | 0.029 | 0.040 | 0.027 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | (0.511,0.996) | (0.229,0.130) | 0.154 | 0.026 | 0.087 | 0.027 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | (1.012,0.992) | (0.271,0.143) | 0.153 | 0.026 | 0.087 | 0.027 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | (0.511,0.992) | (0.211,0.117) | 0.103 | 0.026 | 0.041 | 0.024 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | (0.999,1.001) | (0.245,0.131) | 0.104 | 0.026 | 0.044 | 0.026 |

Table 3.2: Non-stationary Model Parameters Estimation Results.

| | | | | Parameter Estimation | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $Nt$ | $\beta$ | $\rho_1$ | $\rho_2$ | $\hat{\beta}$ | $SE_{\hat{\beta}}$ | $\hat{\rho_1}$ | $SE_{\rho_1}$ | $\hat{\rho_2}$ | $SE_{\hat{\rho_2}}$ |
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.20 | 0.05 | (0.494,1.002) | (0.075,0.138) | 0.202 | 0.041 | 0.054 | 0.047 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.20 | 0.05 | (1.000,1.004) | (0.072,0.133) | 0.204 | 0.043 | 0.056 | 0.052 |
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.15 | 0.10 | (0.502,0.992) | (0.071,0.134) | 0.160 | 0.038 | 0.076 | 0.046 |
| N2=N3=N4=N5=2 | (1.0,1.0) | 0.15 | 0.10 | (0.998,1.003) | (0.073,0.131) | 0.163 | 0.044 | 0.070 | 0.051 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | (0.500,0.997) | (0.069,0.131) | 0.151 | 0.030 | 0.032 | 0.030 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | (0.997,1.003) | (0.071,0.126) | 0.150 | 0.032 | 0.035 | 0.034 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | (0.498,1.000) | (0.072,0.134) | 0.105 | 0.026 | 0.043 | 0.028 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0,1.0) | 0.10 | 0.05 | (1.002,0.990) | (0.072,0.128) | 0.106 | 0.028 | 0.040 | 0.031 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | (0.500,0.995) | (0.074,0.134) | 0.150 | 0.029 | 0.033 | 0.030 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | (0.997,1.006) | (0.071,0.125) | 0.151 | 0.032 | 0.034 | 0.034 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | (0.495,1.005) | (0.073,0.141) | 0.107 | 0.027 | 0.043 | 0.029 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0,1.0) | 0.10 | 0.05 | (0.997,1.005) | (0.072,0.130) | 0.107 | 0.028 | 0.041 | 0.032 |
| N2=N3=N5=2,N4 =3 | (0.5, 1.0) | 0.20 | 0.05 | (0.499,1.001) | (0.073,0.135) | 0.198 | 0.038 | 0.049 | 0.042 |
| N2=N3=N5=2,N4 =3 | (1.0, 1.0) | 0.20 | 0.05 | (1.001,0.997) | (0.072,0.126) | 0.200 | 0.040 | 0.053 | 0.048 |
| N2=N3=N5=2,N4 =3 | (0.5, 1.0) | 0.15 | 0.10 | (0.500,0.993) | (0.073,0.135) | 0.160 | 0.034 | 0.078 | 0.040 |
| N2=N3=N5=2,N4 =3 | (1.0, 1.0) | 0.15 | 0.10 | (0.997,1.005) | (0.072,0.129) | 0.161 | 0.036 | 0.074 | 0.046 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | (0.500,0.997) | (0.071,0.136) | 0.162 | 0.038 | 0.074 | 0.045 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | (0.996,1.001) | (0.071,0.129) | 0.161 | 0.042 | 0.068 | 0.050 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | (0.504,0.994) | (0.071,0.134) | 0.113 | 0.036 | 0.041 | 0.035 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | (0.997,1.005) | (0.076,0.132) | 0.118 | 0.037 | 0.042 | 0.038 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | (0.501,0.991) | (0.077,0.144) | 0.159 | 0.034 | 0.082 | 0.041 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | (0.995,1.010) | (0.072,0.126) | 0.161 | 0.036 | 0.074 | 0.045 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | (0.496,1.005) | (0.073,0.129) | 0.110 | 0.033 | 0.041 | 0.032 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | (0.998,1.000) | (0.069,0.124) | 0.111 | 0.034 | 0.044 | 0.035 |

