Old Dominion University

ODU Digital Commons

Mathematics & Statistics Theses & Dissertations

Mathematics & Statistics

Winter 2010

Analysis of Models for Longitudinal and Clustered Binary Data

Weiming Yang
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/mathstat_etds

 Part of the Analysis Commons, Applied Statistics Commons, and the Longitudinal Data Analysis and Time Series Commons

Recommended Citation

Yang, Weiming. "Analysis of Models for Longitudinal and Clustered Binary Data" (2010). Doctor of Philosophy (PhD), Dissertation, Mathematics & Statistics, Old Dominion University, DOI: 10.25777/69t2-c482
https://digitalcommons.odu.edu/mathstat_etds/74

This Dissertation is brought to you for free and open access by the Mathematics & Statistics at ODU Digital Commons. It has been accepted for inclusion in Mathematics & Statistics Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

ANALYSIS OF MODELS FOR LONGITUDINAL AND CLUSTERED BINARY DATA

by

Weiming Yang
M.S. Aug 2005, University of Toledo
M.S. Aug 2009, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

DOCTOR OF PHILOSOPHY

MATHEMATICS AND STATISTICS

OLD DOMINION UNIVERSITY
December 2010

Approved by:

N. Rao Chaganty (Director)

Davanand Naik (Member)

Michael Doviak (Member)

Larry Filer (Member)

ABSTRACT

ANALYSIS OF MODELS FOR LONGITUDINAL AND CLUSTERED BINARY DATA

Weiming Yang
Old Dominion University, 2010
Director: Dr. N. Rao Chaganty

This dissertation deals with modeling and statistical analysis of longitudinal and clustered binary data. Such data consists of observations on a dichotomous response variable generated from multiple time or cluster points, that exhibit either decaying correlation or equi-correlated dependence. The current literature addresses modeling the dependence using an appropriate correlation structure, but ignores the feasible bounds on the correlation parameter imposed by the marginal means.

The first part of this dissertation deals with two multivariate probability models, the first order Markov chain model and the multivariate probit model, that adhere to the feasible bounds on the correlation. For both the models we obtain maximum likelihood estimates for the regression and correlation parameters, and study both asymptotic and small-sample properties of the estimates. Through simulations we compare the efficiency of the two methods and demonstrate that neither is uniformly superior over the other.

The second part of this dissertation deals with marginal models, an alternative to multivariate probability models. We discuss the generalized estimating equations and the quadratic inference function methods for estimating the regression parameter in marginal models. Relative efficiency calculations show these methods when compared to the likelihood estimates could result in significant loss in efficiency for highly correlated data. We also propose a modified quadratic inference function method and demonstrate through efficiency calculations this is an improvement of the original quadratic inference function approach. The final part of this dissertation deals with methods for constructing higher order Markov chain models using copulas.

ACKNOWLEDGMENTS

I would like to thank first and foremost my advisor, Professor N. Rao Chaganty. This would not have been possible without his support. Not only were his encouragement and assistance were essential to both the completion of this work and to my academic career, his persistence for scientific precision influenced my own philosophy on academic investigations. I am forever grateful and would like to express my deepest gratitude to him.

I would like to thank Professors Dayanand Naik, Michael Doviak, and Larry Filer for serving on my dissertation committee. Additionally, I would like to express my gratitude to other statistics faculty, friends, and my fellow students in the Department of Mathematics and Statistics, and to the Chairman Professor Dorrepaal, for providing a caring and pleasant learning environment.

I also wish to acknowledge the staff members Barbara Jeffrey and Sheila Hegwood, for their help during my graduate studies. Finally, I have been very fortunate that the Department of Mathematics and Statistics provided financial support for my graduate studies, and would like to say thank you to the department for giving me this wonderful opportunity.

I dedicate this thesis to my parents and my wife Shu Liao.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

I.1 Repeated measurements

Many practical research areas utilize regression analysis of correlated longitudinal data. In particular, longitudinal data is frequently encountered in many subject-matter areas such as biology, medicine and public health, and social sciences. Longitudinal data is essentially data observed sequentially over time. It may be collected either from a designed experiment or an observational study, where the outcome variables are related to a sequence of events or responses recorded at certain time points during a study period. In essence, longitudinal data may be regarded as a collection of many short time series, one for each subject.

Some research areas are focused on correlated clustered data. Clustered data refers to a set of measurements collected from subjects that are organized in groups, where a group of related subjects constitutes a cluster. An example of cluster or group is a genetically related members from a family pedigree. Obviously, situations where clustered data arise can be independent of time. Longitudinal data may be thought of as a special kind of clustered data by treating a subject as a cluster, so each subject's time series forms a set of correlated observations. This perspective is mainly for technical convenience, because similar tools can be applied to analyze longitudinal data or clustered data with different modeling of dependence structures. In longitudinal data analysis, serial correlation is common, whereas in clustered data analysis exchangeable pairwise within-cluster correlation model is popular.

Although there are many approaches for the analysis of continuous longitudinal data, the development of the methods for categorical longitudinal data has received less attention, and the methodology is not nearly as well-developed as for continuous data. In this dissertation, we will study likelihood and some other alternative methods for analyzing longitudinal binary data.

Table 1.1 shows the general layout for longitudinal data that will be used in this

This dissertation follows the style of *Journal of the American Statistical Association.*

dissertation. Let y_{ij} be the response at time point j on subject i, where $j = 1, \cdots, t_i$ and $i = 1, \cdots, n$. We assume that the n subjects are independent. The data are said to be balanced if $t_i = t$ for $1 \leq i \leq n$. Let k be the number of covariates, and $x_{ij} = (x_{ij1}, \cdots, x_{ijk})'$ be the corresponding covariate vector associated with y_{ij}. In general, x_{ij}'s are time-dependent, the values of x_{ij} vary at different time points. Some real life binary longitudinal data examples are given in next section.

Table 1.1: General layout for longitudinal data

Subject	Time	Response	Covariates		
1	1	y_{11}	x_{111}	$\cdots$	x_{11k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	y_{1j}	x_{1j1}	$\cdots$	x_{1jk}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_1	y_{1t_1}	x_{1t_11}	$\cdots$	x_{1t_1k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	1	y_{i1}	x_{i11}	$\cdots$	x_{i1k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	y_{ij}	x_{ij1}	$\cdots$	x_{ijk}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_i	y_{it_i}	x_{it_i1}	$\cdots$	x_{it_ik}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	1	y_{n1}	x_{n11}	$\cdots$	x_{n1k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	j	y_{nj}	x_{nj1}	$\cdots$	x_{njk}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	t_n	y_{nt_n}	x_{nt_n1}	$\cdots$	x_{nt_nk}

I.2 Data examples

To motivate both methodological and theoretical developments in the subsequent chapters, a few real world data sets with binary outcomes will be used for illustration in this dissertation.

I.2.1 Six city respiratory infection study

Table 1.2: Six-city data

No maternal smoking					Maternal smoking				
Age				Count	Age				Count
7	8	9	10		7	8	9	10	
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	11	1	1	1	1	7

Table 1.2 shows a longitudinal data set from a Harvard University technical report by N. M. Laird, G. J. Beck and J. H. Ware. These data are part of a study of the respiratory health effects of indoor and outdoor air pollution in six U.S. cities.

The main research interest is the effect of maternal smoking on children's respiratory illness. The serial response variable for children from 7 to 10 is recorded as a binary outcome denoting the absence (0) and presence (1) of respiratory illness. The maternal smoking is also a dichotomous variable with 0 as an indicator of smoking and 1 indicates non-smoker. This is a time-independent covariate, i.e. it does not change during the time of the study for each subject. The data includes only those children who had all four responses at ages 7, 8, 9 and 10, and therefore balanced.

I.2.2 Indonesian children's health study

Table 1.3: Indonesian children's health data

		Age						
Xerophthalmia	RI	1	2	3	4	5	6	7
No	No	90	236	330	176	143	65	5
	Yes	8	36	39	9	7	1	0
Yes	No	0	2	18	15	8	4	1
	Yes	0	0	7	0	0	0	0

Sommer et al. (1984) reported a study in West Java, Indonesia to determine the causes and effect of vitamin A deficiency in preschool children. More than 3000 children were medically examined quarterly for up to six visits to assess whether they suffered from respiratory or diarrheal infection (RI) and xerophthalmia, an ocular manifestation of vitamin A deficiency. Weight and height were also measured. The data on 275 children are summarized in Table 1.3. This longitudinal data is recorded at equally spaced time points where the binary response variable y_{ij} is equal to 1 if the child i had RI at time point j and 0 otherwise. The main covariate of interest is Xerophthalmia which is represented as a binary variable with 1 for presence and 0 for the absence of the Xerophthalmia symptom.

The main objective of this study was to assess the increase in risk for RI among kids who were vitamin A deficient, which was measured indirectly via Xerophthalmia. It was also of interest to evaluate the degree of heterogeneity in the risk of disease

among the kids.

I.2.3 Hamilton's depression study

The data for this example were taken from a double-blind, European, multi-center, placebo and active treatment controlled, randomized, 5 arm parallel group, 7 week dose-finding study to evaluate safety and efficacy for three fixed doses of a new drug in patients with major depressive disorder. The dependent variable in this study is a binary function of a patient's average score on the Hamilton's Depression Scale (Ham-D), taking the value 1 if the Ham-D value at time $i = 1, \cdots, 8$ is less than or equal to 80% of the baseline value, and 0 otherwise. The baseline Ham-D values for all subjects are greater than 18, implying that all subjects are initially diagnosed as severely depressed with a dichotomous baseline value of 0. The assessment was made each week on the patients, starting from the beginning of the first week (baseline) and continuing for the next seven weeks, for a total of 8 measurements. The primary objective is the change in Ham-D rating from baseline to the final visit (the subject might left the study early). The main covariates we will use in our analysis are the Treatment (active or placebo) and Time (in number of weeks from the baseline measurement). Table 1.4 shows the observation of all 8 visits of the first patient.

Table 1.4: Hamilton's depression data

ID	Time	Ham-D	COUNTRY	Age	Gender	Treatment	Baseline	Y
1	1	24	BULGARIA	42	M	1	22	0
1	2	22	BULGARIA	42	M	1	22	0
1	3	21	BULGARIA	42	M	1	22	0
1	4	14	BULGARIA	42	M	1	22	1
1	5	12	BULGARIA	42	M	1	22	1
1	6	10	BULGARIA	42	M	1	22	1
1	7	7	BULGARIA	42	M	1	22	1
1	8	8	BULGARIA	42	M	1	22	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

I.3 Overview of methods for the analysis of repeated measurements

Statistical researchers have developed several related types of extensions of generalized linear models and quasi-likelihood methods for the analysis of repeated measurements. These methods are useful for both discrete and continuous response variables, including normal, Poisson, binary, and gamma responses. Three general types of extensions of generalized linear models methodology to the analysis of repeated measurements are

- Marginal models;
- Random-effects models;
- Transition models.

We will discuss next these models briefly.

I.3.1 Marginal models

In marginal models, the marginal expectation $\mu_{ij} = \mathrm{E}(y_{ij})$ is modeled as a function of explanatory variables. The marginal expectation is the average response over the subpopulation that shares a common value of the covariate vector. Associations among repeated observations are modeled separately from the marginal mean and variance of the response vector.

The assumptions can be outlined as follows:

1. The marginal expectation μ_{ij} is related to the covariates through a known link function h:
$$\mu_{ij} = h(x'_{ij}\beta),$$
where β is a $k \times 1$ vector of regression parameters.
2. The marginal variance of y_{ij} is related to the marginal expectation μ_{ij} via
$$\mathrm{Var}(y_{ij}) = \phi \mathrm{V}(\mu_{ij}),$$
where V is a known variance function and ϕ is a dispersion scale parameter.

3. The covariance between y_{ij} and y_{ik} is a known function of μ_{ij} and μ_{ik} , and a vector α of unknown parameters.

I.3.2 Random-effects models

In random-effects models, heterogeneity between individuals arising from unmeasured variables is accounted for by including subject-specific random effects in the model. These random effects are assumed to account for all of the within-subject correlation present in the data. Conditional on the values of the random effects, the responses are assumed to be independent.

The assumptions in the random-effects models are as follows:

1. Given a vector b_i of subject-specific effects for the ith subject, the conditional mean of y_{ij} satisfies the model

$$h(\mathrm{E}(y_{ij}|b_i)) = x'_{ij}\,\beta + z'_{ij}\,b_i,$$

   where h is a known link function and z_{ij} is a vector of covariates for subject i at time j.

2. Outcomes $y_{i1}, \cdots, y_{it_i}$ are independent given b_i for each $i = 1, \cdots, n$.

3. Random effects $b_1, \cdots, b_n$ are independent and identically distributed.

I.3.3 Transition models

Transition models account for heterogeneity by tracing the development of a dependent variable over time and they represent the distribution of its current value as a function of its history. In transition models, for the analysis of repeated measurements, the observations $y_{i1}, \cdots, y_{it_i}$ from subject i are correlated because y_{ij} is explicitly influenced by the past values $y_{i1}, \cdots, y_{ij-1}$. Suppose h_j is the history of the subject up to time point j. With this information, we can calculate the probability of observing $y_i = (y_{i1}, \cdots, y_{it})$ as

$$\pi(y_i) = p(y_{i1}) \prod_{j=2}^{t} p(y_{ij}|h_j),$$

where $p(y_{i1})$ is the marginal probability of the first observation and $p(y_{ij}|h_j)$ is the conditional probability of y_{ij} given its history. Specifically, if the transition form y_{ij-1} to y_{ij} only depends on the values of y_{ij-1}, this transition model is known as the first order Markov chain model.

I.4 Overview of thesis

This dissertation is organized as follows. In Chapter II, we introduce a conditional linear family of multivariate binary distributions due to Qaqish (2003). We show that this family results in a first order Markov chain model when the correlation structure is the first-order autoregressive model. We derive the maximum likelihood estimates and the Fisher information matrix for the regression and correlation parameters in this model. We also develop test of hypothesis. Next, we study the multivariate probit probability model and show that these two models generate different probability mass functions. We make large and small sample efficiency comparisons between the two likelihood models and show that no one model is uniformly superior over the other. Real life data examples are presented to contrast parameter estimates for both the models.

In Chapter III, we discuss marginal models. As the name indicates these models specify only the marginal distributions and correlation structure but not the joint distribution. A popular method of estimation for the marginal regression parameter is the generalized estimating equations. We briefly describe the GEE and outline some theoretical drawbacks with the method. An alternative to the GEE is the quadratic inference function (QIF) approach due to Qu et al. (2000). This method uses basis matrices in lieu of the working correlation parameter and bypasses estimating the correlation parameter. We give explicit expression for the asymptotic variance of the QIF regression parameter estimator. Through large and small sample simulations we show that the QIF method is inefficient for estimating the regression parameter for highly correlated data when compared with the maximum likelihood estimator from the Markov chain model. We also propose a modified QIF (mQIF) method using a correlation estimator that pools information from all the subjects. We show that mQIF is an improvement over the QIF method. We also study performance of the MC and mQIF when the correlation model is misspecified.

In Chapter IV, we give a short introduction and discuss possible methods of constructing higher order Markov chain models using copulas. The dissertation ends with a discussion in Chapter V and an appendix containing the SAS macro TMMLE that we developed for fitting the first order Markov chain model for real data.

CHAPTER II

TRANSISTION MODELS FOR BINARY DATA

II.1 Introduction

Binary longitudinal data are often collected in clinical trials and genetic studies where interest is on assessing the effect of a treatment over time. A canonical problem is to determine a regression relationship between the measured responses and a set of time dependent/independent covariates. The longitudinal binary data can be viewed as a short discrete time series. Unlike Gaussian time series, modeling discrete variate time series is very challenging and difficult. And the statistical methods are not well developed for discrete variate time series. The fact that variate values are integer renders most traditional representations of dependence either impossible or impractical. There have been a number of efforts to develop a suitable class of time series models for binary data. In this chapter we first briefly review the literature and then introduce a fully specified transistion model for the analysis of longitudinal binary data.

II.2 Survey of transistion models for binary data

Longitudinal binary data consists of a sequence of binary variables, which can be thought as a sequence of states of a two-state stochastic process. A natural way to model the joint distribution of a two-state stochastic process is by a probability model which describes the transition from one state to another. Several researchers have developed transition models for the analysis of longitudinal binary data. Here we will briefly survey some of these models. Muenz and Rubinstein (1985) introduced a two-state Markov chain for a discrete-time binary sequence. Their motivation was a study of the impact of mastectomy which measured the binary response of "distress" or "no distress" for five groups of women who had surgery for different breast disease. The transition matrix of the chain is given by

$$M = \begin{pmatrix} p_{00} & 1-p_{00} \\ p_{10} & 1-p_{10} \end{pmatrix},$$

where p_{00} is the $0 \to 0$ transition probability and p_{10} is the $1 \to 0$ transition probability. It was assumed that the process is stationary and thus the transition matrix M remains the same at different time points. Muenz and Rubinstein (1985) modeled the transition probabilities p_{00} and p_{10} by logistic regressions with a covariate vector $\mathbf{x}$ and two different regression parameters β and γ as follows

$$p_{00}(\beta) = \frac{\exp(\mathbf{x}'\beta)}{1+\exp(\mathbf{x}'\beta)} \quad \text{and} \quad p_{10}(\gamma) = \frac{\exp(\mathbf{x}'\gamma)}{1+\exp(\mathbf{x}'\gamma)}. \tag{2.2.1}$$

The likelihood function corresponding to model (2.2.1) is then given by

$$\begin{aligned} L(\beta,\,\gamma) &= \prod_{i=1}^{n} p_{00}(\beta)^{n_{i00}}(1-p_{00}(\beta))^{n_{i01}} p_{10}(\gamma)^{n_{i10}}(1-p_{10}(\gamma))^{n_{i11}} \\ &= \prod_{i=1}^{n} p_{00}(\beta)^{n_{i00}}(1-p_{00}(\beta))^{n_{i01}} \prod_{i=1}^{n} p_{10}(\gamma)^{n_{i10}}(1-p_{10}(\gamma))^{n_{i11}}, \end{aligned} \tag{2.2.2}$$

where n_{i00}, n_{i01}, n_{i10} and n_{i11} are the number of $0 \to 0$, $0 \to 1$, $1 \to 0$ and $1 \to 1$ transitions observed on the ith subject. Clearly the likelihood function (2.2.2) depends only on $p_{00}(\beta)$ and $p_{10}(\gamma)$, so we can get the maximum likelihood estimates of β and γ by maximizing the following two log-likelihood functions separately

$$\begin{aligned} \ell_0(\beta) &= \sum_{i=1}^{n} \{n_{i00}\mathbf{x}_i'\beta - (n_{i00}+n_{i01})\log[1+\exp(\mathbf{x}_i'\beta)]\}, \\ \ell_1(\gamma) &= \sum_{i=1}^{n} \{n_{i10}\mathbf{x}_i'\gamma - (n_{i10}+n_{i11})\log[1+\exp(\mathbf{x}_i'\gamma)]\}. \end{aligned}$$

Albert and Waclawiw (1998) proposed a quasi-likelihood transitional model. They assume the transitional probabilities are random satisfying some moment conditions. Their model can be thought of as a random-effects model. In their model the transition matrix for the ith subject takes the form

$$\begin{pmatrix} 1-P_{i01} & P_{i01} \\ P_{i10} & 1-P_{i01} \end{pmatrix}, \tag{2.2.3}$$

where P_{i01} and P_{i10} are random quantities satisfying the moment conditions

$$E[P_{i01}] = \mu_{01}, \quad E[P_{i10}] = \mu_{10}$$

and

$$var(P_{i01}) = \sigma_{01}^2, \quad var(P_{i10}) = \sigma_{10}, \text{ and } \text{corr}(P_{i01}, P_{i10}) = \rho.$$

If there are observed covariates, the means μ_{01} and μ_{10} could be modeled as a function of the covariates and the regression parameter. Let $n_i = (n_{i00}, n_{i01}, n_{i10})'$ be a vector of the number of $0-0$, $0-1$ and $1-0$ transitions over t time points for the ith subject. Note that the $1-1$ transitions are determined by the sum of the other transitions. Let s_i be the vector which consists of the squares and cross-products of n_i along with n_{i11}, that is,

$$s_i = (n_{i00}^2, n_{i00}n_{i01}, n_{i00}n_{i10}, n_{i01}^2, n_{i01}n_{i10}, n_{i10}^2, n_{i11}n_{i00}, n_{i11}n_{i01}, n_{i11}n_{i10}, n_{i11}^2)'.$$

Albert and Waclawiw (1998) proposed an estimating equation approach for estimating the means μ_{01}, μ_{10}, variances σ_{01}, σ_{10} and the correlation ρ based on s_i's for $1 \leq i \leq n$. They also provide methods of estimation for the regression parameter in the situations where the means are functions of the covariates and the regression parameter.

Azzalini (1994) introduced another model based on first order Markov chain to analyze serial correlated binary observations. Instead of using correlation to measure the dependence, he used odds ratio to measure the dependence between two consecutive binary variables. The odds ratio is defined as

$$\psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}, \tag{2.2.4}$$

where $p_j = P(Y_t = 1|Y_{t-1} = j)$ for $j = 0, 1$. A technical reason in favor of using this odds ratio as a measure of dependence as opposed to the correlation is given by Fitzmaurice and Laird (1993). They observed when the association between observations is modeled using ψ, the estimates of the mean are relatively insensitive to changes of the association parameter. Moreover, the range of feasible values for ψ is independent of the value of $\mu = E(Y_t)$. In the stationary case, given the mean μ and the odds ratio ψ, we can obtain p_0 and p_1 by solving the equations (2.2.4) and

$$\mu = \mu p_1 + (1-\mu)p_0. \tag{2.2.5}$$

In the non-stationary case $\mu_t(\beta) = E(Y_t)$ varies at different time points and it is related to the covariates x_t and the regression parameter β via some link function.

Then equation (2.2.5) is replaced by

$$\mu_t(\beta) = \mu_{t-1}(\beta)\, p_1 + (1 - \mu_{t-1}(\beta))\, p_0, \tag{2.2.6}$$

where p_0 and p_1 now vary with t and β as well. Solving equations (2.2.4) and (2.2.6), we obtain

$$p_j = \begin{cases} \mu_t & \text{for } \psi = 1 \\ \frac{\delta - 1 + (\psi-1)(\mu_t - \mu_{t-1})}{2(\psi-1)(1-\mu_{t-1})} + j\frac{1-\delta+(\psi-1)(\mu_t-\mu_{t-1}-2\mu_t\mu_{t-1})}{2(\psi-1)\mu_t(1-\mu_{t-1})} & \text{for } \psi \neq 1 \end{cases} \tag{2.2.7}$$

where $\delta^2 = 1 + (\psi - 1)[(\mu_t - \mu_{t-1})^2\psi - (\mu_t + \mu_{t-1})^2 + 2(\mu_t + \mu_{t-1})]$. In the above we suppressed the dependence of p_j on t, β and ψ.

Given a sequence of observations $y_1, y_2, \cdots, y_t$, the log-likelihood function for β and ψ is

$$\ell(\beta, \psi) = \sum_{i=1}^{t} \ell_i(\beta, \psi) = \sum_{i=1}^{t} [y_i \operatorname{logit}(p_{y_{i-1}}) + \log(1 - p_{y_{i-1}})]. \tag{2.2.8}$$

This can be summed over all the subjects and maximized to get the maximum likelihood estimates of β and ψ. Even though this approach has desirable features but there are some disadvantages. If the marginal means depend on time-varying covariates, the assumption of odds ratio being fixed over all possible marginal probabilities is unlikely to be so.

In the next section, we discuss a Markov chain model which generates binary random vectors with correlation structure that is the first order autoregressive.

II.3 First order autoregressive structure

A reasonable correlation model for a sequence of serially correlated binary measurements $y_1, y_2, \cdots, y_t$ is the first order autoregressive (AR(1)) structure with parameter ρ which is given by

$$R(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho^t \\ \rho & 1 & \cdots & \rho^{t-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^t & \rho^{t-1} & \cdots & 1 \end{bmatrix}. \tag{2.3.1}$$

In the next two sections we derive the feasible range for the correlation parameter ρ and a method of constructing a joint probability distribution with correlation structure (2.3.1).

II.3.1 Conditional linear family

Recently, Qaqish (2003) introduced a conditional linear family of binary distributions. The idea behind the construction of this family is that the expected value of the current observation is taken as a linear function of the past observations. That is,

$$E(Y_i|Y_j = y_j,\ 1 \le j \le i-1) = p_i + \sum_{j=1}^{i-1} b_{ij}(Y_j - p_j). \tag{2.3.2}$$

To find the coefficients b_{ij} we multiply both sides of the above equation by $(Y_j - p_j)$, $j = 1, \cdots, i-1$. Then taking expectations we get

$$\begin{aligned}
Cov(Y_i, Y_1) &= b_{i1}\, Var(Y_1) + b_{i2}\, Cov(Y_1, Y_2) + \cdots + b_{i(i-1)}\, Cov(Y_1, Y_{i-1}) \\
Cov(Y_i, Y_2) &= b_{i1}\, Cov(Y_2, Y_1) + b_{i2}\, Var(Y_2) + \cdots + b_{i(i-1)}\, Cov(Y_2, Y_{i-1}) \\
&\vdots \\
Cov(Y_i, Y_{i-1}) &= b_{i1}\, Cov(Y_{i-1}, Y_1) + b_{i2}\, Cov(Y_{i-1}, Y_2) + \cdots + b_{i(i-1)}\, Var(Y_{i-1}).
\end{aligned}$$

The above identities can be written in a matrix form as

$$\begin{bmatrix} Cov(Y_i, Y_1) \\ Cov(Y_i, Y_2) \\ \vdots \\ Cov(Y_i, Y_{i-1}) \end{bmatrix} = \Sigma_{i-1}\, \mathbf{b},$$

where $\mathbf{b} = (b_{i1}, b_{i2}, \cdots, b_{i(i-1)})'$ and

$$\Sigma_{i-1} = \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_{i-1}) \\ Cov(Y_2, Y_1) & Var(Y_2) & \cdots & Cov(Y_2, Y_{i-1}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_{i-1}, Y_1) & Cov(Y_{i-1}, Y_2) & \cdots & Var(Y_{i-1}) \end{bmatrix}.$$

Thus

$$\mathbf{b} = \Sigma_{i-1}^{-1} \begin{bmatrix} Cov(Y_i, Y_1) \\ Cov(Y_i, Y_2) \\ \vdots \\ Cov(Y_i, Y_{i-1}) \end{bmatrix}.$$

In the case the correlation matrix of $(y_1, \ldots, y_{i-1})$ is $R(\rho)$, we have

$$\Sigma_{i-1} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{i-1} \end{bmatrix} R(\rho) \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{i-1} \end{bmatrix}.$$

Therefore

$$\begin{aligned} \mathbf{b} &= \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{i-1}} \end{bmatrix} R^{-1}(\rho) \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_{i-1}} \end{bmatrix} \begin{bmatrix} \sigma_1\sigma_i\rho^{i-1} \\ \sigma_2\sigma_i\rho^{i-2} \\ \vdots \\ \sigma_{i-1}\sigma_i\rho \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sigma_i}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_i}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sigma_i}{\sigma_{i-1}} \end{bmatrix} R^{-1}(\rho) \begin{bmatrix} \rho^{i-1} \\ \rho^{i-2} \\ \vdots \\ \rho \end{bmatrix}. \end{aligned}$$

For the AR(1) correlation matrix $R(\rho)$, the inverse is given by

$$R^{-1}(\rho) = \frac{1}{1-\rho^2}[(1+\rho^2)M_1 - \rho M_2 - \rho^2 M_3],$$

where M_1 is the identity matrix, M_2 is a tri-diagonal matrix with ones right above and below the diagonal, zero everywhere else. The matrix M_2 is a matrix of zeros except the first and last diagonal elements which are equal to 1. Note that

$$R^{-1}(\rho) \begin{bmatrix} \rho^{i-1} \\ \rho^{i-2} \\ \vdots \\ \rho^2 \\ \rho \end{bmatrix} = \frac{1}{1-\rho^2} \left((1+\rho^2) \begin{bmatrix} \rho^{i-1} \\ \rho^{i-2} \\ \vdots \\ \rho^2 \\ \rho \end{bmatrix} - \begin{bmatrix} \rho^{i-1} \\ \rho^i + \rho^{i-2} \\ \vdots \\ \rho^4 \\ \rho^3 \end{bmatrix} - \begin{bmatrix} \rho^{i+1} \\ 0 \\ \vdots \\ 0 \\ \rho^3 \end{bmatrix} \right)$$

$$= \frac{1}{1-\rho^2} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \rho - \rho^3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \rho \end{bmatrix}.$$

Therefore

$$\mathbf{b} = \begin{bmatrix} \frac{\sigma_i}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_i}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sigma_i}{\sigma_{i-1}} \end{bmatrix} R^{-1}(\rho) \begin{bmatrix} \rho^{i-1} \\ \rho^{i-2} \\ \vdots \\ \rho \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \rho\frac{\sigma_i}{\sigma_{i-1}} \end{bmatrix}.$$

Thus for the AR(1) correlation structure the conditional linear family of binary distributions satisfy the relation

$$\begin{aligned} E(Y_i|Y_j = y_j,\ 1 \le j \le i-1) &= \mathrm{P}(Y_i = 1|Y_j = y_j) \\ &= p_i + \rho\frac{\sigma_i}{\sigma_{i-1}}(y_{i-1} - p_{i-1}). \end{aligned} \tag{2.3.3}$$

This shows that if the correlation structure is AR(1) then Y_i depends only y_{i-1} and is independent of the past observations. Thus the resulting process is a first order Markov chain.

II.3.2 Probability mass function of the conditional linear family

In the previous section we have seen that when the correlation structure is AR(1), the resulting conditional linear family is a first order Markov chain. For this chain, the transition probabilities of moving from y_{i-1} to y_i can be written explicitly as follows.

$$\begin{aligned} P_{i01} &= p_i - \rho\frac{\sigma_i}{\sigma_{i-1}}(0 - p_{i-1}) = p_i - \rho\frac{\sigma_{i-1}\sigma_i}{q_{i-1}} \\ P_{i11} &= p_i - \rho\frac{\sigma_i}{\sigma_{i-1}}(1 - p_{i-1}) = p_i + \rho\frac{\sigma_{i-1}\sigma_i}{p_{i-1}} \\ P_{i00} &= 1 - P_{i01} = q_i + \rho\frac{\sigma_{i-1}\sigma_i}{q_{i-1}} \\ P_{i10} &= 1 - P_{i11} = q_i - \rho\frac{\sigma_{i-1}\sigma_i}{p_{i-1}}. \end{aligned}$$

We require that these four probabilities lie between 0 and 1 for all i. This leads to the bounds on the correlation parameter ρ as

$$\max_{2\le i\le t} L(p_{i-1}, p_i) \le \rho \le \min_{2\le i\le t} U(p_{i-1}, p_i), \tag{2.3.4}$$

where

$$L(a,b) = \max\left\{-\sqrt{\frac{ab}{(1-a)(1-b)}}, -\sqrt{\frac{(1-a)(1-b)}{ab}}\right\},$$
$$U(a,b) = \min\left\{\sqrt{\frac{a(1-b)}{(1-a)b}}, \sqrt{\frac{(1-a)b}{a(1-b)}}\right\}.$$

The correlation bounds (2.3.4) were given in Chaganty and Joe (2006). For fixed p_i, $1 \leq i \leq t$ and ρ satisfying the bounds (2.3.4), the joint probability distribution with AR(1) structure can be obtained starting with an initial Bernoulli distribution with mean p_1 and by the first order Markov chain with transition matrix

$$H_{(i-1),i} = \begin{array}{c} \\ y_{i-1} \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} q_i + \rho\frac{\sigma_{i-1}\sigma_i}{q_{i-1}} & p_i - \rho\frac{\sigma_{i-1}\sigma_i}{q_{i-1}} \\ q_i - \rho\frac{\sigma_{i-1}\sigma_i}{p_{i-1}} & p_i + \rho\frac{\sigma_{i-1}\sigma_i}{p_{i-1}} \end{bmatrix} \end{array} \begin{array}{c} y_i \\ \begin{array}{c} 0 \\ 1 \end{array} \end{array} \tag{2.3.5}$$

where $q_i = 1 - p_i, \sigma_i = (p_i q_i)^{1/2}, i = 1, 2, \cdots, t$. The transition matrix $H_{(i-1),i}$ gives us the conditional distribution of Y_i given the previous state Y_{i-1}. Note that $H_{(i-1),i}$ is a function of marginal probabilities p_i and p_{i-1} as well as the correlation parameter ρ. If the marginal distribution of Y_{i-1} is Bernoulli with mean p_{i-1} then the marginal distribution of Y_i is again Bernoulli with mean p_i, since

$$(q_{i-1},\, p_{i-1})\, H_{(i-1),i} = (q_i,\, p_i).$$

Further the transition probability matrix from state y_{i-1} to state y_j for $j \geq i$, is given by

$$\begin{aligned} H_{(i-1),j} &= H_{(i-1),i} H_{i,(i+1)} \cdots H_{(j-1),j} \\ &= \begin{bmatrix} q_j + \rho^{j-i+1}\frac{\sigma_{i-1}\sigma_j}{q_{i-1}} & q_j - \rho^{j-i+1}\frac{\sigma_{i-1}\sigma_j}{p_{i-1}} \\ p_j - \rho^{j-i+1}\frac{\sigma_{i-1}\sigma_j}{q_{i-1}} & p_j + \rho^{j-i+1}\frac{\sigma_{i-1}\sigma_j}{p_{i-1}} \end{bmatrix}. \end{aligned} \tag{2.3.6}$$

Thus the correlation between Y_{i-1} and Y_j for $j \geq i$ is ρ^{j-i+1}. When $p_1 = p_2 = \cdots = p_t = p$, that is, all the marginal probabilities are the same, the matrix $H_{(i-1),i}$ does not depend on i, and the Markov chain is a homogeneous, stationary chain.

For the above first order Markov chain model, the joint probability mass function for $Y = (Y_1, \cdots, Y_t)$ is given by

$$\begin{aligned} \pi(y) &= P(Y = y) \\ &= p_1^{y_1} q_1^{1-y_1} \prod_{i=2}^{t} \left[p_i^{y_i} q_i^{1-y_i} + (-1)^{y_{i-1}+y_i} \rho \frac{\sigma_{i-1}\sigma_i}{p_{i-1}^{y_{i-1}} q_{i-1}^{1-y_{i-1}}} \right]. \end{aligned} \tag{2.3.7}$$

This joint distribution is such that the marginal of Y_i is Bernoulli with mean p_i and $Corr(Y_i, Y_j) = \rho^{|i-j|}$ for $1 \leq i \neq j \leq t$ as shown above.

II.3.3 Generalization to Markov structures

The construction of the joint distribution for repeated binary observations with specified means and AR(1) correlation structure in the previous section can be generalized readily as follows. Assume that the initial distribution of Y_1 is Bernoulli with mean p_1, and for $i \geq 2$, let the transition from state Y_{i-1} to Y_i follow a first order Markov chain with transition matrix given by

$$H_{(i-1),i} = \begin{bmatrix} q_i + \rho_{i-1}\frac{\sigma_{i-1}\sigma_i}{q_{i-1}} & q_i - \rho_{i-1}\frac{\sigma_{i-1}\sigma_i}{p_{i-1}} \\ p_i - \rho_{i-1}\frac{\sigma_{i-1}\sigma_i}{q_{i-1}} & p_i + \rho_{i-1}\frac{\sigma_{i-1}\sigma_i}{p_{i-1}} \end{bmatrix}. \tag{2.3.8}$$

Matrix (2.3.8) is a legitimate transition matrix if $L(p_{(i-1)}, p_i) \leq \rho_{i-1} \leq U(p_{(i-1)}, p_i)$. As before, it is easy to see that from the above transition matrix the marginal distribution of Y_i is Bernoulli with mean p_i if the marginal distribution of $Y_{(i-1)}$ is Bernoulli with mean p_{i-1} and the correlation between $Y_{(i-1)}$ and Y_i is ρ_{i-1}. For the above transition matrices we have

$$\begin{aligned} H_{(i-1),j} &= H_{(i-1),i} H_{i,(i+1)} \cdots H_{(j-1),j} \\ &= \begin{bmatrix} q_j + \left(\prod_{k=(i-1)}^{(j-1)} \rho_k \right) \frac{\sigma_{i-1}\sigma_j}{q_{i-1}} & q_j - \left(\prod_{k=(i-1)}^{(j-1)} \rho_k \right) \frac{\sigma_{i-1}\sigma_j}{p_{i-1}} \\ p_j - \left(\prod_{k=(i-1)}^{(j-1)} \rho_k \right) \frac{\sigma_{i-1}\sigma_j}{q_{i-1}} & p_j + \left(\prod_{k=(i-1)}^{(j-1)} \rho_k \right) \frac{\sigma_{i-1}\sigma_j}{p_{i-1}} \end{bmatrix}. \end{aligned} \tag{2.3.9}$$

This is the transition matrix for moving from state y_{i-1} to y_j, for $j \geq i$. Hence $Corr(Y_{(i-1)}, Y_j) = \left(\prod_{k=(i-1)}^{(j-1)} \rho_k\right)$. In particular if the repeated binary observations are taken at time points $e_1 < e_2 < \cdots < e_t$, we could take $\rho_k = \rho^{e_{k+1}-e_k}$. Then $Corr(Y_{(i-1)}, Y_j)$ equals $\rho^{e_j - e_{i-1}}$, that is, the resulting correlation structure is the Markov structure. The Markov structure is an appropriate model for longitudinal data that is collected at different time points for different subjects. Similarly, we could also construct multivariate binary distributions with specified marginals and Generalized Markov structure. However note that for these structures the parameters must satisfy stringent bounds which are lot more restrictive than positive definiteness. See Crowder and Hand (1990) and Nunèz-Anton and Woodworth (1994) for a description and use of the Markov and Generalized Markov structures in the analysis of continuous longitudinal data.

II.3.4 Maximum likelihood estimation

Let $Y_i = (y_{i1}, y_{i2}, \cdots, y_{it_i})'$ be a vector of binary response with marginal means $p_i = (p_{i1}, p_{i2}, \cdots, p_{it_i})' \in (0,1)^{t_i}$ and let $X'_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ijk})$ be the corresponding k-dimensional row vector of covariates measured at time $j = 1, 2, \cdots, t_i$ for subject $i = 1, 2, \cdots, n$. Assume that the n subjects are independent. Suppose that the responses are related to the covariates as

$$E(y_{ij}) = p_{ij} = h(X'_{ij}\beta),$$

where $h(\cdot)$ is either the probit or the logit link function and β is the k-dimensional regression parameter.

Let $\theta = (\beta, \rho)$, then the likelihood for subject i is given by

$$\begin{aligned} L_i(\theta) &= \pi(y_i; X_i, \theta) \\ &= p_{i1}^{y_{i1}} q_{i1}^{1-y_{i1}} \prod_{j=2}^{t_i} \left[p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \rho \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}} \right], \end{aligned} \tag{2.3.10}$$

where $y_i = (y_{i1}, y_{i2}, \cdots, y_{it_i})'$ and $X'_i = (X_{i1}, X_{i2}, \ldots, X_{it_i})$. The maximum likelihood estimate $\widehat{\theta}_{MLE}$ is obtained by maximizing the likelihood function

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} \pi(y_i; X_i, \theta).$$

The log-likelihood function can be written as

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{n}\left[y_{i1}\log(p_{i1})+(1-y_{i1})\log(q_{i1})\right] \\ &\quad +\sum_{i=1}^{n}\sum_{j=2}^{t_i}\log\left[p_{ij}^{y_{ij}}q_{ij}^{1-y_{ij}}+(-1)^{y_{i(j-1)}+y_{ij}}\,\rho\,\frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}}\,q_{i(j-1)}^{1-y_{i(j-1)}}}\right] \\ &= \sum_{i=1}^{n}M_{i1}+\sum_{i=1}^{n}M_{i2}. \qquad (2.3.11)\end{aligned}$$

From the results in Anderson and Goodman (1957) it follows that there exists a consistent solution $\widehat{\theta}=(\widehat{\beta},\,\widehat{\rho})$ to the likelihood equation

$$\frac{\partial}{\partial\theta}\ell(\theta)=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ell_i(\theta)=0.$$

We will now derive a simplified expression for $\frac{\partial}{\partial\theta}\ell(\theta)$ and the Hessian matrix that can be used for computational purposes.

The Hessian matrix is the second derivative of likelihood function, and it is of the form

$$\begin{aligned}H(\theta) &= \frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'} \\ &= \sum_{i=1}^{n}\frac{\partial^2 M_{i1}}{\partial\theta\partial\theta'}+\sum_{i=1}^{n}\frac{\partial^2 M_{i2}}{\partial\theta\partial\theta'}. \qquad (2.3.12)\end{aligned}$$

The first and second derivatives of M_{i1} are as follows:

$$\frac{\partial M_{i1}}{\partial\beta} = \sum_{i=1}^{n}\left[y_{i1}\frac{1}{p_{i1}}\frac{\partial p_{i1}}{\partial\beta}-(1-y_{i1})\frac{1}{q_{i1}}\frac{\partial p_{i1}}{\partial\beta}\right], \qquad (2.3.13)$$

$$\begin{aligned}\frac{\partial^2 M_{i1}}{\partial\beta\partial\beta'} &= \sum_{i=1}^{n}\left[-y_{i1}\frac{1}{p_{i1}^2}\frac{\partial p_{i1}}{\partial\beta}\frac{\partial p_{i1}}{\partial\beta'}+y_{i1}\frac{1}{p_{i1}}\frac{\partial^2 p_{it}}{\partial\beta\partial\beta'}\right. \\ &\quad \left.-(1-y_{i1})\frac{1}{q_{i1}^2}\frac{\partial p_{i1}}{\partial\beta}\frac{\partial p_{i1}}{\partial\beta'}-(1-y_{i1})\frac{1}{q_{i1}}\frac{\partial^2 p_{i1}}{\partial\beta\partial\beta'}\right], \qquad (2.3.14)\end{aligned}$$

$$\frac{\partial M_{i1}}{\partial\rho} = 0. \qquad (2.3.15)$$

To get the first and second derivative of M_{i2} in (2.3.11), let P^*_{ij} be the conditional probability of $Y_{ij} = y_{ij}$ given that the previous state equals to $y_{i(j-1)}$, that is,

$$\begin{aligned} P^*_{ij} &= P\left\{Y_{ij} = y_{ij} | Y_{i(j-1)} = y_{i(j-1)}\right\} \\ &= p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \rho \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}}. \end{aligned}$$

The derivative of M_{i2} is a sum of the derivatives of $\log(P^*_{ij})$ at each time points j for each of the n subjects. Let $G_{ij} = p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}}$, we then have

$$\begin{aligned} \frac{\partial G_{ij}}{\partial \beta} &= \left[y_{ij} p_{ij}^{y_{ij}-1} \frac{\partial p_{ij}}{\partial \beta}\right] q_{ij}^{1-y_{ij}} + p_{ij}^{y_{ij}} \left[(1-y_{ij}) q_{ij}^{-y_{ij}} \left(-\frac{\partial p_{ij}}{\partial \beta}\right)\right] \\ &= \left[y_{ij}\, p_{ij}^{y_{ij}-1} q_{ij}^{1-y_{ij}} - (1-y_{ij}) p_{ij}^{y_{ij}} q_{ij}^{-y_{ij}}\right] \frac{\partial p_{ij}}{\partial \beta}, \end{aligned} \tag{2.3.16}$$

$$\begin{aligned} \frac{\partial^2 G_{ij}}{\partial \beta \partial \beta'} &= y_{ij}(y_{ij}-1) p_{ij}^{y_{ij}-2} q_{ij}^{1-y_{ij}} \frac{\partial p_{ij}}{\partial \beta} \frac{\partial p_{ij}}{\partial \beta'} + y_{ij} p_{ij}^{y_{ij}-1} q_{ij}^{1-y_{ij}} \frac{\partial^2 p_{ij}}{\partial \beta \partial \beta'} \\ &\quad -2y_{ij}(1-y_{ij}) p_{ij}^{y_{ij}-1} q_{ij}^{-y_{ij}} \frac{\partial p_{ij}}{\partial \beta} \frac{\partial p_{ij}}{\partial \beta'} \\ &\quad -y_{ij}(1-y_{ij}) p_{ij}^{y_{ij}} q_{ij}^{-y_{ij}-1} \frac{\partial p_{ij}}{\partial \beta} \frac{\partial p_{ij}}{\partial \beta'} - (1-y_{ij}) p_{ij}^{y_{ij}} q_{ij}^{-y_{ij}} \frac{\partial^2 p_{ij}}{\partial \beta \partial \beta'}. \end{aligned} \tag{2.3.17}$$