From Table 3.1 and Table 3.2, we can see that all the estimates for $\beta$ are close to its true value. In the nonstationary case, the estimates for $\rho_1$ and $\rho_2$ are less accurate than the stationary case in general. For some combinations of $\rho_1$ and $\rho_2$, such as $\rho_1 = 0.15$, $\rho_2 = 0.03$ with $n_2 = 2, n_3 = n_5 = 3, n_4 = 4$ in stationary case, the estimates of $\beta$ are slightly less accurate than the others. This is because $\rho_2$ is close to the lower bound. It is clear that for each stage of iterations, $\rho_1$ and $\rho_2$ have to satisfy the range restrictions, which in our case is $0 \leq \rho_2 \leq \rho_1$ and $0 \leq \rho_1 \leq min\left(\frac{\mu_{i2}}{n_t\mu_{i1}}, \frac{\mu_{it}-\rho_2 n_t \mu_{i,t-2}}{n_t \mu_{i,t-1}}, 1\right)$. Therefore, when $\rho_1$ or $\rho_2$ is close to their boundary, the estimates become less accurate. Those less accurate estimates for $\rho_1$ and $\rho_2$ will affect the estimates for $\beta$.

For the purpose of examining the forecast performance of the model (3.1), we use the parameter estimates obtained by using only the first four observations, $Y_{i1}, Y_{i2}, Y_{i3}$

and $Y_{i4}$ for $i = 1, 2, ..., 100$ in the forecasting function in Section 3.2.3 to compute a one-step ahead forecast of fifth observation. Next, we compute the sum of squares of the forecast error as well as the variance of the forecast error for these 100 communities. These calculation were repeated 1000 times as well for a total of 1000 estimates of the parameters. We denote the average sum of squares of the forecast errors and the average variance of the forecast error by ASS and AV respectively. The results are reported in Table 3.3 and Table 3.4.

Table 3.3: Stationary Model Forecasting Error.

| $Nt$ | $\beta$ | $\rho_1$ | $\rho_2$ | $ASS$ | $AV$ |
|---|---|---|---|---|---|
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.20 | 0.05 | 1.972 | 1.949 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.20 | 0.05 | 2.365 | 2.322 |
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.15 | 0.10 | 2.027 | 2.000 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.15 | 0.10 | 2.413 | 2.372 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | 1.996 | 1.978 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | 2.378 | 2.365 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | 2.077 | 2.055 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.10 | 0.05 | 2.484 | 2.446 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | 2.038 | 2.035 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | 2.486 | 2.427 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | 2.109 | 2.078 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.10 | 0.05 | 2.521 | 2.483 |
| N2 =N3 =N5 =2,N4 =3 | (0.5, 1.0) | 0.20 | 0.05 | 1.983 | 1.955 |
| N2 =N3 =N5 =2,N4 =3 | (1.0, 1.0) | 0.20 | 0.05 | 2.364 | 2.331 |
| N2 =N3 =N5 =2,N4 =3 | (0.5, 1.0) | 0.15 | 0.10 | 2.001 | 1.991 |
| N2 =N3 =N5 =2,N4 =3 | (1.0, 1.0) | 0.15 | 0.10 | 2.411 | 2.377 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | 1.970 | 1.920 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | 2.341 | 2.286 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | 2.104 | 2.048 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | 2.484 | 2.438 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | 1.959 | 1.928 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | 2.334 | 2.297 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | 2.091 | 2.047 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | 2.482 | 2.441 |

Table 3.4: Non-stationary Model Forecasting Error.