Then

$$\begin{aligned} \frac{\partial M_{i2}}{\partial \beta} &= \sum_{j=2}^{t_i} \frac{\partial}{\partial \beta} \log(P^*_{ij}) = \sum_{j=2}^{t_i} \frac{1}{P^*_{ij}} \frac{\partial P_{ij}^{y_{i(j-1)} y_{ij}}}{\partial \beta} \\ &= \sum_{j=2}^{t_i} \frac{1}{P^*_{ij}} \left[\frac{\partial G_{ij}}{\partial \beta} + (-1)^{y_{i(j-1)}+y_{ij}} \rho \frac{\frac{\partial \sigma_{i(j-1)}\sigma_{ij}}{\partial \beta} G_{i(j-1)} - \sigma_{i(j-1)}\sigma_{ij} \frac{\partial G_{i(j-1)}}{\partial \beta}}{(G_{i(j-1)})^2}\right], \end{aligned}$$

$$\frac{\partial^2 M_{i2}}{\partial\beta\partial\beta'} = \sum_{j=2}^{t_i} -\frac{1}{(P^*_{ij})^2}\frac{\partial P^*_{ij}}{\partial\beta}\frac{\partial P^*_{ij}}{\partial\beta'} + \frac{1}{P^*_{ij}}\left[\frac{\partial^2 G_{ij}}{\partial\beta\partial\beta'} + (-1)^{y_{i(j-1)}+y_{ij}}\rho\frac{\frac{\partial}{\partial\beta'}\left(\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta}G_{i(j-1)} - \sigma_{i(j-1)}\sigma_{ij}\frac{\partial G_{i(j-1)}}{\partial\beta}\right)}{(G_{i(j-1)})^2}\right.$$
$$\left. + (-1)^{y_{i(j-1)}+y_{ij}}2\rho\frac{\left(\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta}G_{i(j-1)} - \sigma_{i(j-1)}\sigma_{ij}\frac{\partial G_{i(j-1)}}{\partial\beta}\right)}{G^3_{i(j-1)}}\right], \tag{2.3.18}$$

$$\frac{\partial M_{i2}}{\partial\rho} = \sum_{j=2}^{t_i}\frac{1}{P^*_{ij}}\left[(-1)^{y_{i(j-1)}+y_{ij}}\frac{\sigma_{i(j-1)}\sigma_{ij}}{p^{y_{i(j-1)}}_{i(j-1)}\,q^{1-y_{i(j-1)}}_{i(j-1)}}\right], \tag{2.3.19}$$

$$\frac{\partial^2 M_{i2}}{\partial\rho^2} = \sum_{j=2}^{t_i} -\left\{\frac{1}{P^*_{ij}}\left[(-1)^{y_{i(j-1)}+y_{ij}}\frac{\sigma_{i(j-1)}\sigma_{ij}}{p^{y_{i(j-1)}}_{i(j-1)}\,q^{1-y_{i(j-1)}}_{i(j-1)}}\right]\right\}^2, \tag{2.3.20}$$

$$\frac{\partial^2 M_{i2}}{\partial\beta\partial\rho} = \sum_{j=2}^{t_i}\frac{1}{P^*_{ij}}\left[(-1)^{y_{i(j-1)}+y_{ij}}\frac{\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta}G_{i(j-1)} - \sigma_{i(j-1)}\sigma_{ij}\frac{\partial G_{i(j-1)}}{\partial\beta}}{(G_{i(j-1)})^2}\right], \tag{2.3.21}$$

where

$$\frac{\partial}{\partial\beta'}\left(\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta}G_{i(j-1)} - \sigma_{i(j-1)}\sigma_{ij}\frac{\partial G_{i(j-1)}}{\partial\beta}\right)$$
$$= \frac{\partial^2\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta\partial\beta'}G_{i(j-1)} + \frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta}\frac{\partial G_{i(j-1)}}{\partial\beta'} - \frac{\partial G_{i(j-1)}}{\partial\beta}\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta'} + \sigma_{i(j-1)}\sigma_{ij}\frac{\partial^2 G_{i(j-1)}}{\partial\beta\partial\beta'},$$

$$\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta} = \frac{\partial}{\partial\beta}\sqrt{p_{ij}q_{ij}p_{i(j-1)}q_{i(j-1)}}$$
$$= \frac{1}{2\sigma_{i(j-1)}\sigma_{ij}}\left[(1-2p_{i(j-1)})p_{ij}q_{ij}\frac{\partial p_{i(j-1)}}{\partial\beta} + p_{i(j-1)}q_{i(j-1)}(1-2p_{ij})\frac{\partial p_{ij}}{\partial\beta}\right],$$

$$\begin{aligned}\frac{\partial^2\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta\partial\beta'} &= -\frac{1}{\sigma_{i(j-1)}\sigma_{ij}}\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta}\frac{\partial\sigma_{i(j-1)}\sigma_{ij}}{\partial\beta'} \\ &+\frac{1}{2\sigma_{i(j-1)}\sigma_{ij}}\left[\left(\frac{\partial^2 p_{i(j-1)}}{\partial\beta\partial\beta'}-2\frac{\partial p_{i(j-1)}}{\partial\beta}\frac{\partial p_{i(j-1)}}{\partial\beta'}-2p_{i(j-1)}\frac{\partial^2 p_{i(j-1)}}{\partial\beta\partial\beta'}\right)(p_{ij}-p_{ij}^2)\right. \\ &+(1-2p_{i(j-1)})(1-2p_{ij})\frac{\partial p_{i(j-1)}}{\partial\beta}\frac{\partial p_{ij}}{\partial\beta'}+(1-2p_{i(j-1)})(1-2p_{ij})\frac{\partial p_{ij}}{\partial\beta}\frac{\partial p_{i(j-1)}}{\partial\beta'} \\ &\left.+(p_{i(j-1)}-p_{i(j-1)}^2)\left(\frac{\partial^2 p_{ij}}{\partial\beta\partial\beta'}-2\frac{\partial p_{ij}}{\partial\beta}\frac{\partial p_{ij}}{\partial\beta'}-2p_{ij}\frac{\partial^2 p_{ij}}{\partial\beta\partial\beta'}\right)\right].\end{aligned}$$

Using the above expressions we can write the derivative of the log-likelihood function and the Hessian matrix as

$$\frac{\partial}{\partial\theta}\ell(\theta) = \begin{pmatrix} \sum_{i=1}^{n}\frac{\partial M_{i1}}{\partial\beta}+\sum_{i=1}^{n}\frac{\partial M_{i2}}{\partial\beta} \\ \\ \sum_{i=1}^{n}\frac{\partial M_{i2}}{\partial\rho}\end{pmatrix}, \tag{2.3.22}$$

and

$$H = \begin{pmatrix} \sum_{i=1}^{n}\frac{\partial^2 M_{i1}}{\partial\beta\partial\beta'}+\sum_{i=1}^{n}\frac{\partial^2 M_{i2}}{\partial\beta\partial\beta'} & \sum_{i=1}^{n}\frac{\partial^2 M_{i2}}{\partial\beta\partial\rho} \\ \\ \sum_{i=1}^{n}\frac{\partial^2 M_{i2}}{\partial\beta\partial\rho} & \sum_{i=1}^{n}\frac{\partial^2 M_{i2}}{\partial\rho^2}\end{pmatrix}. \tag{2.3.23}$$

We have used the above expressions (2.3.22) and (2.3.23) to develop a computer program to find the maximum likelihood estimates for the first order Markov chain model.

It is well known that the maximum likelihood estimator $\hat{\theta}_{MLE}$ has an asymptotic normal distribution:

$$(\hat{\theta}_{MLE}-\theta)\sim\mathcal{N}(0,\mathcal{I}^{-1}),$$

where $\mathcal{I}$ is the Fisher information in n subjects, and it can be calculated as

$$\mathcal{I} = \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial \ell(\theta)}{\partial \theta}\frac{\partial \ell(\theta)}{\partial \theta'}\right] = \sum_{i=1}^{n}\sum_{y_i} \pi(y_i; X_i, \theta)\left[\frac{\partial \ell(\theta)}{\partial \theta}\frac{\partial \ell(\theta)}{\partial \theta'}\right], \qquad (2.3.24)$$

where the inner sum is taken over the 2^{t_i} possible vectors of y_i. The diagonal elements of $\mathcal{I}^{-1}$ are the asymptotic variances of the parameter estimates. We compute the Fisher information matrix in the software that we developed to estimate the standard errors of the maximum likelihood estimates.

II.4 Hypothesis testing

In this section we develop hypothesis testing procedures for testing the significance of the correlation parameter ρ. We develop hypothesis tests for general functions of the correlation parameters, and then concentrate upon specific examples and compare their performance through simulations.

II.4.1 Likelihood ratio test

Knowledge of the likelihood function allows us to utilize a likelihood ratio test for hypothesis tests regarding the correlation parameters.

Generally, we would like to test a null hypothesis that the correlation parameter ρ is equal to some constant, or $H_0 : \rho = \rho_0$. To do this, we take the ratio of the likelihood evaluated with the maximum likelihood estimates under H_0 (the restricted MLE's) against the likelihood evaluated with the so-called unrestricted maximum likelihood estimates. Let $\widehat{\theta}_0 = (\widehat{\beta}_0, \widehat{\rho}_0)$ be the restricted and $\widehat{\theta} = (\widehat{\beta}, \widehat{\rho})$ the

unrestricted MLE's, respectively. Then the likelihood ratio test statistic is

$$\begin{aligned}
\lambda(\widehat{\theta}_0, \widehat{\theta}) &= \frac{\prod_{i=1}^{n} \pi_i(y_i|\widehat{\theta}_0)}{\prod_{i=1}^{n} \pi_i(y_i|\widehat{\theta})} \\
&= \frac{p_{i1}^{y_{i1}} q_{i1}^{1-y_{i1}} \prod_{j=2}^{t_i} \left[p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \widehat{\rho}_0 \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}} \right]}{p_{i1}^{y_{i1}} q_{i1}^{1-y_{i1}} \prod_{j=2}^{t_i} \left[p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \widehat{\rho} \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}} \right]} \\
&= \prod_{j=2}^{t_i} \frac{\left[p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \widehat{\rho}_0 \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}} \right]}{\left[p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \widehat{\rho} \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}} \right]}.
\end{aligned}$$

For ease of notation we wrote p_{ij} and q_{ij} in the above for the estimated probabilities that depend on the estimated regression parameter β. The most obvious special case is the hypothesis $H_0 : \rho = 0$. Thus, the test statistic becomes

$$\lambda(\widehat{\theta}_0, \widehat{\theta}) = \prod_{j=2}^{t_i} \frac{p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}}}{p_{ij}^{y_{ij}} q_{ij}^{1-y_{ij}} + (-1)^{y_{i(j-1)}+y_{ij}} \widehat{\rho} \frac{\sigma_{i(j-1)}\sigma_{ij}}{p_{i(j-1)}^{y_{i(j-1)}} q_{i(j-1)}^{1-y_{i(j-1)}}}}.$$

Recall that that $-2\ln(\lambda(\widehat{\theta}_0, \widehat{\theta}))$ has an asymptotic chi-square distribution with $d = d_{ur} - d_r$ degrees of freedom, where d_{ur} is the number of parameters in the unrestricted model and d_r is the number of parameters under the null hypothesis. Since the difference in the number of parameters between $\widehat{\theta}_0 = (\widehat{\beta}_0, 0)$ and $\widehat{\theta} = (\widehat{\beta}, \widehat{\rho})$ is 1, then $-2\ln(\lambda(\widehat{\theta}_0, \widehat{\theta}))$ is asymptotically chi-square with one degree of freedom.

II.4.2 Wald's test

Another hypothesis testing procedure based on the maximum likelihood estimates is the Wald's test. Under this test procedure, the maximum likelihood estimate $\widehat{\rho}$ of the correlation parameter ρ is compared with the hypothesized value ρ_0, with the assumption that the difference between the two will be approximately normal. Typically, the square of the difference is compared to a chi-squared distribution. In the univariate case, the Wald's test statistic is

$$\frac{(\widehat{\rho} - \rho_0)^2}{Var(\widehat{\rho})}.$$

Alternatively, we could use the test statistic

$$\frac{(\widehat{\rho} - \rho_0)}{\mathrm{SE}(\widehat{\rho})},$$

where $\mathrm{SE}(\widehat{\rho})$ is the standard error of the maximum likelihood estimate, and it is estimated by the inverse of the Fisher information for the parameter ρ. This statistic is approximately distributed as standard normal.

II.4.3 Estimated power of the test statistics

To gauge the performance of the likelihood ratio test and the Wald's test, we make use of simulations to estimate the power of these tests. We simulated $n = 30$ observations of size $t = 4$. For the likelihood ratio test, we calculated both the restricted and unrestricted maximum likelihood estimates, which should be similar as the simulated data reflect the conditions stated in the null hypothesis. For the Wald's test, we calculated the maximum likelihood estimate and the Fisher information using the simulated data. Since we are simulating data using nonzero value for ρ and the null hypothesis is $H_0 : \rho = 0$, we expect to reject the null hypothesis. Recall that for each test we reject H_0 if the test statistic is greater than a chi-square critical value χ^2_1 for a particular significance level. We chose the significance level to be $\alpha = 0.05$. If we repeat these simulations a large number times for a particular value of the non-zero correlation parameter, then the estimated power of the test is the ratio of the number of times we reject the null hypothesis to the total number of repeated simulations. If we then repeat this procedure over a wide range of values for the non-zero correlation parameter, we can get an idea of how the test performs in many scenarios.

The estimated power of the likelihood ratio test (LRT) and the Wald's test are graphed in Figure 2.1. This shows that for testing $H_0 : \rho = 0$, the power functions for both the tests are increasing with ρ, converging to one as ρ moves far away from zero. The graph of the power functions for both the tests are almost identical. Thus for testing $H_0 : \rho = 0$, the two LRT and Wald's test perform similarly.

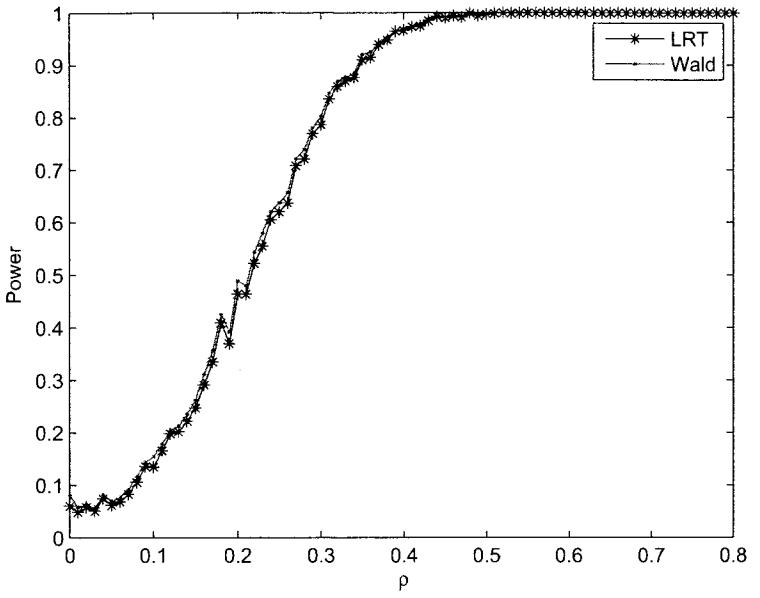


Figure 2.1: Estimated power for LRT and Wald test

II.5 SAS macro TMMLE

To fit the first order Markov chain model for longitudinal binary data, we developed a SAS MACRO called TMMLE. It uses the estimate of the generalized estimating equation (GEE) method from PROC GENMOD as the initial value for nonlinear optimization by Newton-Raphson ridge method. The main calculations concerning the search of the maximum likelihood estimate and an output delivery system are carried out by PROC IML. At each iteration, the macro calculates feasible range of the correlation parameter ρ based on the current estimate of the regression parameter and checks whether the estimated ρ is within the range, if not, we replace the estimated ρ by the midpoint of the feasible range.

The current version of the macro implements two link functions: logit and probit. The macro can be invoked using the command:

%tmmle(data=, yvar=, xvar=, id=, fun=, outpar=, outllf=) where

data	the SAS data set containing the data to be analyzed
yvar	name of the dependent variable
xvar	names of the independent variables
id	group or cluster id, should be a vector of positive integers.
fun	options of link functions
outpar	output of the parameter estimator, standard error and p-values
outllf	output of the value of log-likelihood function and AIC,BIC values

II.6 Multivariate probit model

Another fully specified model for binary data is the multivariate probit model. Multivariate probit model belongs to the class of latent variable thresholds models for analyzing binary dependent data. The model assumes that the binary response is the indicator of an unobserved latent variable exceeding a given threshold. Estimation of the regression and latent correlation parameters can be done in a likelihood framework.

II.6.1 Likelihood function

The multivariate probit model can be described as follows. Let $Y = (y_1, y_2, \cdots, y_t)'$ be a vector of binary random variables. We assume that there exists a corresponding latent continuous multivariate normal random variable $Z = (z_1, z_2, \cdots, z_t)'$ such that

$$Y_j = \begin{cases} 1 & Z_j > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We assume that $Z_j = \mu_j + \varepsilon_j$, for $j = 1, 2, \cdots, t$, and

$$(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_t)' \sim MVN(0, R),$$

where R is a correlation matrix known as the latent correlation matrix. The marginal mean of y_j is given by

$$E(y_j) = P(y_j = 1) = p_j = P(\mu_j + \varepsilon_j > 0) = 1 - \Phi(-\mu_j) = \Phi(\mu_j), \qquad (2.6.1)$$

where Φ is the standard normal cumulative distribution function. Thus the probability mass function of Y can be expressed as

$$P(y;\mu,\ R) = \int_{D_t} \cdots \int_{D_1} \frac{1}{(2\pi)^{t/2}|R|^{1/2}} \exp\left\{-\frac{\varepsilon' R^{-1}\varepsilon}{2}\right\} d\varepsilon, \tag{2.6.2}$$

where $y = (y_1, \ldots, y_t)'$ is a binary vector, $\mu = (\mu_1, \ldots, \mu_t)'$ and

$$D_j = \begin{cases} (-\infty, \mu_j) & \text{if } y_j = 1 \\ (\mu_j, \infty) & \text{if } y_j = 0 \end{cases}$$

for $j = 1, 2, \cdots, t$. Let Φ_t be the t-variate normal cumulative distribution function, then the equation (2.6.2) can also be written as

$$P(y;\mu,R) = \Phi_t[(-1)^{1-y_1}\mu_1, \cdots, (-1)^{1-y_t}\mu_t; 0, CRC'], \tag{2.6.3}$$

where C is a diagonal matrix with the jth diagonal element is $(-1)^{1-y_j}$. Although equation (2.6.2) defines a proper probability distribution function for any positive definite matrix R, in order to retain likelihood identifiability R is restricted to be a correlation matrix.

Suppose the correlation structure of $Y = (y_1, \ldots, y_t)$ is the first order autoregressive (AR(1)) with parameter ρ, then

$$\text{Corr}(y_j,\ y_k) = \rho^{|j-k|} = \frac{\Phi_2(\mu_j, \mu_k; r_{jk}) - p_j p_k}{[p_j(1-p_j)p_j(1-p_j)]^{1/2}}. \tag{2.6.4}$$

Here Φ_2 is the bivariate normal distribution function with density function

$$f(\varepsilon_j, \varepsilon_k, r_{jk}) = \frac{1}{2\pi\sqrt{1-r_{jk}^2}} \exp\left\{-\frac{\varepsilon_j^2 - 2r_{jk}\varepsilon_j\varepsilon_k + \varepsilon_k^2}{2(1-r_{jk}^2)}\right\}.$$

For any ρ satisfying the range restrictions (2.3.4), equation (2.6.4) can be solved uniquely for r_{jk} for $j, k = 1, \cdots, t$ (Emrich and Piedmonte (1991)), using the bisection method. We can then construct a probability distribution with specified mean and AR(1) correlation structure with specified ρ using the multivariate probit model (2.6.1) with correlation matrix $R = (r_{jk})$.

II.7 Comparison of the transistion and multivariate probit models

In this section we shall compare the probability mass functions generated by the transistion and multivariate probit models. Suppose $Y = (y_1, y_2, \ldots, y_t)$ is a binary

Table 2.1: Mass function generated from transition model and multivariate probit model

Y				$P_{TM}(Y)$	$P_{MPROBIT}(Y)$
0	0	0	0	0.0538321	0.0538567
0	0	0	1	0.1643539	0.1654672
0	0	1	0	0.0090897	0.0095023
0	0	1	1	0.3407123	0.3391005
0	1	0	0	0.0005554	0.0004391
0	1	0	1	0.0016957	0.0006700
0	1	1	0	0.0025923	0.0022505
0	1	1	1	0.0971685	0.0987233
1	0	0	0	0.0163028	0.0156603
1	0	0	1	0.0497737	0.0492761
1	0	1	0	0.0027528	0.0029030
1	0	1	1	0.1031828	0.1041499
1	1	0	0	0.0008602	0.0015572
1	1	0	1	0.0026262	0.0030677
1	1	1	0	0.0040148	0.0037775
1	1	1	1	0.1504868	0.1496076

NOTE: Parameters are $\rho = 0.35$, $p = (0.33, 0.26, 0.71, 0.91)$. Latent correlations are $r_{12} = 0.549$, $r_{13} = 0.216$, $r_{14} = 0.101$, $r_{23} = 0.746$, $r_{24} = 0.350$ and $r_{34} = 0.668$.

vector with marginal mean $p = (p_1, p_2, \ldots, p_t)$ and AR(1) correlation structure with a parameter ρ. For given p and feasible ρ, we can find a vector μ and a unique latent correlation matrix R by solving equation (2.6.4). Then μ and R can be used to generate a probability distribution for Y with mean vector p and AR(1) correlation structure with parameter ρ.

Table 2.1 shows the mass function generated by Markov chain model and multivariate probit model with marginal probability $p = (0.33, 0.26, 0.71, 0.91)$ and correlation parameter $\rho = 0.35$, which is within the feasible range $(-0.2010, 0.3788)$. We can see that the values of the two mass functions are very close, but are not exactly the same. For example, when $y = (0, 1, 0, 1)$ the corresponding probabilities generated by two models are 0.00169 and 0.00067, respectively. It can be shown theorctically that the mass functions generated by these two models are different.

Here, we give the proof when $d = 3$, similar proof holds for higher dimensions.

First, it is clear that $\mathrm{P}(y_i = 1) = p_i = \mathrm{P}(Z_i \leq z_i) = \Phi(z_i)$, $i = 1, 2, 3$. Since the latent correlation r_{ij} is obtained by solving equation (2.6.4), we have

$$\mathrm{P}(y_i = 1, y_j = 1) = p_i p_j + \rho^{j-i}\sigma_i\sigma_j = \Phi(z_i, z_j; r_{ij}) = \mathrm{P}(Z_i \leq z_i, Z_j \leq z_j),$$

for $1 \leq i < j \leq 3$. For the first order Markov chain model, we have

$$\begin{aligned}
\mathrm{P}(y_1 = 1, y_2 = 1, y_3 = 1) &= \mathrm{P}(y_3 = 1 | y_2 = 1, y_1 = 1)\mathrm{P}(y_1 = 1, y_2 = 1) \\
&= \mathrm{P}(y_3 = 1 | y_2 = 1)\mathrm{P}(y_1 = 1, y_2 = 1) \\
&= \left[p_3 + \frac{\sigma_2\sigma_3}{p_2}\right] \mathrm{P}(y_1 = 1, y_2 = 1).
\end{aligned}$$

And for the multivariate probit model, we have

$$\begin{aligned}
\mathrm{P}(y_1 = 1, y_2 = 1, y_3 = 1) &= \mathrm{P}(Z_1 \leq z_1, Z_2 \leq z_2, Z_3 \leq z_3) \\
&= \mathrm{P}(Z_3 \leq z_3 | Z_2 \leq z_2, Z_1 \leq z_1)\mathrm{P}(Z_1 \leq z_1, Z_2 \leq z_2).
\end{aligned}$$

The two models will generate the same probability distribution if and only if

$$\mathrm{P}(Z_3 \leq z_3 | Z_2 \leq z_2, Z_1 \leq z_1) = p_3 + \frac{\sigma_2\sigma_3}{p_2}. \tag{2.7.1}$$

Let $I_i = I(Z_i \leq z_i)$, $i = 1, 2, 3$ be the indicator functions. Then

$$\begin{aligned}
\mathrm{P}(Z_3 \leq z_3 | Z_2 \leq z_2, Z_1 \leq z_1) &= \mathrm{P}(I_3 = 1 | I_2 = 1, I_1 = 1) \\
&= \mathrm{E}(I_3 = 1 | I_2 = 1, I_1 = 1).
\end{aligned} \tag{2.7.2}$$

According to Drezner (1990) and Joe (1995), an approximation to (2.7.2) is given by

$$\mathrm{E}(I_3) + \Omega_{21}\Omega_{11}^{-1} \begin{pmatrix} 1 - \mathrm{E}(I_1) \\ 1 - \mathrm{E}(I_2) \end{pmatrix}, \tag{2.7.3}$$

where

$$\begin{aligned}
\Omega_{21} &= (\mathrm{Cov}(I_3, I_1), \mathrm{Cov}(I_3, I_2)), \\
\Omega_{11} &= \begin{pmatrix} \mathrm{Var}(I_1) & \mathrm{Cov}(I_1, I_2) \\ \mathrm{Cov}(I_1, I_2) & \mathrm{Var}(I_2) \end{pmatrix}.
\end{aligned}$$

Notice that

$$\begin{aligned}\mathrm{Cov}(I_i, I_j) &= \mathrm{E}(I_i I_j) - \mathrm{E}(I_i)\mathrm{E}(I_j) \\ &= \mathrm{P}(Z_i \le z_i, Z_j \le z_j) - p_i p_j \\ &= p_i p_j + \rho^{j-i}\sigma_i\sigma_j - p_i p_j \\ &= \rho^{j-i}\sigma_i\sigma_j, \end{aligned} \tag{2.7.4}$$

for $1 \le i < j \le 3$. Thus, we have

$$\begin{aligned}\Omega_{21} &= (\rho^2\sigma_1\sigma_3 \;\; \rho\sigma_2\sigma_3) = \rho\sigma_3(\rho\sigma_1 \;\; \sigma_2), \\ \Omega_{11} &= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}, \\ \Omega_{11}^{-1} &= \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}\frac{1}{1-\rho^2}, \end{aligned}$$

$$\begin{aligned}\Omega_{21}\Omega_{11}^{-1} &= \rho\sigma_3(\rho\sigma_1 \;\; \sigma_2)\begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}\frac{1}{1-\rho^2} \\ &= \rho\sigma_3(\rho \;\; 1)\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}\frac{1}{1-\rho^2} \\ &= \rho\sigma_3(0 \;\; 1-\rho^2)\begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}\frac{1}{1-\rho^2} \\ &= \rho\sigma_3(0 \;\; 1)\begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix} \\ &= \rho\sigma_3\begin{pmatrix} 0 & \frac{1}{\sigma_2} \end{pmatrix}. \end{aligned}$$

Thus, (2.7.3) is equal to

$$p_3 + \rho\sigma_3 \begin{pmatrix} 0 & \frac{1}{\sigma_2} \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = p_3 + \rho\frac{\sigma_2\sigma_3}{p_2},$$

which is the right hand side of (2.7.1). Thus the first Markov chain model is approximately equal to the multivariate probit model but they are not the same.

Even though equation (2.6.4) can be solved unique for r_{jk} for any ρ satisfying the range restrictions, but the solution does not guarantee that the latent correlation matrix will be a positive definite matrix. And multivariate probit model may fail to generate a probability distribution of Y if latent correlation matrix is not a positive definite matrix. For example if $p = (0.26, 0.36, 0.25, 0.24)$, the feasible range for ρ is $(-0.3244, 0.7698)$. For $\rho = 0.7$, solving equation (2.6.4) we get

$$R = \begin{bmatrix} 1 & 0.9378 & 0.7511 & 0.5869 \\ 0.9378 & 1 & 0.9460 & 0.7657 \\ 0.7511 & 0.9460 & 1 & 0.9157 \\ 0.5869 & 0.7657 & 0.9157 & 1 \end{bmatrix}.$$

But this is not a positive definite matrix.

II.7.1 Maximum likelihood estimate

For the multivariate probit model (2.6.1), suppose that $\mu_j = x_j'\beta$ where x_j is the covariate vector and β is the regression parameter. In many application it is reasonable to assume that the latent correlation matrix R is a structured matrix characterized common parameter α. Then the unknown parameter $\theta = (\beta, \alpha)$ can be estimated using the maximum likelihood estimation. Note that if we have n subjects $Y_1, \cdots, Y_n$, and the correlation structures of Y_i are AR(1) with a common parameter ρ, then it is unlikely to have a common latent correlation matrix R for all subjects Y_i, since the solution of equation (2.6.4) changes when the marginal means P_i are different for different subjects.

Let Y_i be a t_i-dimensional vector of binary response with marginal means p_i and let X_{ij}' be the corresponding k-dimensional row vector of covariates measured at

time $j = 1, 2, \cdots, t_i$ for subject $i = 1, 2, \cdots, n$. Assume that the n subjects are independent. Suppose that the responses are related to the covariates as

$$E(y_{ij}) = p_{ij} = \Phi(X'_{ij}\beta) = \Phi(\mu_{ij}),$$

where $\Phi(\cdot)$ is standard normal distribution function and β is the k-dimensional regression parameter.

Then the likelihood for subject i is given by the form of (2.6.2)

$$\begin{aligned} L_i(\theta) &= P(y_i;\theta) \\ &= P(y_i;\theta) = \int_{D_{it_i}} \cdots \int_{D_{i1}} \frac{1}{(2\pi)^{t/2}|R|^{1/2}} \exp\left\{-\frac{\varepsilon' R^{-1}\varepsilon}{2}\right\} d\varepsilon, \quad (2.7.5) \end{aligned}$$

where $y_i = (y_{i1}, \ldots, y_{it_i})'$ is a binary vector, $\mu_i = (\mu_{i1}, \ldots, \mu_{it_i})'$ and

$$D_{ij} = \begin{cases} (-\infty, \mu_{ij}) & \text{if } y_{ij} = 1 \\ (\mu_{ij}, \infty) & \text{if } y_{ij} = 0 \end{cases}$$

for $j = 1, 2, \cdots, t_i$.

The maximum likelihood estimate $\widehat{\theta}_{MLE}$ is obtained by maximizing the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log L_i(\theta) = \sum_{i=1}^{n} \log P(y_i;\theta).$$

The log-likelihood function can be written as

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{P(y_i;\theta)} J_i(\theta)' \nabla P(y_i;\theta),$$

where the jacobian matrix $J_i(\theta) = J_i$ is independent of θ and it is given by

$$J_i = \begin{bmatrix} X_i & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$\nabla P(y_i;\theta) = \left[\frac{\partial P(y_i;\theta)}{\partial \mu_i}, \frac{\partial P(y_i;\theta)}{\partial \alpha}\right]'.$$

The j-th element of $\partial P(y_i;\theta)/\partial \mu_i$ is

$$\frac{\partial P(y_i;\beta)}{\partial \mu_{ij}} = (-1)^{1+y_{ij}} \phi(\mu_{ij}) \mathrm{P}(y_{i,j^*}; \mu_{i,j^*|j}, R_{i,j^*|j}),$$

where j^* is the vector of index that is complementary to j. $\mu_{i,j^*|j}$ and $R_{i,j^*|j}$ are the conditional mean and conditional covariance matrix of $Z_{i,j*}$ given Z_{ij}.

It is well known that $\hat{\theta}$ is AMVN(θ, $\mathcal{I}^{-1}(\theta)$) where

$$\mathcal{I}(\theta) = \sum_{i=1}^{n} J_i' E\left[\frac{1}{P(y_i;\theta)} \nabla P(y_i;\theta) \nabla P(y_i;\theta)'\right] J_i. \tag{2.7.6}$$

II.8 Comparison of asymptotic performance

The two estimating procedures that we discussed yield consistent estimates of the regression parameters. In this section we study their asymptotic performance via calculating and comparing the asymptotic variance of the regression parameter for both the methods. The asymptotic relative efficiency (ARE) is calculated using the following formula:

$$\text{ARE} = \frac{\text{Asymptotic variance of Markov chain model}}{\text{Asymptotic variance of multivariate probit model}}.$$

If the above quantity is less than one, we could conclude that the Markov chain model is better than the multivariate probit model and vice versa if the quantity is more than one. In the simulations we used the probit link function and a large sample size $n = 500$ so that the simulated variances will approximate the asymptotic variances. We used the following mean function for the repeated binary variables in our efficiency calculations

$$\Phi^{-1}(p_{it}) = \mu_{it} = \beta_0 + \beta_1 x_{it}^C + \beta_2 x_{it}^D. \tag{2.8.1}$$

To begin with we set the number of repeated measurements as three. The covariates are chosen as follows: the first covariate x_{it}^C is continuous and taken as uniform on (0,1), the second covariate x_{it}^D is discrete and takes values $\{-1.5, -0.5, 0.5\}$. We fixed the true regression coefficients as $\beta_0 = 1.2$, $\beta_1 = 0.34$, $\beta_2 = -0.15$. With these values we calculated the means and the bounds for the correlation parameter. For different values ρ within the bounds we calculated the Fisher information matrices for the regression parameter for both the transition model and the probit model. Table 2.2 shows the the results for the model that contains only the intercept and the continuous covariate x_{it}^C. The bound of ρ is at the bottom of the table. For

Table 2.2: Asymptotic variance and ARE with one covariate, $t = 3$

	$nV(\beta_0)$			$nV(\beta_1)$		
	MC	MPROBIT	ARE	MC	MPROBIT	ARE
$\rho = 0.0$	3.7432	3.7432	1.000	12.7559	12.7559	1.000
$\rho = 0.1$	3.7832	3.8009	0.995	12.3962	12.5057	0.991
$\rho = 0.2$	3.7614	3.7883	0.993	11.7153	11.9092	0.984
$\rho = 0.3$	3.6967	3.7193	0.994	10.8064	11.0117	0.981
$\rho = 0.4$	3.5915	3.6050	0.996	9.6753	9.8466	0.983
$\rho = 0.5$	3.4444	3.4549	0.997	8.3081	8.4367	0.985
$\rho = 0.6$	3.2472	3.2752	0.991	6.6565	6.7823	0.981

NOTE: Range of ρ is $(-0.0661, 0.7170)$. The parameter values are $\beta_0 = 1.2$, $\beta_1 = 0.34$ and $n = 500$.

different values of ρ within the bounds, the asymptotic variances scaled by n for both the methods as well as the asymptotic relative efficiency (ARE) of the Markov chain model with respect to the multivariate probit model are in Table 2.2. Since the two models are identical when ρ is zero the ARE is one. However, for all positive values of ρ the values of ARE are all less than 1 for both the regression coefficients. Hence in this case for estimating β, the Markov chain model is more efficient than the multivariate probit model.

We considered next the model which includes the discrete covariate x_{it}^D as well. The simulation results are given in Table 2.3. When $\rho = 0$, the situation is same as before, the two method are identical and hence the AE is equal to 1. But for nonzero values of ρ, the performance of Markov chain model is not uniformly better than the multivariate probit model. Table 2.3 shows that the ARE for β_0, is less than one at the beginning and is decreasing as ρ increases from 0.1 to 0.3, and then it increases to 1.013 as ρ increases to 0.7. Thus when there is a high correlation the multivariate probit model is better at estimating the slope. However, the Markov chain model is uniformly better than the multivariate probit model for estimating the slopes β_1 and β_2.

We carried out simulations increasing the number of repeated measurements to four. In this case, we chose the first covariate x_{it}^C to be the same as before; continuous

Table 2.3: Asymptotic variance and ARE with two covariates, $t = 3$

	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.0$	3.5750 (1.000)	9.5811 (1.000)	1.2192 (1.000)
$\rho = 0.1$	3.6446 (0.997)	9.4030 (0.995)	1.2057 (0.996)
$\rho = 0.2$	3.6486 (0.994)	8.9695 (0.987)	1.1663 (0.992)
$\rho = 0.3$	3.5972 (0.996)	8.3081 (0.986)	1.1026 (0.991)
$\rho = 0.4$	3.4989 (1.000)	7.4457 (0.987)	1.0156 (0.991)
$\rho = 0.5$	3.3621 (1.007)	6.4089 (0.989)	0.9059 (0.993)
$\rho = 0.6$	3.1948 (1.013)	5.2226 (0.988)	0.7728 (0.994)
$\rho = 0.7$	3.0034 (1.013)	3.9060 (0.969)	0.6152 (0.990)

NOTE: Range of ρ is $(-0.1306, 0.7941)$. The parameter values $\beta_0 = 1.2$, $\beta_1 = 0.34$, $\beta_2 = -0.15$, and $n = 500$; efficiencies are given in parentheses.

uniform on (0,1). The second covariate x_{it}^D to taken to be a discrete covariate taking values $\{-1.5, -0.5, 0.5, 1.5\}$. We fixed the regression parameters as $\beta_0 = 0.8$, $\beta_1 = -0.1$, $\beta_2 = 0.15$. Two simulations were done excluding and including the second covariate x_{it}^D. Table 2.4 shows the results of simulation with regression parameter are intercept and x_{it}^C. The ARE is one when $\rho = 0$ as before. In this case we can see that multivariate probit model is more efficient for estimating β_0 while the first order Markov chain model is more efficient when estimating β_1 for all values of ρ. In the second simulation, we added the discrete covariate x_{it}^D to the model. From Table 2.5, we can see that the Markov chain model is more efficient for estimating both β_1 and β_2. Further the ARE is steadily decreasing as ρ increases.

When we increase the repeated measurements to five, in both simulations excluding and including the discrete covariate, the first order Markov chain model seems to be doing better than the multivariate probit model. The results are shown in Table 2.6 and Table 2.7. Please note that the feasible range of ρ becomes narrow as the number of repeated measurements increases.

In conclusion the asymptotic relative efficiencies are in the range 0.90 and 1.02 and thus both the methods are good for modeling serially correlated repeated binary measurements. But the first order Markov chain model seems to have a slight edge

Table 2.4: Asymptotic variance and ARE with one covariate, $t = 4$

	$nV(\beta_0)$			$nV(\beta_1)$		
	MC	MPROBIT	ARE	MC	MPROBIT	ARE
$\rho = 0.0$	0.5010	0.5010	1.000	0.4015	0.4015	1.000
$\rho = 0.1$	0.5792	0.5801	1.001	0.4176	0.4160	0.996
$\rho = 0.2$	0.6649	0.6679	1.004	0.4252	0.4213	0.991
$\rho = 0.3$	0.7560	0.7661	1.008	0.4233	0.4180	0.988
$\rho = 0.4$	0.8665	0.8766	1.011	0.4105	0.4053	0.987
$\rho = 0.5$	0.9873	1.0019	1.015	0.3853	0.3813	0.990
$\rho = 0.6$	1.1258	1.1454	1.017	0.3460	0.3438	0.994
$\rho = 0.7$	1.2870	1.3113	1.019	0.2900	0.2901	0.994
$\rho = 0.8$	1.1258	1.1454	1.017	0.3460	0.3438	0.994

NOTE: Range of ρ is $(-0.2255, 0.9152)$. The parameter values are $\beta_0 = 0.8$, $\beta_1 = -0.1$ and $n = 500$.

over the multivariate probit model for estimating the regression slopes.

II.9 Comparison of small-sample performance

To evaluate the small-sample performance, we chose the continuous covariate x_{it}^C to be standard normal and the discrete covariate x_{it}^D same as in the large sample simulations. The sample size was fixed as $n = 30$. For the number of repeated measurements t we took 3,4 and 5. For each combination of the parameters we simulated 1000 samples and for each sample we estimated the regression parameters. We then calculated the average squared deviation of the estimated parameter value from the true population values, i.e. the mean square error (MSE). The small sample efficiencies are calculated by taking the ratio of the MSE for the two estimating procedures as

$$\text{RE} = \frac{\text{MSE of the Markov chain model estimator}}{\text{MSE of the multivariate probit model estimator}}.$$

The relative efficiencies, when $t = 3$, for the three regression parameters are plotted in Figure 2.2, Figure 2.3 and Figure 2.4. We can see from Figure 2.2, the relative efficiency for the intercept is greater than 1 and is increasing as ρ increases.

Table 2.5: Asymptotic variance and ARE with two covariates, $t = 4$

	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.0$	0.5051 (1.000)	0.4013 (1.000)	1.5253 (1.000)
$\rho = 0.1$	0.5837 (1.001)	0.4174 (0.996)	1.4937 (0.997)
$\rho = 0.2$	0.6695 (1.004)	0.4249 (0.991)	1.4173 (0.992)
$\rho = 0.3$	0.7647 (1.008)	0.4230 (0.987)	1.3013 (0.990)
$\rho = 0.4$	0.8713 (1.011)	0.4101 (0.986)	1.1513 (0.990)
$\rho = 0.5$	0.9920 (1.015)	0.3849 (0.985)	0.9728 (0.990)
$\rho = 0.6$	1.1303 (1.017)	0.3455 (0.980)	0.7693 (0.983)
$\rho = 0.7$	1.1412 (1.020)	0.3274 (0.980)	0.5787 (0.981)

NOTE: Range of ρ is $(-0.1726, 0.7105)$. The parameter values are $\beta_0 = 0.8$, $\beta_1 = -0.1$, $\beta_2 = 0.15$, and $n = 500$; ARE are given in parentheses.

Table 2.6: Asymptotic variance and ARE with one covariate, $t = 5$

	$nV(\beta_0)$			$nV(\beta_1)$		
	MC	MPROBIT	ARE	MC	MPROBIT	ARE
$\rho = 0.0$	1.3162	1.3162	1.000	0.6217	0.6217	1.000
$\rho = 0.1$	1.5272	1.5054	0.986	0.6674	0.6541	0.980
$\rho = 0.2$	1.7506	1.7037	0.973	0.7004	0.6729	0.961
$\rho = 0.3$	1.9969	1.9264	0.965	0.7205	0.6802	0.944
$\rho = 0.4$	2.2754	2.1613	0.950	0.7258	0.6584	0.907

NOTE: Range of ρ is $(-0.0080, 0.4373)$. The parameter values are, $\beta_0 = 1.5$, $\beta_1 = -0.6$ and $n = 500$.

Table 2.7: Asymptotic variance and ARE with two covariates, $t = 5$

	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.0$	3.1740 (1.000)	0.5861 (1.000)	7.6772 (1.000)
$\rho = 0.1$	3.3390 (0.989)	0.6296 (0.981)	7.5164 (0.992)
$\rho = 0.2$	3.4568 (0.975)	0.6612 (0.962)	7.1167 (0.979)
$\rho = 0.3$	3.5437 (0.953)	0.6806 (0.945)	6.5038 (0.947)

NOTE: Range of ρ is $(-0.0080, 0.3890)$. The parameter values are $\beta_0 = 1.5$, $\beta_1 = -0.6$, $\beta_2 = -0.1$, and $n = 500$; ARE are given in parentheses.

This shows that the multivariate probit model is estimating the intercept better than the Markov chain model in small samples, especially when there is a high correlation. Figure 2.3 shows the relative efficiency for the slope $\beta 1$ of the continuous covariate. The plot is above one except in an small interval of ρ, indicating that even in this case the multivariate probit estimator is outperforming the Markov chain model estimator. Note that the figure shows a roughly decreasing trend in the efficiencies. Unlike the previous two coefficients, the relative efficiency for β_2 is less than 1 and is decreasing as the correlation increases, indicating that the Markov chain model is better at estimating the slope for the discrete covariate than the multivariate probit model. The simulated MSE values and the relative efficiencies for different values of ρ are presented in Table 2.8.

We now consider the case $t = 4$. Figure 2.5 has the plot of the relative efficiency for the intercept β_0. The efficiency is more than for almost every value of ρ and is increasing as ρ increases, which shows the multivariate probit model is better than the Markov chain model for estimating β_0. The efficiency plots for β_1 and β_2 are in Figures 2.6, 2.7 respectively. These two plots show that for highly correlated data the Markov chain model is better than multivariate probit model for estimating the slopes. Table 2.9 contains the MSE's and the efficiencies for the three regression coefficients in the case $t = 4$.

Lastly, Figures 2.8, 2.9 and 2.10, contains the plots of the efficiencies for the three regression coefficients when $t = 5$. For all the three regression coefficients the relative

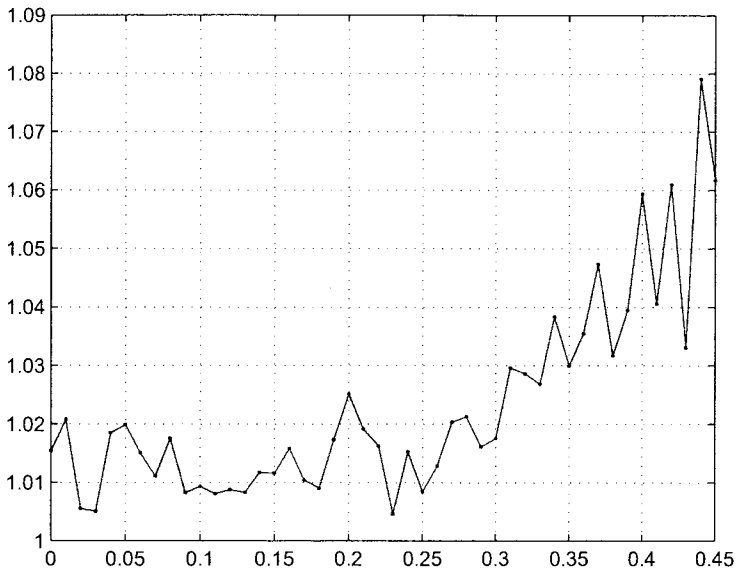


Figure 2.2: Plot of efficiency of β_0

efficiency is flat at one for small and moderate values of ρ. However, for the intercept β_0 and slope β_1 the relative efficiency is more than one for values close to the upper boundary value of ρ, whereas it is less than one for β_2. Thus for moderate correlation both methods perform equally well and first order multivariate probit model is more efficient for estimating β_0 and β_1 for highly correlated data, whereas the Markov chain outperforms estimating β_2 for highly correlated data. Table 2.10 contains the MSE's and the efficiencies for the three regression coefficients in the case $t = 5$.

To summarize, in the small sample case we see that both the models are equally good for small and moderately correlated data especially if the number of repeated measurements is about five. For highly correlated data, the multivariate probit model is a good choice if the covariates are continuous and the Markov chain model is better for discrete covariates.