| $Nt$ | $\beta$ | $\rho_1$ | $\rho_2$ | $ASS$ | $AV$ |
|---|---|---|---|---|---|
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.20 | 0.05 | 2.990 | 2.887 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.20 | 0.05 | 4.950 | 4.781 |
| N2=N3=N4=N5=2 | (0.5, 1.0) | 0.15 | 0.10 | 3.039 | 2.967 |
| N2=N3=N4=N5=2 | (1.0, 1.0) | 0.15 | 0.10 | 5.049 | 4.900 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | 3.059 | 2.946 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | 5.020 | 4.859 |
| N2 =2,N3 =N5 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | 3.134 | 3.044 |
| N2 =2,N3 =N5 =3,N4 =4 | (1.0, 1.0) | 0.10 | 0.05 | 5.178 | 4.999 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.15 | 0.03 | 3.087 | 3.014 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.15 | 0.03 | 5.081 | 4.982 |
| N2 =N5 =2,N3 =3,N4 =4 | (0.5, 1.0) | 0.10 | 0.05 | 3.172 | 3.084 |
| N2 =N5 =2,N3 =3,N4 =4 | (1.0, 1.0) | 0.10 | 0.05 | 5.210 | 5.087 |
| N2 =N3 =N5 =2,N4 =3 | (0.5, 1.0) | 0.20 | 0.05 | 2.997 | 2.917 |
| N2 =N3 =N5 =2,N4 =3 | (1.0, 1.0) | 0.20 | 0.05 | 4.919 | 4.787 |
| N2 =N3 =N5 =2,N4 =3 | (0.5, 1.0) | 0.15 | 0.10 | 3.046 | 2.967 |
| N2 =N3 =N5 =2,N4 =3 | (1.0, 1.0) | 0.15 | 0.10 | 5.082 | 4.918 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | 3.036 | 2.875 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | 5.040 | 4.728 |
| N2 =N3 =N4 =2,N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | 3.136 | 3.024 |
| N2 =N3 =N4 =2,N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | 5.208 | 4.972 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.15 | 0.10 | 3.004 | 2.874 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.15 | 0.10 | 5.002 | 4.775 |
| N2 =N3 =2,N4 =N5 =3 | (0.5, 1.0) | 0.10 | 0.05 | 3.136 | 3.033 |
| N2 =N3 =2,N4 =N5 =3 | (1.0, 1.0) | 0.10 | 0.05 | 5.174 | 4.986 |

From Table 3.3 and Table 3.4, we see that the value of the average sum of squares and the average variance of the forecast errors are very close to each other for all different combinations of parameters. This indicates that the average sum of squares of the forecast errors can closely estimate the average variance of the forecast errors. Note that the values for average variance and sum of squares of forecast errors are generally smaller under stationary covariates. It could be also seen that the difference between average variance and sum of squares of forecast errors are also smaller than the nonstationary case. This means that we have better estimates for the parameters in the stationary case. This is because the covariates do not change with respect to time $t$ in stationary case. By similar discussion of results of simulated variance of forecasting errors in Section 2.2.4, for our particular setup, the average variance of

forecasting error for $\beta = (0.5, 1)$ is smaller than that of $\beta = (1, 1)$ as expected.

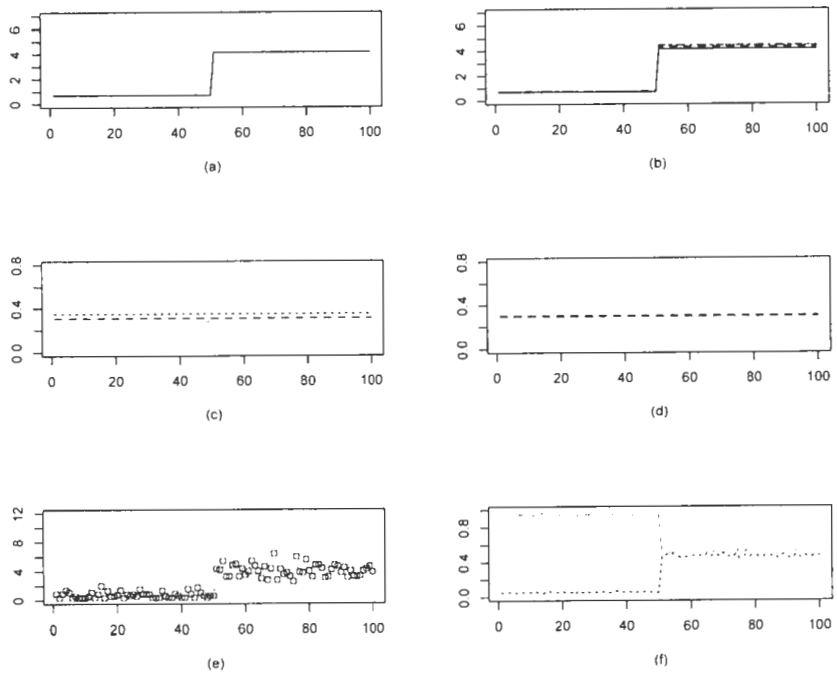


Figure 3.1: A plot of (a) values of stationary mean for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (b) values of stationary variance for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (c) values of stationary lag 1 correlation for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line); (d) values of stationary lag 2 correlation for $t = 1$ (solid line), $t = 2$ (dashed line); (e) Average forecast overlaid on average of longitudinal data; and (f) proportion of absolute values of forecast error that are 0 or 1 (solid line) and $> 1$ (dotted line); by communities obtained from 1000 simulations with $\rho_1 = 0.15$, $\rho_2$=0.10, $\beta$=(1,1), stationary covariates (2.23)-(2.24) and $n_1 = 1, n_2 =, ..., n_5 = 2$.