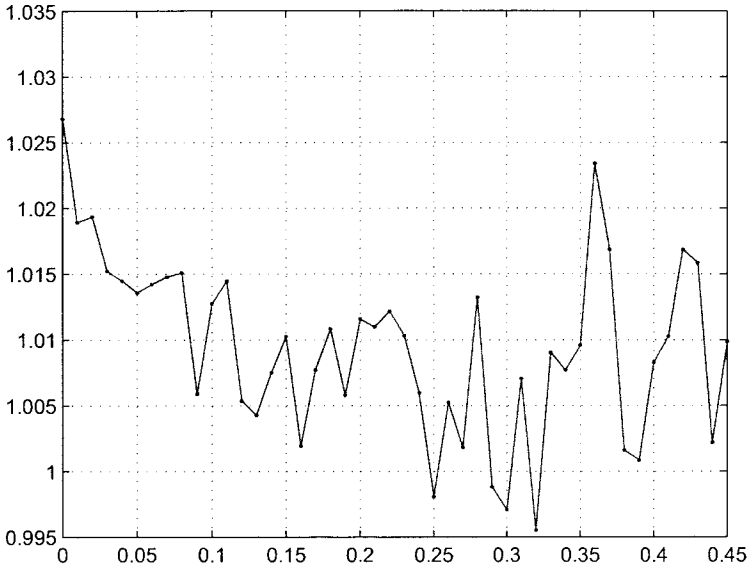


Figure 2.3: Plot of efficiency of β_1

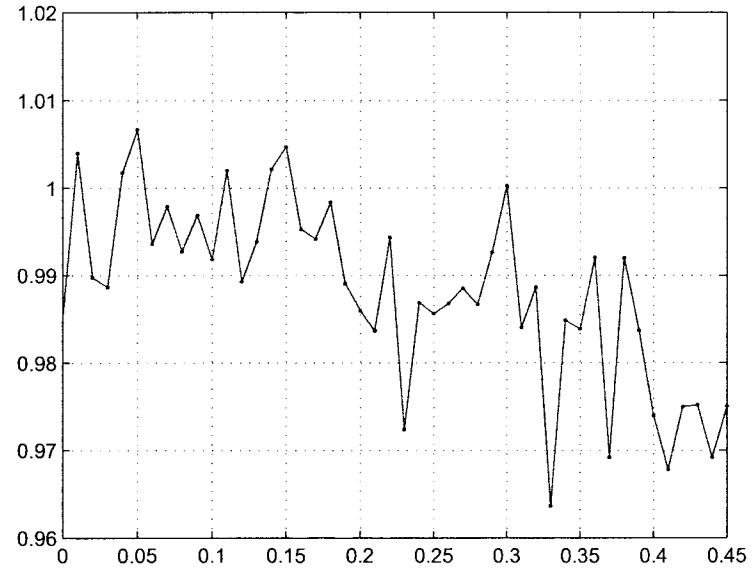


Figure 2.4: Plot of efficiency of β_2

Table 2.8: MSE and efficiencies of MC and MP estimate, $t = 3$

	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.00$	0.0247 (1.015)	0.0357 (1.027)	0.0348 (0.986)
$\rho = 0.05$	0.0285 (1.019)	0.0323 (1.014)	0.0357 (1.007)
$\rho = 0.10$	0.0274 (1.009)	0.0352 (1.013)	0.0349 (0.992)
$\rho = 0.15$	0.0310 (1.011)	0.0342 (1.010)	0.0294 (1.005)
$\rho = 0.20$	0.0358 (1.025)	0.0343 (1.012)	0.0320 (0.986)
$\rho = 0.25$	0.0360 (1.008)	0.0308 (0.985)	0.0280 (0.986)
$\rho = 0.30$	0.0334 (1.017)	0.0325 (0.997)	0.0280 (1.000)
$\rho = 0.35$	0.0375 (1.030)	0.0301 (1.010)	0.0279 (0.984)
$\rho = 0.40$	0.0418 (1.059)	0.0324 (1.008)	0.0276 (0.974)
$\rho = 0.45$	0.0478 (1.062)	0.0283 (1.010)	0.0250 (0.975)

NOTE: Range of ρ is $(-0.207, 0.4849)$. The parameter values are $\beta_0 = 0.8$, $\beta_1 = -0.1$.

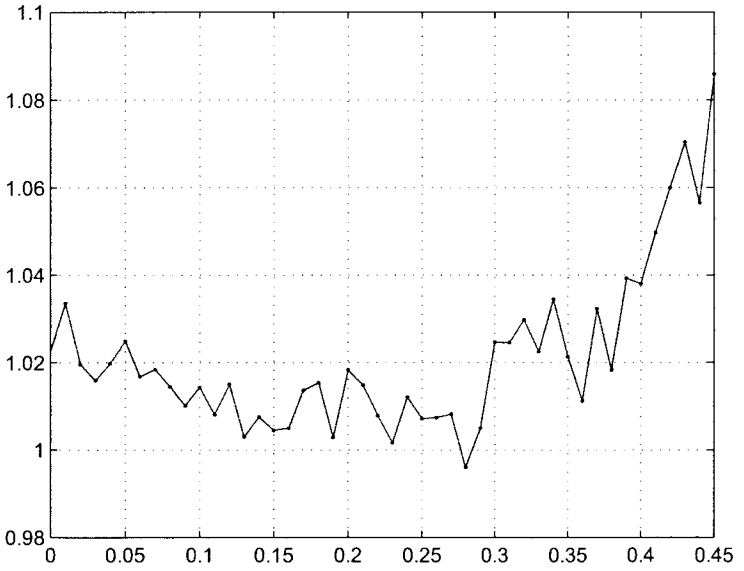


Figure 2.5: Plot of efficiency of β_0

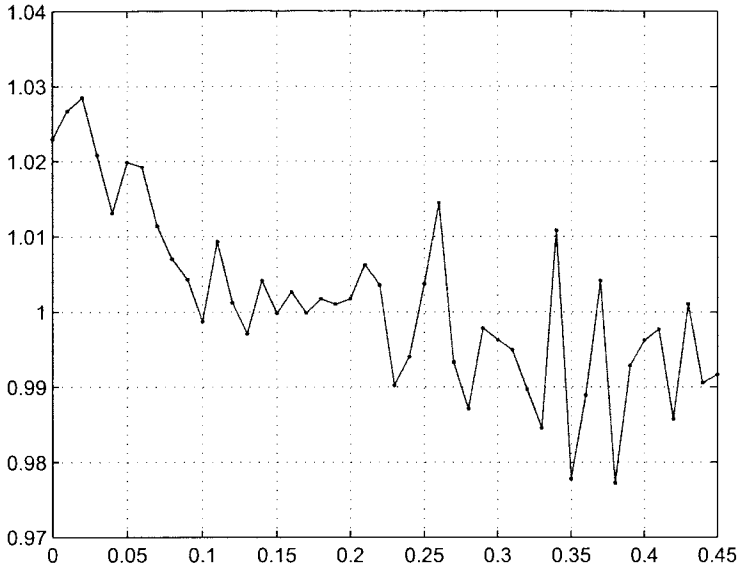


Figure 2.6: Plot of efficiency of β_1

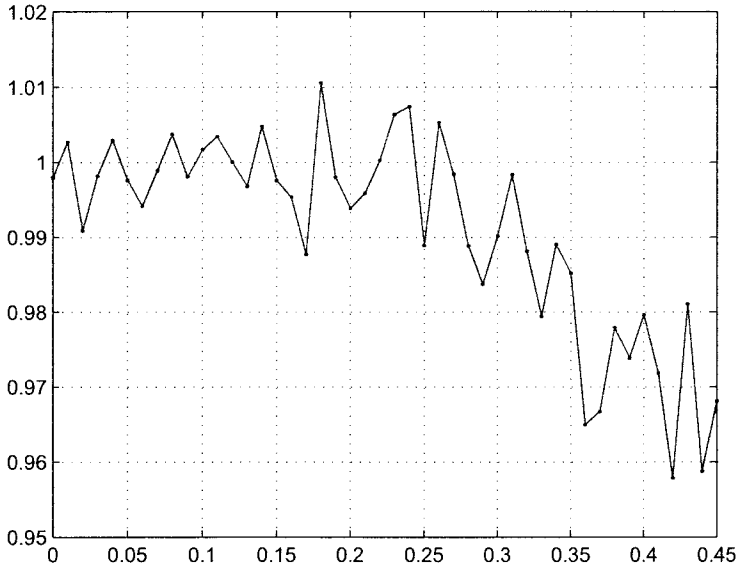


Figure 2.7: Plot of efficiency of β_2

Table 2.9: MSE and efficiencies of MC and MP estimate, $t = 4$

Method	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.00$	0.0180(1.023)	0.0147(1.023)	0.0206(0.998)
$\rho = 0.05$	0.0193(1.025)	0.0152(1.020)	0.0203(0.998)
$\rho = 0.10$	0.0210(1.014)	0.0151(0.999)	0.0191(1.002)
$\rho = 0.15$	0.0217(1.004)	0.0161(1.000)	0.0200(0.998)
$\rho = 0.20$	0.0257(1.018)	0.0172(1.002)	0.0199(0.994)
$\rho = 0.25$	0.0224(1.007)	0.0153(1.004)	0.0173(0.989)
$\rho = 0.30$	0.0281(1.025)	0.0157(0.996)	0.0188(0.990)
$\rho = 0.35$	0.0299(1.021)	0.0140(0.978)	0.0177(0.985)
$\rho = 0.40$	0.0320(1.038)	0.0150(0.996)	0.0162(0.980)
$\rho = 0.45$	0.0416(1.085)	0.0148(0.992)	0.0147(0.968)

NOTE: Range of ρ is $(-0.121, 0.498)$. The parameter values are $\beta_0 = 0.8$, $\beta_1 = -0.1$.

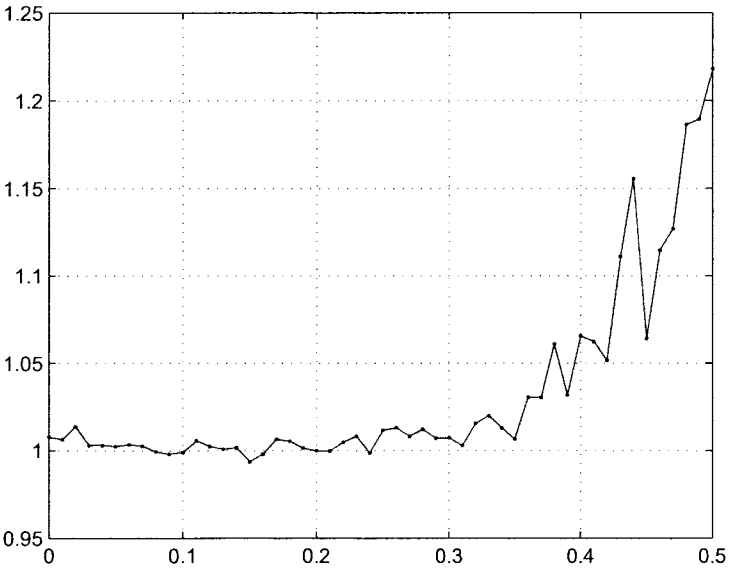


Figure 2.8: Plot of efficiency of β_0

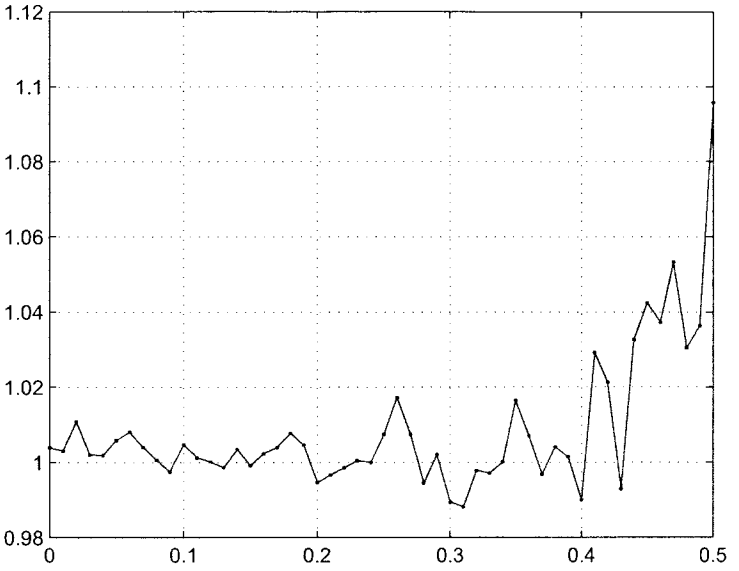


Figure 2.9: Plot of efficiency of β_1

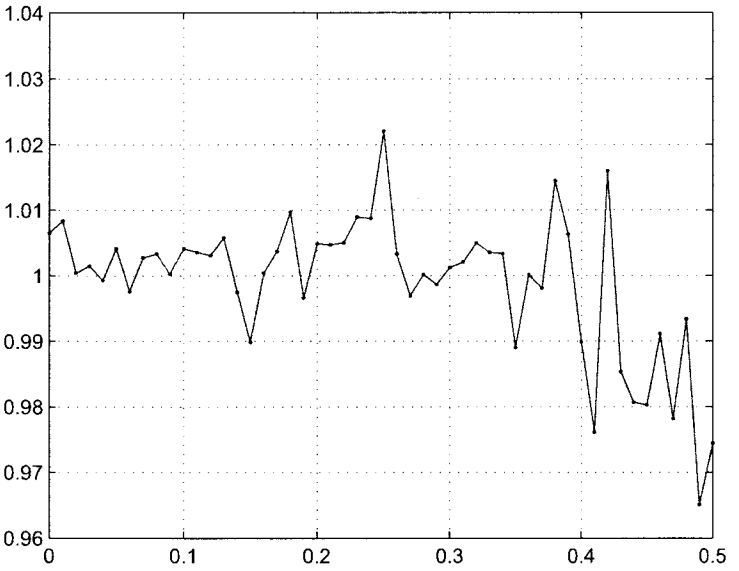


Figure 2.10: Plot of efficiency of β_2

Table 2.10: MSE and efficiencies of MC and MP estimate, $t = 5$

Method	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.00$	0.0119(1.008)	0.0056(1.004)	0.0106(1.006)
$\rho = 0.05$	0.0122(1.002)	0.0065(1.006)	0.1201(1.004)
$\rho = 0.10$	0.0135(0.999)	0.0061(1.005)	0.0109(1.004)
$\rho = 0.15$	0.0149(0.994)	0.0064(0.999)	0.0113(0.990)
$\rho = 0.20$	0.0158(1.000)	0.0068(0.995)	0.0103(1.005)
$\rho = 0.25$	0.0179(1.012)	0.0070(1.008)	0.0111(1.022)
$\rho = 0.30$	0.0186(1.007)	0.0065(0.989)	0.0098(1.001)
$\rho = 0.35$	0.0198(1.007)	0.0070(1.016)	0.0093(0.989)
$\rho = 0.40$	0.0226(1.065)	0.0067(0.990)	0.0084(0.990)
$\rho = 0.45$	0.0230(1.064)	0.0074(1.042)	0.0079(0.980)
$\rho = 0.50$	0.0302(1.218)	0.0072(1.096)	0.0080(0.974)

NOTE: Range of ρ is $(-0.313, 0.532)$. The parameter values is $\beta_0 = 0.8$, $\beta_1 = -0.1$.

II.9.1 Analysis of real data

To illustrate the application of the two likelihood estimation methods, in this section we present the analysis of the three binary longitudinal data that we discussed in Section I.2.

Example 2.1. Six city data.

For this data the main issue of interest is the effect of maternal smoking on children's respiratory illness. We fit the following regression model to the data:

$$\Phi^{-1}(p_{it}) = \mu_{it} = \beta_0 + \beta_1 x_{it}^{\text{Age}} + \beta_2 x_{it}^{\text{MS}} + \beta_3 x_{it}^{\text{Age}} x_{it}^{\text{MS}},$$

where $i = 1, 2, \cdots, 537$ and $t = 1, 2, 3, 4$. The covariates are the age of the child (x_{it}^{Age}), the maternal smoking habit indicator (x_{it}^{MS}) and their interaction.

Table 2.11: Parameter estimates for the Six-city data

	MC			MPROBIT		
PARAMETER	EST	SE	PVALUE	EST	SE	PVALUE
Intercept	-1.1366	0.0585	0.0000	-1.1368	0.0601	0.0000
Age	-0.0829	0.0699	0.0281	-0.0816	0.0362	0.0282
Smoking	0.1599	0.1708	0.0859	0.1598	0.0972	0.0968
Age×Smoking	0.0453	0.0620	0.4650	0.0438	0.7061	0.4801
ρ	0.3836	0.0313	0.0000	0.6842*	0.0387*	0.0000

NOTE: Range of ρ is (-0.1357, 0.9267).

Table 2.11 provides point estimates, standard errors and p-values for both the first order Markov chain and the multivariate probit model. The estimates of the regression parameters are very similar for the two models. Note that in the table, the value 0.6842 for multivariate probit model is not the correlation parameter for the binary repeated measurements, it is the value of latent correlation parameter. The p-values indicate that the child's age is a significant factor and the regression coefficient is negative, which means that older children are less likely to get respiratory disease. The main covariate of interest, maternal smoking, is not significant in both the models, even though estimate of the regression coefficient in both models is positive, which indicates that the children are more likely to develop respiratory disease if their mother was a smoker as opposed to a nonsmoker.

The values of model selection criteria, AIC and BIC, based on Markov chain model with various combinations of the covariates are given in Table 2.12. The model with only age as the covariate and the one with both age and smoke as covariates have similar AIC values, but the former model has the smallest BIC value, so it is the best model for this data set.

Table 2.12: Model selection based on AIC and BIC

Covariates	**LF**	**df**	**AIC**	**BIC**
Intercept, Age	-815.49	2	1632.97	1637.26
Intercept, Smoke	-816.70	2	1635.39	1639.68
Intercept, Age, Smoke	-814.31	1	1632.62	1641.19
Intercept, Age, Smoke, Interaction	-814.01	0	1634.02	1646.88

Example 2.2. Indonesian children's health study

The second longitudinal binary data that we analyzed is the Indonesian children's health study data. Table 1.3 displays a subset of this data originally studied by Sommer et al. (1984). In this study preschool children were examined up to six consecutive quarters for the presence of respiratory infection. There were 1,200 observations in total. The covariates of interest include: (1) age in months (centered at 36); (2) presence/absence of xerophthalmia, an ocular manifestation of chronic vitamin A deficiency; (3) cosine and sine terms for the annual cycle; (4) gender; (5) height for age, as a percent of the National Center for Health Statistics (NCHS) standard (centered at 90%); and (6) presence of stunting, defined as being below 85% in height for age. The key covariate of interest is the indicator of xerophthalmia symptom. The primary objective of this study was to assess the increase in risk of respiratory infection for kids who were vitamin A deficient, which was measured indirectly via xerophthalmia.

The parameter estimates, standard errors and p-values for the two likelihood methods, Markov chain and multivariate probit, are given in Table 2.13. The estimates and standard errors are very similar for both the methods. The results indicate that respiratory infection is strongly related to age and season. Further, the risk of infection decreases approximately 4% per month for children between one and five years. The xerophthalmia coefficient is about 0.33, but the p-value is approximately 0.2, hence xerophthalmia is not statistically significant factor.

Table 2.13: Parameter estimates for Indonesian children data

	MC			MPROBIT		
PARAMETER	EST	SE	PVALUE	EST	SE	PVALUE
Intercept	-1.3877	0.0870	0.0000	-1.3914	0.0892	0.0000
Age	-0.0159	0.0032	0.0000	-0.0167	0.0034	0.0000
Xerophthalmi	0.3280	0.2411	0.1736	0.3087	0.0972	0.1997
Seasonal cosine	-0.2720	0.0808	0.0007	-0.2740	0.0807	0.0007
Seasonal sine	-0.0827	0.0843	0.3266	-0.0873	0.0828	0.2920
Sex	-0.1897	0.1143	0.0970	-0.1899	0.1178	0.1068
Height	-0.0190	0.0122	0.1184	-0.0200	0.0126	0.1120
Stunted	0.1016	0.2040	0.6186	0.1203	0.2094	0.5657
ρ	0.0583	0.0446	0.1915	0.2029*	0.1066*	0.0570*

NOTE: Range of ρ is (-0.1357, 0.9267).

Example 2.3. Hamilton's depression study

The third longitudinal binary data we analyze is the Hamilton's depression study data. This data were taken from a randomized double-blind, placebo and active treatment, 7 week study in Europe to evaluate safety and efficacy for three fixed doses of a new drug in patients with major depressive disorder. For our analysis, we will study the active treatment and the placebo groups. The dependent variable in this study is a binary function of a patient's average score on the Hamilton's Depression Scale (Ham-D), taking the value 1 if the Ham-D value at time $i = 1, \cdots, 8$ is less than or equal to 80% of the baseline value, and 0 otherwise. The covariates used in this study are the treatment, time, gender, age, and baseline measurement and the country where the patient resides.

The results indicate that Treatment, Time, Country and Baseline are significant covariates to predict effectiveness of the treatment. The coefficient of Treatment is positive, which means that the drug is effective at reducing depression levels. The results are shown in Table 2.14.

Table 2.14: Parameter estimates for Ham-D data

	MC		
PARAMETER	EST	SE	PVALUE
Intercept	-3.1459	0.2691	<0.0001
Treatment	0.3309	0.1048	0.0016
Time	0.5610	0.0257	<0.0001
Age	-0.0322	0.0512	0.5293
Gender	-0.1654	0.1116	0.1386
Country	0.0663	0.0239	0.0055
Baseline	0.1296	0.0438	0.0030
ρ	0.3923	0.0304	<0.0001

NOTE: Range of ρ is (-0.0093, 0.3923).

II.10 Other models for correlated binary variables

In this section, we discuss other methods for generating dependent binary variables.

II.10.1 Generating dependent binary variables using circle plus operation

Let $\Omega = \{0, 1\}$. Suppose $x, y \in \Omega$. Define the circle plus operation as

$$x \oplus y = \begin{cases} x + y & \text{if } 0 \leq x + y \leq 1 \\ x + y - 1 & \text{if } x + y > 1. \end{cases} \tag{2.10.1}$$

Note $x \oplus y \in \Omega$.

Lemma 2.1. *Suppose X_i's are independent and distributed as* Binomial$(1, p_i)$, $1 \leq i \leq t$. *Let $Y_1 = X_1$ and $Y_j = Y_{j-1} \oplus X_j$, for $2 \leq j \leq t$. Then*

$$P(Y_k = 1 | Y_j = 1) = P(Y_k = 1 | Y_{k-1} = 1) = 1.$$

Proof. By definition of the circle plus operation, we have

$$Y_i = \begin{cases} 0 & \text{if } (Y_{i-1}, X_i) = (0,0) \\ 1 & \text{otherwise.} \end{cases}$$

It is clear that $P(Y_i = 1 | Y_{i-1} = 1) = 1$. Therefore

$$\begin{aligned} P(Y_k = 1 | Y_j = 1) &= P(Y_k = 1, Y_{j+1} = 1 | Y_j = 1) + P(Y_k = 1, Y_{j+1} = 0 | Y_j = 1) \\ &= P(Y_k = 1, Y_{j+1} = 1 | Y_j = 1) \\ &= P(Y_k = 1 | Y_{j+1} = 1, Y_j = 1) P(Y_{j+1} = 1 | Y_j = 1) \\ &= P(Y_k = 1 | Y_{j+1} = 1, Y_j = 1) \\ &= P(Y_k = 1, Y_{j+2} = 1 | Y_{j+1}, Y_j = 1) + P(Y_k = 1, Y_{j+2} = 0 | Y_{j+1}, Y_j = 1) \\ &= \cdots \\ &= P(Y_k = 1 | Y_{k-1} = 1, \cdots, Y_j = 1) \\ &= P(Y_k = 1 | Y_{k-1} = 1) \\ &= 1. \end{aligned}$$

This completes the proof of the lemma. □

Theorem 2.1. *Suppose X_i's are independent and distributed as* Binomial$(1, p_i)$, $1 \leq i \leq t$. *Let $Y_1 = X_1$ and $Y_j = Y_{j-1} \oplus X_j$, for $2 \leq j \leq t$. Then*

1. $P(Y_j = 1) = p_j^* = 1 - \prod_{i=1}^{j} q_i$, *where* $q_i = 1 - p_i$.
2. $P(Y_j = 1, Y_k = 1) = p_j^*$.
3. $\text{Corr}(Y_j, Y_k) = \rho_{jk} = \sqrt{\frac{p_j^* q_k^*}{q_j^* p_k^*}}$, *for* $1 \leq j < k \leq t$.

Proof. The first assertion is proved by induction.

For $i = 1$, we have $P(Y_1 = 1) = P(X_1 = 1) = p_1 = 1 - q_1$.

Assume that for $i = j - 1$

$$P(Y_{j-1} = 1) = 1 - \prod_{i=1}^{j-1} q_i.$$

Note that

$$\begin{aligned} P(Y_j = 1) &= 1 - P(Y_j = 0) \\ &= 1 - P(Y_{j-1} = 0, X_j = 0). \end{aligned}$$

Since X_i's are independent and Y_{j-1} depends on X_{j-1}, we have Y_{j-1} is independent of X_j. Therefore

$$\begin{aligned} P(Y_j = 1) &= 1 - P(X_j = 0)P(Y_{j-1} = 0) \\ &= 1 - q_j\, P(Y_{j-1} = 0). \end{aligned}$$

By the induction hypothesis we get

$$P(Y_j = 1) = 1 - \prod_{i=1}^{j} q_i.$$

This proves the first part of the theorem.

To prove the second part, let $j < k$. By Lemma 2.1, we have

$$P(Y_k = 1, Y_j = 1) = P(Y_k = 1 | Y_j = 1)P(Y_j = 1) = P(Y_j = 1) = p_j^* = 1 - \prod_{i=1}^{j} q_i.$$

Since $P(Y_j = 1) = p_j^*$ we have

$$\begin{aligned} Corr(Y_j, Y_k) = \rho_{jk} &= \frac{P(Y_k = 1, Y_j = 1) - P(Y_j = 1)P(Y_k = 1)}{\sqrt{\mathrm{Var}(Y_j)\mathrm{Var}(Y_k)}} \\ &= \frac{p_j^* - p_j^* p_k^*}{\sqrt{p_j^* q_j^* p_k^* q_k^*}} \\ &= \sqrt{\frac{p_j^* q_k^*}{q_j^* p_k^*}}. \end{aligned} \tag{2.10.2}$$

This completes the proof of the theorem. □

Note that in Theorem 2.1, the correlation between Y_j and Y_k is the maximum possible correlation with marginals fixed as p_j^* and p_k^*. Thus given $p_1, \cdots, p_t$, we can generate a dependent binary random vector $Y = (Y_1, \ldots, Y_k)$ such that the correlation between Y_j and Y_k is the maximum possible with marginals fixed as $p_j^* = 1 - \prod_{i=1}^{j} q_i$ and $p_k^* = 1 - \prod_{i=1}^{k} q_i$. Since $p_k^* = 1 - q_k q_{k-1}^*$, we have $q_k^* = q_k q_{k-1}^*$ and therefore $q_k^* \leq q_{k-1}^*$,

i.e. $p_k^* \geq p_{k-1}^*$. Thus the circle plus operation generates binary variables with the ordering $P(Y_1 = 1) \leq \cdots \leq P(Y_t = 1)$, and the maximum possible correlation between any pair.

A special case of this model is when we have $p_1 = \cdots = p_t = p$. In this case,

$$P(Y_j = 1) = 1 - q^j$$

and

$$Corr(Y_j, Y_k) = \left(\frac{1-q^j}{1-q^k}\right) q^{|k-j|}.$$

This is similar to the autoregressive structure of order one.

II.10.2 Quadratic exponential family

An alternative model for correlated binary variables is the quadratic exponential family of distributions. The probability mass function for this family is of the form

$$\pi(y;\lambda) = \exp(\lambda_1 y_1 + \cdots + \lambda_t y_t + \lambda_{12} y_1 y_2 + \cdots + \lambda_{t-1,t} y_{t-1} y_t)/k(\lambda), \quad (2.10.3)$$

with $y \in \{0,1\}^t$, $\lambda = (\lambda_1, \cdots, \lambda_t, \lambda_{12}, \cdots, \lambda_{t-1,t})$ and $k(\lambda)$ is the normalizing constant. This probability mass function maximizes the entropy function

$$-\sum_y \pi(y)\, \log(\pi(y))$$

in the class of multivariate binary distributions $\pi(y)$ which satisfy the moment constraints $\sum_y y_i \pi(y) = p_i$, for $i = 1, \cdots, t$, and $\sum_y y_i y_j \pi(y) = p_{ij}$, for $1 \leq i \leq j \leq t$ (see Kapur and Kesavan (1992), Chaganty and Joe (2006)).

Given a vector $\mu = (p_1, \cdots, p_t, p_{12}, \cdots, p_{t-1,t})'$ that specifies the univariate and bivariate marginals, if there is a multivariate binary distribution then the distribution that maximizes the entropy, can be obtained solving the equation

$$\sum_y \pi(y;\lambda) s = \mu, \quad (2.10.4)$$

where $s = (y_1, \cdots, y_t, y_{12}, \cdots, y_{t-1,t})'$, and λ is a vector of Lagrange multipliers. The above summation is taken over 2^t possible binary vectors y.

For a given $p = (p_1, \cdots, p_t)'$, we have checked numerically the quadratic exponential model generates the same binary distribution as the Markov chain model if the correlation structure is AR(1) with parameter ρ satisfying the constraints

$$\max_{2 \le i \le t} L(p_{i-1}, p_i) \le \rho \le \min_{2 \le i \le t} U(p_{i-1}, p_i). \tag{2.10.5}$$

CHAPTER III

MARGINAL MODELS

III.1 Introduction

In Chapter II we have studied multivariate likelihoods for modeling and analysing longitudinal and clustered data. We should use likelihood methods whenever possible, because they have strong theoretical underpinnings and maximum likelihood estimation is the optimal estimation procedure. The marginal models are an alternative in situations where the likelihood methods are difficult to construct and prohibitively difficult to implement. Motivated by quasi-likelihood methods, Liang and Zeger (1986) have introduced marginal models and generalized estimating equations (GEE) for estimating the regression coefficients primarily for non-normal repeated observations. The marginal models avoid specification of the joint distribution but model the dependence between the repeated measurements using a working correlation structure, that may not be the true correlation structure. However, under misspecification of the correlation structure the estimator of the regression parameter can be inefficient. Qu et al. (2000) introduced a new method, known as quadratic inference functions (QIF), that does not involve direct estimation of the correlation parameter but produces highly efficient estimate of the regression parameter even if the working correlation structure is misspecified. Instead of estimating the nuisance parameter α in the working correlation matrix, their method models the inverse of the working correlation matrix as a linear combination of a class of basis matrices. This gives a sufficiently rich class that accommodates or at least approximates the correlation structures most commonly used. Qu et al. (2000) showed that even if the correlation is misspecified, their method remains optimal within the assumed family, and hence more efficient than Liang and Zeger's GEE regression estimator under misspecification.

The outline of this chapter is as follows. In Section III.2 we briefly describe the generalized estimating equations method. In the next Section III.3 we introduce quadratic inference functions method, establish asymptotic variances for the estimates, then study their asymptotic relative efficiency. In Section III.6 we derive

a new estimation for the covariance matrix of the score function and Section III.5 contains the small-sample efficiency comparisons.

III.2 Generalized estimating equation

In this section, we give a brief introduction to the commonly used estimation method to analyze longitudinal and clustered data, known as the generalized estimating equations (GEE) proposed by Liang and Zeger (1986). The method can be regarded as an extension to the multivariate case the quasi-likelihood methods described in the book McCullagh and Nelder (1989).

The framework of the generalized estimating equations is as follows. Let $Y_i = (y_{i1}, \ldots, y_{it})$ be a vector of correlated observations on the ith subject. Assume that $\mathrm{E}(y_{ij}) = \mu_{ij} = h(x'_{ij}\beta)$ and $Var(y_{ij}) = v_{ij} = v(\mu_{ij})$, where β is a k-dimensional regression parameter. The mean and variance functions $h(\cdot)$ and $v(\cdot)$ are assumed to be known. Suppose that $\mathrm{Cov}(Y_i) = \phi W_i(\beta, \alpha)$, where α is an nuisance parameter. The GEE framework assumes that $W_i(\beta, \alpha) = A_i^{1/2} R(\alpha) A_i^{1/2}$, where $A_i = diag(v_{ij})$ is the diagonal matrix of variances that depends on β, and $R(\alpha)$ is a working correlation matrix determined by α. The generalized estimating equation is given by

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' W_i^{-1}(y_i - \mu_i) \;=\; 0, \tag{3.2.1}$$

where $W_i = W_i(\beta, \alpha)$. Equation (3.2.1) is an unbiased estimating equation for fixed α. Under some regularity conditions, the estimate of β obtained solving equation (3.2.1) is consistent with mean equal to β and covariance matrix

$$Cov(\hat{\beta}_{GEE}) = \left(\sum_{i=1}^{n} D_i' W_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i' W_i^{-1} \Sigma_i W_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i' W_i^{-1} D_i\right)^{-1} \tag{3.2.2}$$

where $D_i = \partial \mu_i / \partial \beta$. Clearly, if the working correlation matrix $R(\alpha)$ equals the true correlation structure, then $W_i = \Sigma_i$ and equation (3.2.1) is the optimal unbiased estimating equation. In this case (3.2.2) reduces to

$$\mathrm{Cov}(\hat{\beta}_{opt}) \;=\; \left(\sum_{i=1}^{n} D_i' \Sigma_i^{-1} D_i\right)^{-1}. \tag{3.2.3}$$

GEE method has been very popular for analyzing longitudinal data because it is computational less demanding than the fully specified model. However, GEE method has several theoretical shortcomings due to some critical underlying assumptions. The working correlation, when misspecified, lacks a proper definition and thus causes a breakdown of the asymptotic properties of the estimator. Further, for binary random variables there is no guarantee that the working correlation parameter α will fall within the feasible bounds, that is the value of α may not compatible with the marginal means, that is, a multivariate binary distribution with specified means and correlation structure may not exist. This could course a series problem because the resulting standard errors and p-values could lead to misleading conclusions. Sabo and Chaganty (2010) gave examples where an infeasible correlation estimate leads to wrong conclusions. To rectify the problem Sabo and Chaganty (2010) suggested to run GEE with working independent structure, compute the correlation bounds using the estimated marginal means, and then rerun GEE with a correlation value selected within those bounds using some objective criteria. Alternatively, they suggest use of likelihood methods, such as the multivariate probit model, to avoid these infeasibility problems altogether. In the next section we study efficiency of GEE as compared to the Markov chain likelihood procedure.

III.2.1 Comparison of asymptotic performance

The asymptotic relative efficiency comparisons between the multivariate probit model and GEE were done by Chaganty and Joe (2004), where they showed that the multivariate probit model is uniformly superior to the GEE method. In this section, we will study the efficiency of GEE estimates with respect to the maximum likelihood estimator based on the first order Markov chain likelihood model.

For comparisons of efficiency, we used the marginal mean model for the binary observations:

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 x_{it}^D + \beta_2 x_{it}^C, \tag{3.2.4}$$

where x_{it}^D is a discrete covariate taking values $1, 2, \cdots, t$ and x_{it}^C is continuous distributed as standard normal. We fixed the values of regression coefficients as $\beta_0 = 1$, $\beta_1 = 0.3$ and $\beta_2 = -0.1$. We chose a large value for n (=1000), so that the calculated

efficiencies will not depend much on the simulated covariates and they approximate the true asymptotic relative efficiency. To cover most practical situations, we consider four cases: (1) small number of repeated measurements with low value of correlation, (2) small number of repeated measurements with high value of correlation, (3) large number of repeated measurements with low value of correlation, and (4) large number of repeated measurements with high value of correlation. Specifically, we choose $t = 4$ and 8 and correlation parameter $\rho = 0.2$ and 0.7 respectively. These values are within the AR(1) correlation bounds (-0.118, 0.780).

Table 3.1 shows the simulation results for the first case $t = 4$ and $\rho = 0.2$. The values of the first row are the diagonal elements of the inverse of the Fisher Information matrix given by (2.3.24). These values are the asymptotic variances of the maximum likelihood estimates of the first order Markov chain model. The second row of the table gives the asymptotic variances of the estimates using the optimal GEE. These values are the diagonal elements of the matrix (3.2.3). The asymptotic relative efficiencies are in parenthesis. We can see that even the optimal GEE is less efficient than the maximum likelihood estimator for the Markov chain model. In the case where GEE uses a working AR(1) structure we present the asymptotic variances computed taking the diagonal elements of the asymptotic covariance matrix (3.2.2) for various parameter values of α in the working correlation matrix. We can see that the relative efficiencies decreases when the value of α gets far from the true value 0.2 of the correlation. The worst case efficiency is when $\alpha = 0.9$, for example, the efficiency of the regression coefficient β_2 is only about 70% of that of maximum likelihood estimator.

Next, we took $\rho = 0.7$ but kept $t = 4$ same as before. The results are displayed in Table 3.2. The GEE estimator with optimal choice of the working covariance matrix, is still less efficient than the estimator of the Markov chain model. The table also has the asymptotic variances and the relative efficiencies for different values of the working correlation α. Interestingly, the GEE with identity correlation ($\alpha = 0$) matrix has the worst efficiency (0.376) for the regression coefficient β_2. Recall that the true correlation is 0.7 which is far from zero. Thus we see that for highly correlated data, GEE with identity structure is very inefficient.

We now consider the third case where $t = 8$ and $\rho = 0.2$. The results of the

Table 3.1: Asymptotic variance and ARE of GEE, $t = 4$

Method	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
Maximum likelihood	10.111 (1.000)	1.366 (1.000)	4.564 (1.000)
Optimal	10.220 (0.989)	1.386 (0.986)	4.6202 (0.988)
$\alpha = 0.0$	10.293 (0.982)	1.395 (0.979)	4.914 (0.929)
$\alpha = 0.1$	10.238 (0.988)	1.388 (0.984)	4.692 (0.973)
$\alpha = 0.2$	10.220 (0.989)	1.386 (0.986)	4.620 (0.988)
$\alpha = 0.3$	10.238 (0.988)	1.388 (0.984)	4.685 (0.974)
$\alpha = 0.4$	10.295 (0.982)	1.394 (0.980)	4.859 (0.939)
$\alpha = 0.5$	10.390 (0.973)	1.402 (0.974)	5.113 (0.892)
$\alpha = 0.6$	10.528 (0.960)	1.412 (0.968)	5.416 (0.843)
$\alpha = 0.7$	10.712 (0.944)	1.422 (0.961)	5.740 (0.795)
$\alpha = 0.8$	10.952 (0.923)	1.432 (0.954)	6.064 (0.753)
$\alpha = 0.9$	11.245 (0.899)	1.439 (0.949)	6.373 (0.716)

NOTE: The true correlation parameter is $\rho = 0.2$. ARE are given in parentheses.

Table 3.2: Asymptotic variance and ARE of GEE, $t = 4$

Method	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
Maximum likelihood	9.136 (1.000)	0.918 (1.000)	1.862 (1.000)
Optimal	9.332 (0.979)	0.955 (0.962)	2.014 (0.924)
$\alpha = 0.0$	9.646 (0.947)	0.979 (0.938)	4.957 (0.376)
$\alpha = 0.1$	9.570 (0.955)	0.971 (0.946)	4.010 (0.464)
$\alpha = 0.2$	9.503 (0.961)	0.965 (0.951)	3.171 (0.569)
$\alpha = 0.3$	9.446 (0.967)	0.961 (0.956)	2.733 (0.681)
$\alpha = 0.4$	9.340 (0.972)	0.958 (0.959)	2.370 (0.785)
$\alpha = 0.5$	9.364 (0.976)	0.956 (0.961)	2.152 (0.865)
$\alpha = 0.6$	9.341 (0.978)	0.955 (0.962)	2.044 (0.985)
$\alpha = 0.7$	9.332 (0.979)	0.955 (0.962)	2.014 (0.924)
$\alpha = 0.8$	9.342 (0.978)	0.955 (0.962)	2.036 (0.914)
$\alpha = 0.9$	9.375 (0.974)	0.955 (0.961)	2.090 (0.891)

NOTE: The true correlation parameter is $\rho = 0.7$. ARE are given in parentheses.

simulations are presented in Table 3.3. The results are similar to the case where $t = 4$. The GEE estimator is less efficient than the maximum likelihood estimator for all values of working correlation α. Table 3.4 shows the simulation results with $t = 8$ with $\rho = 0.7$. The GEE estimator is extremely inefficient especially when working correlation is far from the true correlation or even out of bounds.

In summary, the simulation results show the GEE estimator even with the optimal choice of the working covariance is less efficient than the maximum likelihood estimate obtained from the Markov chain model. Furthermore, the efficiency is very low when the working correlation parameter is far from the true correlation. The worse choices for α near 0 for strong dependence, and large α near 1 for weak dependence. In particular, the choice of $\alpha = 0$ is can lead to very low efficiency if there is strong dependence. The results are similar for $t = 4$ and $t = 8$.

Table 3.3: Asymptotic variance and ARE of GEE, $t = 8$

Method	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
Maximum likelihood	4.979 (1.000)	0.216 (1.000)	2.542 (1.000)
Optimal	5.096 (0.977)	0.224 (0.966)	2.585 (0.983)
$\alpha = 0.0$	5.168 (0.964)	0.227 (0.952)	2.773 (0.917)
$\alpha = 0.1$	5.116 (0.973)	0.225 (0.962)	2.630 (0.967)
$\alpha = 0.2$	5.096 (0.977)	0.224 (0.966)	2.585 (0.983)
$\alpha = 0.3$	5.120 (0.972)	0.225 (0.961)	2.623 (0.969)
$\alpha = 0.4$	5.205 (0.957)	0.229 (0.945)	2.719 (0.935)
$\alpha = 0.5$	5.370 (0.927)	0.236 (0.917)	2.847 (0.893)
$\alpha = 0.6$	5.646 (0.882)	0.247 (0.875)	2.984 (0.852)
$\alpha = 0.7$	6.066 (0.821)	0.263 (0.822)	3.115 (0.816)
$\alpha = 0.8$	6.675 (0.746)	0.284 (0.762)	3.230 (0.787)
$\alpha = 0.9$	7.480 (0.665)	0.308 (0.701)	3.325 (0.765)

NOTE: The true correlation parameter is $\rho = 0.2$. ARE are given in parentheses.

III.2.2 An example

Through simulations, we have seen in the previous section that if the correlation estimate used to compute the regression parameter and their standard errors is outside of the feasible range of the correlation determined by the marginal means, the GEE estimator is very inefficient. In this section we are going to present a real data example where violation of the bounds changes the conclusions of the analysis. The data for this example is a subset of the data from Hamilton's depression study discussed example 2.3. For our analysis, we focus on the differences between only two of the arms: the active treatment and placebo. The covariates are the treatment and time.

Since the response is binary, we use GEE with a logit link function and an AR(1) dependence structure. Performing the GEE analysis on this data, ignoring the correlation bounds, we get the results listed in Part (i) of Table 3.5. Here we see that the time effect is significant, but the p-value for treatment is 0.071, suggests that the treatment is not effective for reducing depression levels. Note that the working correlation estimate is 0.404, which is outside the feasible range (-0.030,0.361). The

Table 3.4: Asymptotic variance and ARE of GEE, $t = 8$

Method	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
Maximum likelihood	7.160 (1.000)	0.260 (1.000)	0.959 (1.000)
Optimal	7.394 (0.968)	0.278 (0.935)	1.042 (0.920)
$\alpha = 0.0$	8.211 (0.872)	0.309 (0.842)	2.766 (0.347)
$\alpha = 0.1$	8.060 (0.888)	0.303 (0.859)	2.161 (0.444)
$\alpha = 0.2$	7.909 (0.905)	0.297 (0.876)	1.707 (0.562)
$\alpha = 0.3$	7.764 (0.922)	0.291 (0.893)	1.397 (0.687)
$\alpha = 0.4$	7.629 (0.939)	0.286 (0.909)	1.204 (0.796)
$\alpha = 0.5$	7.513 (0.953)	0.282 (0.922)	1.100 (0.872)
$\alpha = 0.6$	7.428 (0.963)	0.279 (0.931)	1.053 (0.912)
$\alpha = 0.7$	7.394 (0.968)	0.278 (0.935)	1.042 (0.920)
$\alpha = 0.8$	7.441 (0.962)	0.280 (0.931)	1.049 (0.914)
$\alpha = 0.9$	7.614 (0.940)	0.283 (0.920)	1.064 (0.901)

NOTE: The true correlation parameter is $\rho = 0.7$. ARE are given in parentheses.

results of the analysis of the data using the first order Markov chain model which adheres to the correlation bounds is in Part (ii) of Table 3.5. We now see that the treatment effect has a small p-value, providing evidence that the treatment does reduce depression.

Figure 3.1 shows the plot of regression coefficient for treatment (β_T) obtained by GEE with different values of working correlation parameter. We can see that the regression coefficient remains mostly constant within the correlation bounds, but it changes dramatically when the working correlation α is outside the bounds. Further, we can see that when α is larger than 0.7, the sign of the regression coefficient could change as well. The behavior of the standard errors and p-values is similar as seen in Figures 3.2 and 3.3, respectively.

Table 3.5: Analysis of parameter estimates for Ham-D data

(i) Parameter	Est.	SE	p-value	(ii) Parameter	Est.	SE	p-value
Intercept	-4.49	0.237	<0.001	Intercept	-4.49	0.270	<0.001
Treatment	0.38	0.211	0.071	Treatment	0.38	0.188	0.043
Time	1.01	0.058	<0.001	Time	1.02	0.055	<0.001
Est.corr	Correlation bounds			Est.corr	Correlation bounds		
0.404	[-0.030,0.361]			0.361	[-0.030,0.361]		

NOTE: (i) Analysis using GEE (ii) Analysis using Markov chain model.