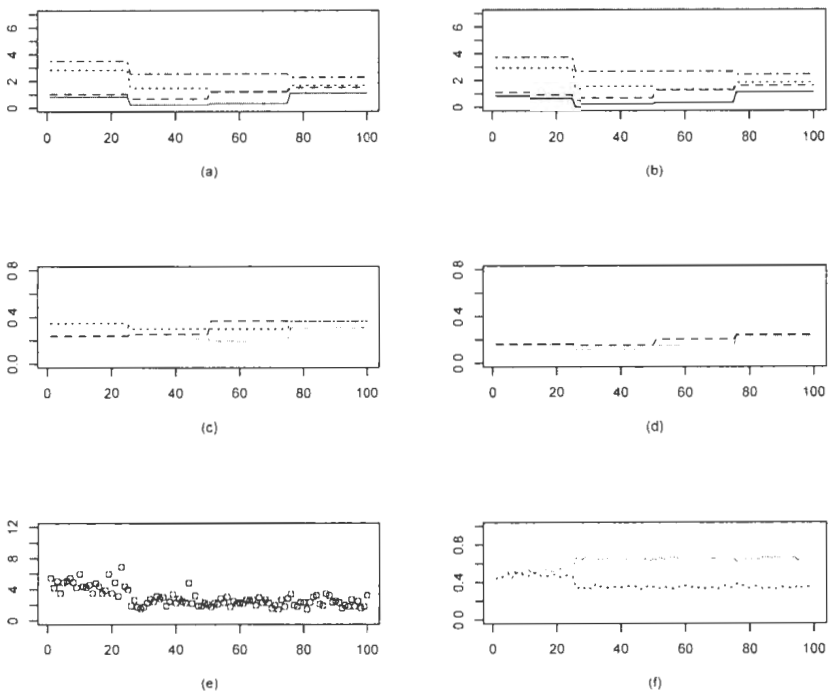


Figure 3.2: A plot of (a) values of nonstationary mean for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (b) values of nonstationary variance for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line), $t = 4$ (dotted dashed line); (c) values of nonstationary lag 1 correlation for $t = 1$ (solid line), $t = 2$ (dashed line), $t = 3$ (dotted line); (d) values of nonstationary lag 2 correlation for $t = 1$ (solid line), $t = 2$ (dashed line): (e) Average forecast overlaid on average of longitudinal data; and (f) proportion of absolute values of forecast error that are 0 or 1 (solid line) and $> 1$ (dotted line); by communities obtained from 1000 simulations with $\rho_1 = 0.15$, $\rho_2$=0.10, $\beta$=(1,1), nonstationary covariates (2.25)-(2.26) and $n_1 = 1, n_2 =, ..., n_5 = 2$.

In Figure 3.1(a),(b),(c) and (d), we let the maximum number of individuals that can be infected $n_t$, $t = 1, 2, ..., 5$ to be time dependent. The graph shows the stationary patterns in the mean $\mu_{it}$, variance $\sigma_{itt}$, lag 1 correlation $\rho_{i,t-1,t}$, and lag 2 correlation $\rho_{i,t-2,t}$. In Figure 3.1(e), we have overlaid a graph of the average of the forecast in 1000 simulations over a scatterplot of the average of the observations $y_{i5}$. The plot shows that the average forecast follows the general pattern of the infections at

the fifth time point. In order to assess the accuracy of out forecasts. We have also displayed a graph showing the average of the proportions of the forecast error $e_{it}$ with absolute deviations 0,1 and greater than 1. Figure 3.1(f) shows that the deviations of magnitude 0 and 1 appear to be over 90% for the first 50 communities and around 55% for the remaining 50 communities. We constructed similar plots in Figure 3.2 for a nonstationary case obtained from covariates generated by using (2.25) and (2.26).