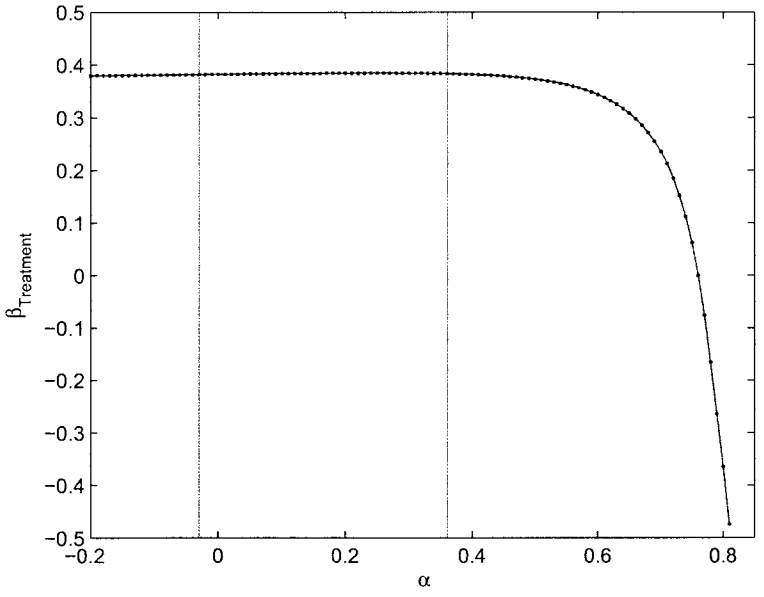


Figure 3.1: Plot of estimate of β_T

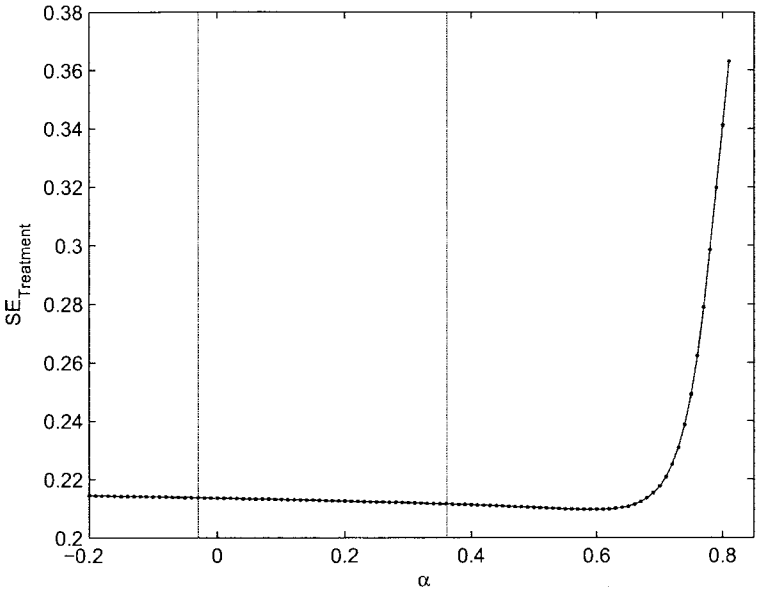


Figure 3.2: Plot of standard error of β_T

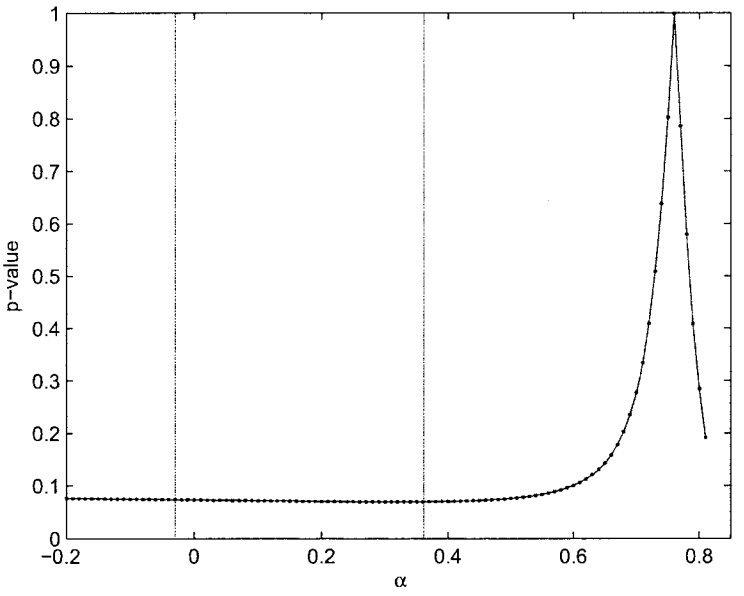


Figure 3.3: Plot of p-value of β_T

III.3 Quadratic inference functions

The generalized estimating equations has been a popular method of estimation for marginal models. However, it has several shortcomings when applied to discrete data. First, the GEE treats dependence as a nuisance and estimates the working correlation parameter with moment estimators constructed using sums of squared Pearson residuals. There is no guarantee that these estimates of the working correlation will lie within the admissible range restricted by the marginal means. There could be considerable loss in efficiency for the regression parameter when the correlation estimate violates the correlation bounds. Another practical shortcoming of GEE is the lack of an objective function which could be used for covariate and model selection purposes. Several authors have suggested modifications and improvements of the GEE to overcome these difficulties. Among these improvements, noteworthy to mention is the quadratic inference function (QIF) method proposed by Qu et al. (2000). This method eliminates estimating the working correlation parameter altogether and borrows ideas from the generalized method of moments estimation procedure due to Hansen (1982). The QIF method also introduces an objective function which is minimized to get the regression parameter estimate. In the next three sections we will present details of the QIF method and study it's efficiency as compared to the likelihood procedures using simulations.

The quadratic inference function is derived by observing that the inverse of the working correlation matrix $R(\alpha)$ can be written as a linear combination of some basis matrices $M_1, \cdots, M_m$, that is,

$$R^{-1}(\alpha) = a_1 M_1 + a_2 M_2 + \cdots + a_m M_m, \tag{3.3.1}$$

where M_i are known symmetric matrices and $a_1, a_2, \cdots, a_m$ are unknown constants, which depend on α. In the next section we give examples of the basis matrices.

III.3.1 Choice of basis matrices

Equicorrelated. Suppose the correlation matrix $R(\alpha)$ of dimension t has an equicorrelated structure, that is, all the off-diagonals are equal to α.

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix} = (1-\alpha)\,\mathbf{I} + b\,\mathbf{J},$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a matrix of ones. The inverse of $R(\alpha)$ is given by

$$R^{-1}(\alpha) = \frac{1}{1-\alpha}\left[\mathbf{I} - \frac{\alpha}{1+(t-1)\alpha}\mathbf{J}\right] = a_1\,M_1 + a_2\,M_2, \tag{3.3.2}$$

where $a_1 = \frac{1}{1-\alpha}$, $a_2 = \frac{-\alpha}{(1-\alpha)(1+(t-1)\alpha)}$, $M_1 = \mathbf{I}$, $M_2 = \mathbf{J}$. The choice of the M matrices is not unique. For example, we could also take $a_1 = \frac{1+(k-2)\alpha}{1-\alpha}$, $a_2 = \frac{-\alpha}{(1-\alpha)(1+(t-1)\alpha)}$, $M_1 = \mathbf{I}$, and

$$M_2 = \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}.$$

First-order autoregressive. Suppose $R(\alpha) = (\alpha^{|i-j|})$ is the first-order autoregressive correlation matrix. The inverse of $R(\alpha)$ equals

$$R^{-1}(\alpha) = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}. \tag{3.3.3}$$

This can be written as a linear combination of three basis matrices,

$$R^{-1}(\alpha) = a_1 M_1 + a_2 M_2 + a_3 M_3$$

$$= a_1 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} + a_2 \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} + a_3 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

where

$$a_1 = \frac{1+\alpha^2}{1-\alpha^2}, \quad a_2 = \frac{-\alpha}{1-\alpha^2}, \quad a_3 = \frac{-\alpha^2}{1-\alpha^2}.$$

The third term in the above expression captures the edge effect of the auto-regressive process. And the third basis matrix M_3, which has only two nonzero values, is a minor boundary correction and can be omitted, that is, we can approximate the inverse by $a_1 M_1 + a_2 M_2$.

We will mainly consider these two correlation structures for comparisons of efficiency between QIF and likelihood methods.

III.3.2 Extended score function

Using the representation (3.3.1) for $R^{-1}(\alpha)$, we can rewrite the generalized estimating equation (3.2.1) as

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2}(a_1 M_1 + a_2 M_2 + \cdots + a_m M_m) A_i^{-1/2}(y_i - \mu_i(\beta)) = 0. \tag{3.3.4}$$

This is equivalent to

$$\sum_{j=1}^{m} a_j \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2} M_j A_i^{-1/2}(y_i - \mu_i(\beta)) = 0. \tag{3.3.5}$$

Thus we see that this is a linear combination of GEE type of estimating equations that involve the basis matrices M_j in place of the inverse of the working correlation

matrix. The QIF method bypasses estimating the coefficients a_j and considers the "extended score' function $\overline{g}_n(\beta)$ given by

$$\overline{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2} M_2 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2} M_m A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}. \tag{3.3.6}$$

The dimension of $\overline{g}_n(\beta)$ is $mk \times 1$, where k is the dimension of the regression parameter β. Therefore (3.3.6) contains more than k equations and hence we cannot obtain a unique estimate of β solving $\overline{g}_n(\beta) = 0$ directly. The approach of QIF is to estimate β by minimizing the quadratic function $\overline{g}_n'(\beta)\,\mathbf{C}^{-1}\,\overline{g}_n(\beta)$ where $\mathbf{C}$ is the covariance matrix of $g_i(\beta)$. A consistent estimate of $\mathbf{C}$ is $\overline{C}_n = (1/n)\sum_{i=1}^{n} g_i(\beta) g_i'(\beta)$. The objective function

$$Q_n(\beta) = n\,\overline{g}_n'(\beta)\,\overline{C}_n^{-1}\,\overline{g}_n(\beta) \tag{3.3.7}$$

is known as the quadratic inference function (Qu et al. (2000)). Thus QIF estimate of β is

$$\widehat{\beta} = \arg\min_{\beta} Q_n(\beta). \tag{3.3.8}$$

Surface plots of quadratic inference function $Q_n(\beta)$ with different parameters are shown in Figure 3.4. To generate these plots we generated binary responses from the model

$$\text{logit}(p_{it}) = \beta_0 + x_{it}\,\beta_1, \tag{3.3.9}$$

where $i = 1, \cdots, 10$; $t = 1, 2, 3, 4$. We took $\beta_0 = 0.9$ and $\beta_1 = -0.3$. The covariates x_{it}'s are generated from standard normal distribution. Correlated binary responses y_i are simulated using the entropy maximization method described in Chaganty and Joe (2006).

First, we simulated responses y_i where the correlation structure among repeated measurements is equicorrelated, and we calculated the QIF defined in (3.3.7) using

the basis matrices M_1 and M_2 designed for equicorrelated structure discussed in Section III.3.1. Next, we simulated data where the correlation structure among repeated measurements is AR(1), and calculated (3.3.7) using the three basis matrices for AR(1) structure. Figure 3.4 shows the plots for these two cases. It is clear that the surface has multiple ridges as well as local minima, but $Q_n(\beta)$ does have a global minimum in both cases which ensures the a unique solution $\widehat{\beta}$ for the minimum (3.3.8).

However, there are situations where $Q_n(\beta)$ does not have a global minimum. For example, we generated binary response variables from the same mean model (3.3.9) given above. But instead of generating x_{it} from standard normal distribution, we chose x_{it} as a fixed categorical covariate. The reason we chose a categorical covariate is they appear frequently in longitudinal data analysis, such as the times when the measurements are taken on subjects. For the data generation we considered both equicorrelated and AR(1) structures. Figure 3.5 shows two examples where the surface of QIF is almost flat and thus does not have a global minimum and therefore a unique solution $\widehat{\beta}$ that minimizes $Q_n(\beta)$ does not exist. This happens because often the matrix $\overline{C}_n$ is almost singular, and its inverse plays a dominant role in the value of $Q_n(\beta)$.

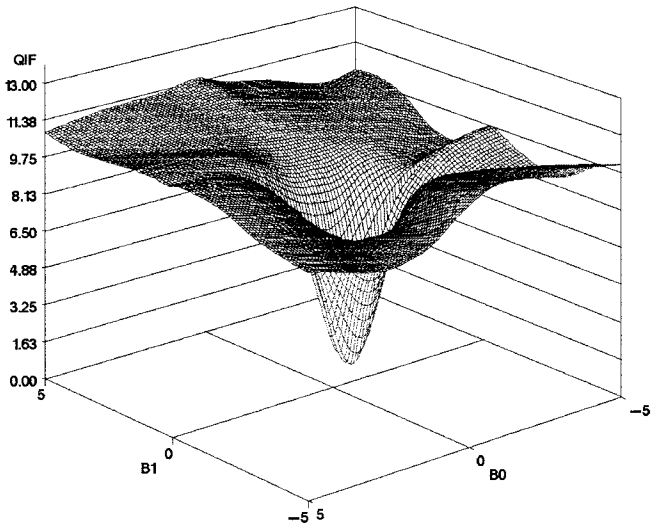
III.3.3 Parameter estimation

As we noted before, a unique estimate of β cannot be obtained solving $\overline{g}_n(\beta) = 0$, since the dimension of $\overline{g}_n(\beta)$ is greater than the number of regression coefficients. So instead, we minimize the quadratic inference function (3.3.7). It is equivalent to setting the partial derivatives of (3.3.7) with respect to β to zero. This leads to the estimating equation

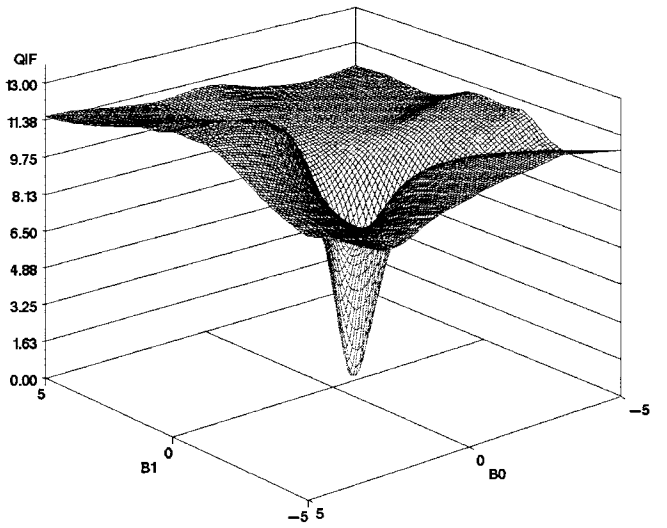
$$\nabla Q_n(\beta) = 0, \tag{3.3.10}$$

where

$$\begin{aligned} n^{-1}\nabla Q_n(\beta) &= 2\nabla\overline{g}_n'\,\overline{C}_n^{-1}\,\overline{g}_n + \overline{g}_n'\,\nabla\overline{C}_n^{-1}\,\overline{g}_n \\ &= 2\nabla\overline{g}_n'\,\overline{C}_n^{-1}\,\overline{g}_n - \overline{g}_n'\,\overline{C}_n^{-1}\,\nabla\overline{C}_n\,\overline{C}_n^{-1}\,\overline{g}_n. \end{aligned} \tag{3.3.11}$$

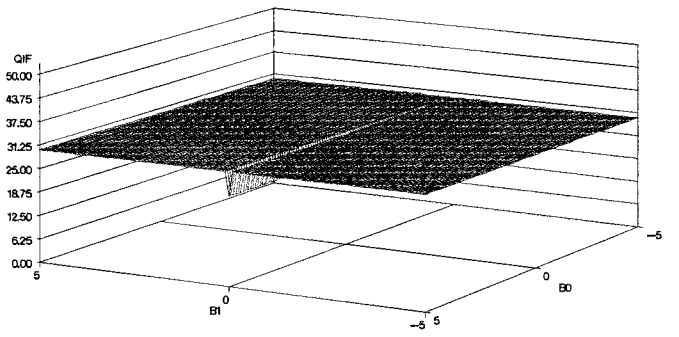


(a) Equicorrelated with $\rho = 0.2$

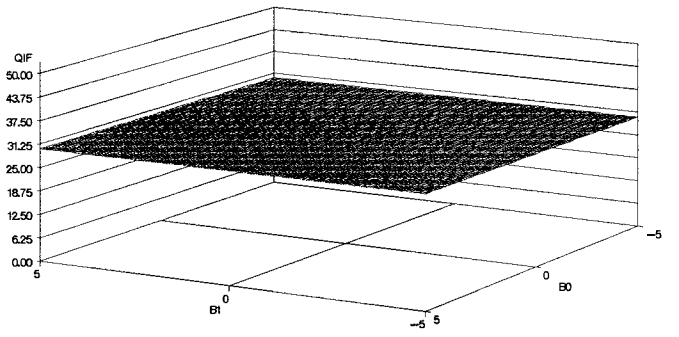


(b) AR(1) with $\rho = 0.8$

Figure 3.4: The surface plot of $Q_n(\beta)$ where a global minimum exists.



(a) Equicorrelated with $\rho = 0.05$



(b) AR(1) with $\rho = 0.28$

Figure 3.5: The surface plot of $Q_n(\beta)$ where no global minimum exists.

Note that $\nabla\overline{g}_n$ is a $mk \times k$ matrix and $\nabla\overline{C}_n$ is a three-dimensional array

$$\left(\frac{\partial\overline{C}_n}{\partial\beta_1}, \frac{\partial\overline{C}_n}{\partial\beta_2}, \cdots, \frac{\partial\overline{C}_n}{\partial\beta_k}\right).$$

We can solve equation (3.3.10) numerically using the Newton-Raphson algorithm, which requires the following second order partial derivatives of $Q_n(\beta)$.

$$\begin{aligned} n^{-1}\nabla^2 Q_n(\beta) &= 2\nabla\overline{g}_n'\,\overline{C}_n^{-1}\,\nabla\overline{g}_n + 2\nabla^2\overline{g}_n'\,\overline{C}_n^{-1}\,\overline{g}_n \\ &\quad -4\nabla\overline{g}_n'\,\overline{C}_n^{-1}\,\nabla\overline{C}_n\,\overline{C}_n^{-1}\,\overline{g}_n + 2\overline{g}_n'\overline{C}_n^{-1}\nabla\overline{C}_n\overline{C}_n^{-1}\nabla\overline{C}_n\overline{C}_n^{-1}\overline{g}_n \\ &\quad -\overline{g}_n'\overline{C}_n^{-1}\nabla^2\overline{C}_n^{-1}\overline{C}_n^{-1}\overline{g}_n. \end{aligned} \tag{3.3.12}$$

Here, $\nabla^2\overline{C}_n$ is a four-dimensional array

$$\left\{\frac{\partial^2\overline{C}_n}{\partial\beta_i\partial\beta_j} : i,j = 1,2,\cdots,k\right\}.$$

The Newton-Raphson method iterates the equation

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} - \{\nabla^2 Q_n(\beta^{(j)})\}^{-1}\nabla Q_n(\beta^{(j)}), \tag{3.3.13}$$

until convergence. To further simplify the iterative algorithm we note that under regularity conditions $\nabla\overline{g}_n$ and $\nabla\overline{C}_n$ have finite limits as $n \to \infty$. Hence they are $O_p(1)$, but $\overline{g}_n$ is $O_p(n^{-1/2})$, so the first term in (3.3.11) is $O_p(n^{-1/2})$, whereas the second term is $O_p(n^{-1})$ and thus it can be asymptotically negligible. Therefore minimizing Q_n is asymptotically equivalent to solving

$$\nabla\overline{g}_n'\,\overline{C}_n^{-1}\,\overline{g}_n = 0. \tag{3.3.14}$$

When $\beta = \widehat{\beta}$, all the terms in $\bigtriangledown^2 Q_n(\beta)$ involving $\overline{g}_n$ are equal to zero except the first term. Therefore we can take

$$\bigtriangledown^2 Q_n(\beta) = 2n \bigtriangledown \overline{g}_n'\overline{C}_n^{-1} \bigtriangledown \overline{g}_n'. \tag{3.3.15}$$

We can use the above expression in the iterative algorithm (3.3.13) to obtain the estimate $\widehat{\beta}$. In the next section we will derive an expression for the asymptotic variance of the QIF regression parameter estimate.

III.3.4 Asymptotic variance

The extended score function (3.3.6) can be written in the form of a weighted estimating equation. Let

$$\Psi(\beta) = n\,\overline{g}_n(\beta) = \sum_{i=1}^{n} \bigtriangledown \boldsymbol{\mu}_i' \, \mathbf{W}_i' \, (y_i - \mu_i), \tag{3.3.16}$$

where

$$\bigtriangledown \boldsymbol{\mu}_i' = \begin{bmatrix} \left(\frac{\partial \mu_i}{\partial \beta}\right)' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \left(\frac{\partial \mu_i}{\partial \beta}\right)' \end{bmatrix}_{mk \times mt} \quad \text{and} \quad \mathbf{W}_i' = \begin{bmatrix} A_i^{-1/2} M_1 A_i^{-1/2} \\ \vdots \\ A_i^{-1/2} M_m A_i^{-1/2} \end{bmatrix}_{mt \times t}.$$

The estimating equation (3.3.16) is clearly an unbiased estimating equation since $\mathrm{E}(y_i - \mu_i) = 0$. By the result stated in Chaganty and Joe (2004), if $\widehat{\beta}$ is the solution of equation (3.3.16) then $\widehat{\beta}$ is asymptotically normal with mean β and covariance matrix equal to inverse of Godambe information matrix $\mathcal{G}^{-1}(\beta) = \left[(-\mathbf{D}_\Psi)' \, \mathbf{V}_\Psi^{-1} \, (-\mathbf{D}_\Psi)\right]^{-1}$. Here

$$\begin{aligned} -\mathbf{D}_\Psi = \mathrm{E}(\partial \Psi(\beta)/\partial \beta') &= \sum_{i=1}^{n} \bigtriangledown \boldsymbol{\mu}_i' \, \mathbf{W}_i' \left(\frac{\partial \mu_i}{\partial \beta'}\right), \\ \mathbf{V}_\Psi = \mathrm{Cov}(\Psi(\beta)) &= \sum_{i=1}^{n} \bigtriangledown \boldsymbol{\mu}_i' \, \mathbf{W}_i' \Sigma_i \mathbf{W}_i \bigtriangledown \boldsymbol{\mu}_i. \end{aligned}$$

Using the matrix Cauchy-Schwarz inequality (Chaganty and Joe (2004)) we can show that the optimal choice for $\mathbf{W}_i$ in (3.3.16) is $\mathbf{W}_i^0 = \left(\frac{\Sigma_i^{-1}}{k}, \frac{\Sigma_i^{-1}}{k}, \ldots, \frac{\Sigma_i^{-1}}{k}\right)'$. In this case the asymptotic covariance matrix $\mathcal{G}^{-1}(\beta)$ of $\widehat{\beta}$ reduces to

$$\left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \beta'}\right)^{-1}.$$

Note that optimal choice $\mathbf{W}_i^0$ depends on the unknown β and on the dependence parameters. But since QIF uses the basis matrices instead of the true covariance matrix, it is not expected to produce the most efficient estimator for the regression parameter.