# Chapter 4

# Concluding Remarks

In this thesis we have investigated the lag 2 binary and binomial sum model for modelling infectious diseases. We began with a simple case by assuming the infected individuals infecting none or only one individual at a time. Model (2.1) was proposed by assuming there is no latent community effect. Model (2.12) was proposed by assuming there is a latent community effect which will affect the mean of our observed number of infections. These binary sum models have limitations because it is more common that the infected individual can infect none, one and more than one individuals at a time in the real situations. Therefore, we proposed model (3.1) by using binomial sum operations instead of binary sum operations. We have assumed the immigration part $d_{it}$ for $t = 2, 3, .., T$ for the $i$th community follow a Poisson distribution. Parzen (1962, p. 118) showed that the Poisson process X(t) satisfies the following five axioms:

Axiom (0) $X(0) = 0$.

Axiom (1) $X(t)$ has independent increments; that is, for all $t_i$ such that $t_0 <$

$t_1 < \cdots < t_n$, the rv's $X(t_i), X(t_{i-1})$, $i = 1, 2, ..., n$, are independent.

Axiom (2) For any $t > 0$, $0 < Pr[X(t) > 0] < 1$.

Axiom (3) For any $t > 0$,

$$\lim_{h \to 0} \frac{Pr[X(t+h) - X(t) \geq 2]}{Pr[X(t+h) - X(t) = 1]} = 0$$

Axiom (4) $X(t)$ has stationary increments; that is, for points $t_i \geq t_j \geq 0$ (and $h > 0$), the random variables $X(t_i) - X(t_j)$ and $X(t_{i+h}) - X(t_{j+h})$ are equidistributed.

In our modelling approach, the number of immigrations are discrete counts which can occur at any point along a continuum. There is no immigration at the initial time point. At any particular point, the probability of the immigration is small. The average number of immigrations is constant over a unit of measure and $d_{i2}$,....$d_{it}$,...,$d_{iT}$ are independent. Therefore, we assume that $d_{it}$, $t = 2, 3, ..., T$, follow Poisson distributions. The Method of Moments and the GQL method was shown to perform well in estimating the parameters. These estimated models can be used to obtain a reasonable forecasts of future spread of disease using the forecasting function when the latent community effect is not present. When the latent community effect is present, the GQL approach still works well for estimating regression parameters and parameter in the latent community effect. However, the moment estimates for the correlation parameters become less accurate.

We remark that by the nature of model (2.1), (2.12) and (3.1), new infections at time point $t$ are completely determined by the number of infections at time point $t-1$

and $t-2$. There may be situations where an individual who was infected at time point $t-k$ continue to infect others at future time points until the individual is discovered or recovered. Furthermore, model (2.12) only considers the latent community effect present in a binary sum infectious disease model. For a binomial sum infectious model, the latent community effect may also present. These situations are subjects for future consideration.

# Bibliography

Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* 82, 407-410.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). New Jersey: Wiley.

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

McKenzie, Ed (1988). Some ARMA models for dependent sequences of Poisson counts. *Probability* 20, no.4, 822-835.

Oyet, A. J. & Sutradhar, B. C. (2011). *Longitudinal Modelling of Infectious Disease.* Unpublished manuscript.

Parzen, E. (1962). *Stochastic Processes with Applications to Science and Engineering.* San Francisco: Holden-Day.

Staudenmayer, J. and Buonaccorsi, J. (2005). Measurement Error in Linear Autoregressive Models. *Journal of the American Statistical Association, Theory and Methods* 100, 841-852.

Sutradhar, B. C. (2011). *Dynamic Mixed Models for Familial Longitudinal Data.* New York: Springer.

Sutradhar, B. C. (2010). Inferences in generalized linear longitudinal mixed models. *The Canadian Journal of Statistisitcs* 38, 174-196.

Sutradhar, B.C., Oyet, A. J. & Gadag, V.G. (2010). On quasi-likelihood estimation for branching processes with immigration. *The Canadian Journal of Statistics* 38, 290-313.

Sutradhar, B. C. & Bari W. (2007). On Generalized quasilikelihood inference in longitudinal mixed model for count data. *Sankhyā: The Indian Journal of Statististics* 69, 671-699.

Sutradhar, B.C. (2004). On exact quasilikelihood inference in generalized linear mixed models. *Sankhyā: The Indian Journal of Statististics* 66, 261-289.

Sutradhar, B.C. (2003). An Overview on Regression Models for Discrete Longitudinal Responses. *Statistical Science* 18, no. 3, 377-393.

Sutradhar, B.C. & Jowaheer, V. (2003). On familial longitudianl Poisson mixed models with gamma random effect. *Journal of Multivariate Analysis* 87, no.2, 398-412.

Sutradhar, B.C. & Das, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika* 86, no.2, 459-465.

Wedderburn, R. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 61, 439-447.