III.4 Comparison of asymptotic performance

In this section we compare asymptotic performance of QIF regression estimator and the maximum likelihood estimator of Markov chain (MC) model discussed in Chapter II, by computing the asymptotic relative efficiency (ARE) of QIF method with respect to the first order Markov chain procedure. In our simulations we chose the number of repeated measurements as $t = 4, 6$ and 8, and a sample size of $n = 1000$. The asymptotic relative efficiency (ARE) is calculated as follows

$$\text{ARE} = \frac{\text{asymptotic variance of MC}}{\text{asymptotic variance of QIF}}.$$

By varying the correlation parameter ρ over its admissible range, we can see how the efficiency changes when the correlation estimate is far from the true correlation.

In our simulations we generated binary responses from the model

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 x_{it}^D + \beta_2 x_{it}^C,$$

where $i = 1, \cdots, n = 1000$; x_{it}^C is a discrete covariate taking values $1, 2, \ldots, t$ and x_{it}^C is generated from uniform distribution on (0, 1).

We start with the case $t = 4$. The true value of the regression coefficients are taken as $\beta_0 = 0.8$, $\beta_1 = -0.1$ and $\beta_2 = 0.15$. The ARE plots of QIF with respect to the MC procedure for the three regression coefficients are in Figure 3.6. When $\rho = 0$ the ARE is 1, that is, when the repeated measurements are independent QIF is as efficient as the Markov chain model. But the ARE is less than 1 and decreases as ρ increases. However, as seen in Figure 3.6(b), the ARE remains high even when ρ is close to 0.7, which indicates that QIF is comparable to the Markov chain model when estimating the coefficient of the discrete covariate. But the ARE for the regression coefficient corresponding to the continuous drops to less than 0.5 as ρ approaches its upper bound of 0.883 as shown in Figure 3.6(c). This shows QIF estimates the regression coefficient corresponding to the continuous covariate very poorly for highly correlated data. Table 3.6 contains the numerical values of the asymptotic relative efficiencies in the case.

We now consider the case $t = 6$. The ARE plots for the three regression coefficients are shown in Figure 3.7. The plots clearly show the QIF estimates are

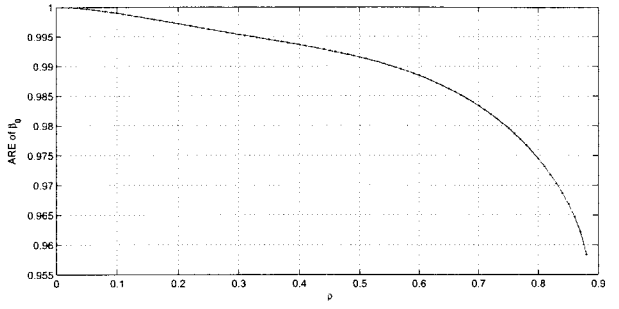
Table 3.6: Asymptotic variance and ARE of QIF, $t = 4$

	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.0$	10.005 (1.000)	0.885 (1.000)	13.280 (1.000)
$\rho = 0.1$	10.363 (0.999)	0.922 (0.999)	13.070 (0.999)
$\rho = 0.2$	10.539 (0.997)	0.940 (0.997)	12.479 (0.997)
$\rho = 0.3$	10.527 (0.995)	0.938 (0.995)	11.552 (0.994)
$\rho = 0.4$	10.326 (0.994)	0.912 (0.994)	10.355 (0.989)
$\rho = 0.5$	9.935 (0.992)	0.857 (0.994)	8.960 (0.977)
$\rho = 0.6$	9.353 (0.988)	0.770 (0.994)	7.440 (0.952)
$\rho = 0.7$	8.573 (0.983)	0.647 (0.993)	5.861 (0.902)
$\rho = 0.8$	7.587 (0.974)	0.481 (0.986)	4.276 (0.787)

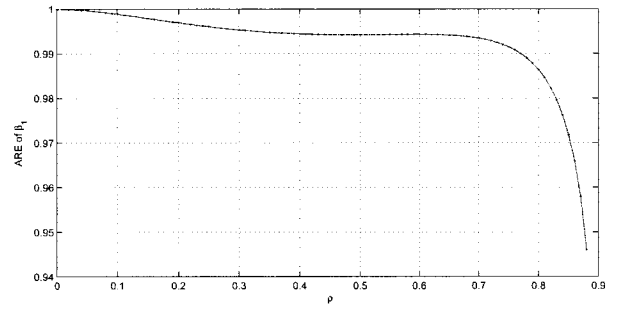
NOTE: Range of ρ is $(-0.450, 0.883)$. The parameter values are $\beta_0 = 0.8$, $\beta_1 = -0.1$, $\beta_2 = 0.15$, and $n = 1000$; AREs are given in parentheses.

uniformly less efficient as compared to the MC estimates for correlated data. In Figures 3.7(a) and (c) the ARE decreases as ρ approaches to its upper bound. But in Figure 3.7(b) we can see the plot of ARE is not monotone, it decreases to 0.935 and increases back to 0.98. This is different from the case when the number of repeated measurements is $t = 4$. Table 3.7 contains the numerical values of the AREs for the three regression coefficients.

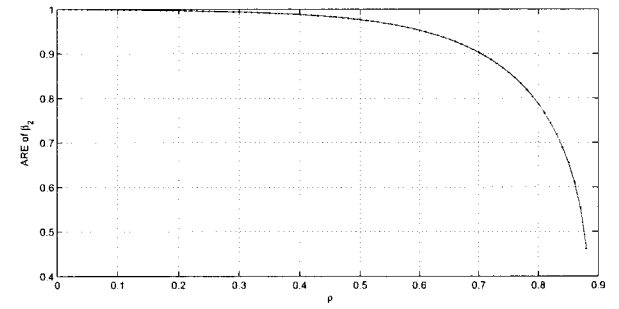
Figure 3.8 and Table 3.8 shows the ARE plots and numerical values, respectively, in the case $t = 8$. The pattern is the same as in the previous two cases. Based on these results, we can conclude that for highly correlated binary data with autoregressive structure, QIF is very inefficient for estimating the regression coefficient for the continuous covariates but reasonably efficient for discrete covariates. However if the correlation is small it is respectable efficiency compared to the maximum likelihood estimator from the Markov chain model.



(a)



(b)



(c)

Figure 3.6: ARE of QIF and MC methods with binary data

Table 3.7: Asymptotic variance of QIF and ARE, $t = 6$

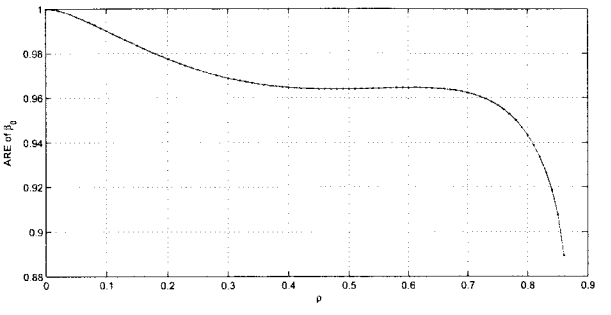
	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.0$	10.905 (1.000)	0.540 (1.000)	18.080 (1.000)
$\rho = 0.1$	11.485 (0.990)	0.593 (0.985)	17.772 (0.990)
$\rho = 0.2$	11.885 (0.978)	0.641 (0.963)	16.893 (0.979)
$\rho = 0.3$	12.097 (0.969)	0.681 (0.947)	15.522 (0.972)
$\rho = 0.4$	12.112 (0.965)	0.708 (0.938)	13.770 (0.966)
$\rho = 0.5$	11.909 (0.964)	0.717 (0.936)	11.761 (0.956)
$\rho = 0.6$	11.447 (0.965)	0.670 (0.943)	9.615 (0.933)
$\rho = 0.7$	10.664 (0.962)	0.644 (0.956)	7.432 (0.874)
$\rho = 0.8$	9.473 (0.943)	0.533 (0.973)	5.293 (0.705)

NOTE: Range of ρ is $(-0.091, 0.862)$. The parameter values are $\beta_0 = 1.2$, $\beta_1 = 0.6$, $\beta_2 = 0.1$, and $n = 1000$; AREs are given in parentheses.

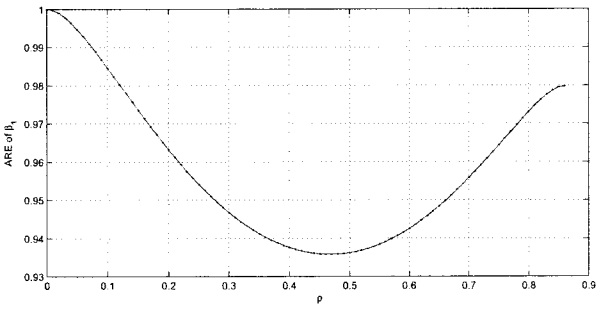
Table 3.8: Asymptotic variance of QIF and ARE, $t = 8$

	$nV(\beta_0)$	$nV(\beta_1)$	$nV(\beta_2)$
$\rho = 0.0$	6.509 (1.000)	0.293 (1.000)	11.200 (1.000)
$\rho = 0.1$	6.905 (0.993)	0.333 (0.983)	10.984 (0.995)
$\rho = 0.2$	7.034 (0.985)	0.375 (0.961)	10.400 (0.987)
$\rho = 0.3$	7.398 (0.977)	0.416 (0.943)	9.505 (0.976)
$\rho = 0.4$	7.480 (0.969)	0.455 (0.931)	8.377 (0.953)
$\rho = 0.5$	7.428 (0.958)	0.487 (0.929)	7.097 (0.899)
$\rho = 0.6$	7.193 (0.935)	0.505 (0.940)	5.744 (0.760)
$\rho = 0.7$	6.703 (0.862)	0.497 (0.972)	4.381 (0.300)

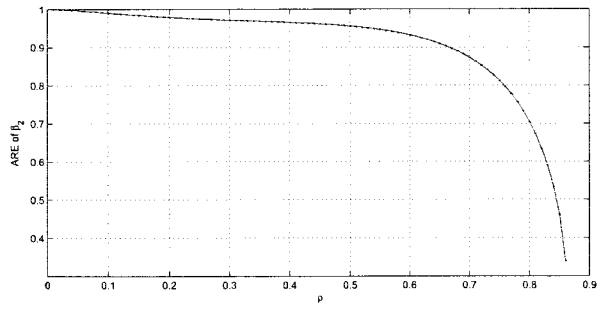
NOTE: Range of ρ is $(-0.027, 0.706)$. The parameter values are $\beta_0 = 1.0$, $\beta_1 = -0.6$, $\beta_2 = -0.1$, and $n = 1000$; AREs are given in parentheses.



(a)

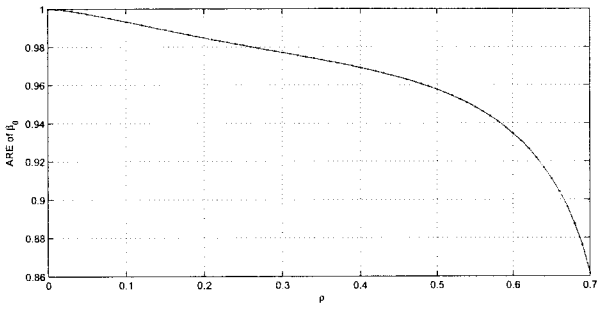


(b)

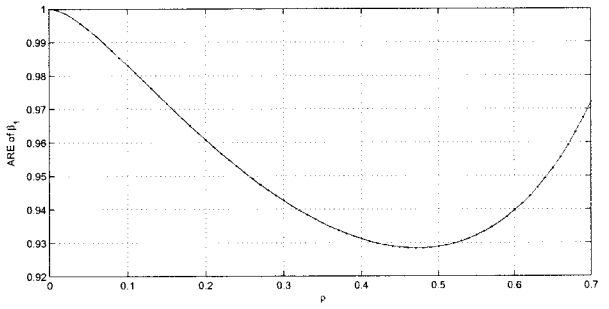


(c)

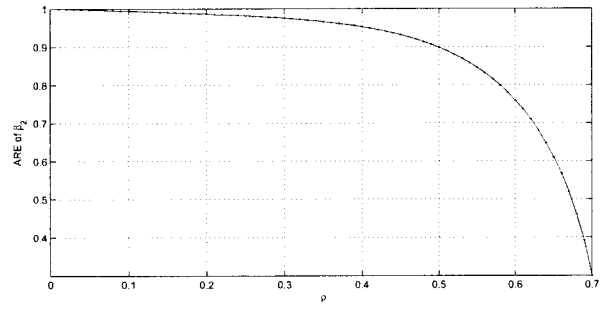
Figure 3.7: ARE of QIF and MC methods with binary data



(a)



(b)



(c)

Figure 3.8: ARE for QIF and MC methods with binary data

III.5 Comparison of small-sample performance

In this section we compare the small-sample performance of QIF estimator with the maximum likelihood estimate from the first order Markov chain (MC) model using simulated data. We fixed the sample size as $n = 30$. Correlated binary random vectors with AR(1) structure are simulated separately from two models. The first model is

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 x_{it}^C, \tag{3.5.1}$$

where x_{it}^C are generated from standard normal distribution. The second model is

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 x_{it}^D + \beta_2 x_{it}^C, \tag{3.5.2}$$

where x_{it}^D take values $1, 2, \ldots, t$. We fixed the true values of the regression coefficients as $\beta_0 = 0.5$ and $\beta_1 = -0.5$. We considered two cases $t = 4$ and $t = 8$. Fixing a value of ρ within the correlation bounds we simulated binary data and calculated QIF and MC estimators for the regression coefficients. We then repeated the procedure 1,000 times for each combination of the parameter choices. The mean square errors (MSE) are calculated averaging the squared deviations of the estimate from the true regression parameter values. The relative efficiencies (RE) are calculated taking ratios as follows:

$$RE = \frac{\text{MSE of MC estimate}}{\text{MSE of QIF estimate}}.$$

Table 3.9 and Table 3.10 shows results from the simulations with $t = 4$ and $t = 8$ respectively. We can see that the values of RE are less than 1 for all values of ρ within the feasible range $(-0.250, 0.465)$, indicating that QIF estimator has larger MSE than the MC estimator. Further the RE of $\widehat{\beta_1}$ is becoming smaller when ρ increases to the upper bound, more specifically, the efficiency is 0.531 when $\rho = 0.45$. For $t = 8$, the efficiencies of QIF are better than those when $t = 4$. The conclusion that we can make from these simulations is that QIF estimates the regression parameter poorly for small samples when there is a high correlation. However, its performance improves when the number of repeated measurements increases.

Next, we did similar simulations using model (3.5.2), fixing $\beta_0 = 0.1$, $\beta_1 = 0.2$ and $\beta_2 = -0.5$. Note that there is a discrete covariate x_{it}^D in this model. The MSEs

Table 3.9: MSE and RE of QIF and MC with one covariate, $t = 4$

	MSE($\widehat{\beta}_0$)			MSE($\widehat{\beta}_1$)		
ρ	QIF	MC	RE	QIF	MC	RE
0.00	0.0464	0.0373	0.805	0.0588	0.0436	0.741
0.05	0.0457	0.0359	0.786	0.0690	0.0522	0.757
0.10	0.0631	0.0515	0.817	0.0494	0.0418	0.846
0.15	0.0652	0.0466	0.715	0.0624	0.0436	0.699
0.20	0.0656	0.0485	0.740	0.0534	0.0417	0.780
0.25	0.0687	0.0541	0.788	0.0464	0.0332	0.715
0.30	0.0678	0.0541	0.798	0.0500	0.0355	0.711
0.35	0.0851	0.0623	0.732	0.0409	0.0284	0.694
0.40	0.0918	0.0637	0.693	0.0466	0.0276	0.591
0.45	0.1019	0.0768	0.753	0.0419	0.0222	0.531

NOTE: Range of ρ is $(-0.250, 0.465)$ and $\beta = (0.5, -0.5)'$.

Table 3.10: MSE and RE of QIF and MC with one covariate, $t = 8$

	MSE($\widehat{\beta}_0$)			MSE($\widehat{\beta}_1$)		
ρ	QIF	MC	RE	QIF	MC	RE
0	0.0202	0.0180	0.891	0.0255	0.0218	0.853
0.05	0.0231	0.0204	0.883	0.0236	0.0203	0.859
0.10	0.0290	0.0250	0.862	0.0255	0.0218	0.853
0.15	0.0261	0.0231	0.885	0.0255	0.0223	0.872
0.20	0.0303	0.0269	0.888	0.0251	0.0210	0.837
0.25	0.0366	0.0300	0.819	0.0201	0.0165	0.823
0.30	0.0399	0.0332	0.833	0.0222	0.0153	0.691
0.35	0.0435	0.0343	0.790	0.0195	0.0114	0.585

NOTE: Range of ρ is $(-0.283, 0.385)$ and $\beta = (0.5, -0.5)'$.

Table 3.11: MSE and RE of QIF with both covariates, $t = 4$

α	MSE($\hat{\beta}_0$)	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)
0.00	0.328 (0.743)	0.048 (0.730)	0.080 (0.573)
0.05	0.363 (0.687)	0.055 (0.639)	0.086 (0.556)
0.10	0.324 (0.788)	0.043 (0.765)	0.075 (0.645)
0.15	0.296 (0.774)	0.044 (0.756)	0.070 (0.623)
0.20	0.388 (0.654)	0.064 (0.580)	0.094 (0.472)
0.25	0.374 (0.687)	0.045 (0.699)	0.079 (0.487)
0.30	0.410 (0.666)	0.055 (0.637)	0.078 (0.494)
0.35	0.388 (0.679)	0.055 (0.593)	0.070 (0.374)

NOTE: Range of ρ is $(-0.162, 0.358)$ and $\beta = (0.1, 0.2, -0.5)'$. REs are given in the parentheses.

and the AREs are displayed for $t = 4$ and $t = 8$ in Tables 3.11 and 3.12, respectively. We can see that in the second model QIF performance is much worse than in the first model, the relative efficiency fell below 0.35 in some cases.

Changing the true values of the regression coefficients and performing small sample simulations, we found cases where QIF did really a poor job in estimating the regression coefficients. The simulations results from one of those cases are shown in Table 3.13. Here we fixed $\beta_0 = 0.5$ $\beta_1 = 0.5$ and $\beta_2 = -0.5$. The true marginal probabilities in this case do not vary much and are in the range 0.7 to 0.8. In this case the efficiency of QIF has dropped to as low as 0.168.

Table 3.12: MSE and RE of QIF with both covariates, $t = 8$

α	MSE($\hat{\beta}_0$)	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)
0.00	0.121 (0.797)	0.006 (0.739)	0.039 (0.743)
0.05	0.147 (0.819)	0.007 (0.734)	0.041 (0.694)
0.10	0.155 (0.752)	0.001 (0.714)	0.041 (0.712)
0.15	0.157 (0.734)	0.009 (0.614)	0.042 (0.658)
0.20	0.146 (0.812)	0.008 (0.693)	0.032 (0.717)
0.25	0.165 (0.804)	0.008 (0.691)	0.037 (0.571)
0.30	0.183 (0.646)	0.009 (0.635)	0.035 (0.481)
0.35	0.201 (0.749)	0.010 (0.685)	0.037 (0.333)

NOTE: Range of ρ is $(-0.104, 0.351)$ and $\beta = (0.1, 0.2, -0.5)'$. REs are given in parentheses.

Table 3.13: MSE and RE of QIF with both covariates, $t = 8$

α	MSE($\hat{\beta}_0$)	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)
0.00	0.334 (0.455)	0.067 (0.273)	0.122 (0.448)
0.05	0.392 (0.440)	0.109 (0.184)	0.131 (0.416)
0.10	0.442 (0.510)	0.109 (0.228)	0.127 (0.448)
0.15	0.519 (0.411)	0.135 (0.179)	0.141 (0.385)
0.20	0.505 (0.411)	0.117 (0.180)	0.156 (0.377)
0.25	0.548 (0.378)	0.113 (0.234)	0.146 (0.323)
0.30	0.500 (0.425)	0.159 (0.152)	0.133 (0.286)
0.35	0.477 (0.349)	0.119 (0.168)	0.152 (0.287)

NOTE: Range of ρ is $(-0.1034, 0.351)$ and $\beta = (0.5, 0.5, -0.5)'$. REs are given in parentheses.

III.6 A robust variance estimator in QIF

As stated in Section III.3.2, Qu, et al. (2000) estimated the covariance matrix $\mathbf{C}$ by

$$\begin{aligned}\overline{C}_n &= \frac{1}{n}\sum_{i=1}^{n} g_i(\beta)g_i'(\beta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} g_{i1}(\beta)g_{i1}'(\beta) & g_{i1}(\beta)g_{i2}'(\beta) & \cdots & g_{i1}(\beta)g_{im}'(\beta) \\ g_{i2}(\beta)g_{i1}'(\beta) & g_{i2}(\beta)g_{i2}'(\beta) & \cdots & g_{i2}(\beta)g_{im}'(\beta) \\ \vdots & \vdots & \ddots & \\ g_{im}(\beta)g_{i1}'(\beta) & g_{im}(\beta)g_{i2}'(\beta) & \cdots & g_{im}(\beta)g_{im}'(\beta) \end{bmatrix}, \quad (3.6.1)\end{aligned}$$

where

$$\begin{aligned} g_{ij}(\beta) &= \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2} M_j A_i^{-1/2}(y_i - \mu_i) = D_{ij}'(\beta)(y_i - \mu_i), \\ g_{ij}(\beta)g_{ik}'(\beta) &= \left(\frac{\partial \mu_i}{\partial \beta}\right)' A_i^{-1/2} M_j A_i^{-1/2}(y_i - \mu_i)(y_i - \mu_i)' A_i^{-1/2} M_k A_i^{-1/2}\left(\frac{\partial \mu_i}{\partial \beta}\right) \\ &= D_{ij}'(\beta)(y_i - \mu_i)(y_i - \mu_i)' D_{ik}(\beta). \end{aligned}$$

The estimate $\overline{C}_n$ is also called an unstructured estimator since it is obtained without any parametric specification. Recall that the framework of GEE and QIF has the assumption that there is a common correlation structure across all subjects, but they both estimate the covariance for the ith subject using $(y_i - \mu_i)(y_i - \mu_i)'$, that is, using only the ith observation. This is not an optimal estimator for $\mathrm{Cov}(y_i)$, because it is neither consistent nor efficient since it ignores the basic assumption that there is a common correlation for all subjects. Pan (2001) proposed a robust estimator for $\mathrm{Cov}(y_i)$ in the context of GEE. It is obtained by pooling observations across different subjects and it is given by

$$V_i = A_i^{1/2}\left(\sum_{i=1}^{n} A_i^{-1/2}(y_i - \mu_i)(y_i - \mu_i)' A_i^{-1/2}\right) A_i^{1/2}. \quad (3.6.2)$$

This estimator could also be used to estimate the covariance for the repeated measurements on the ith subject in QIF as well. Thus our modified estimate of the

covariance matrix of $g_i(\beta)$ is

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} D'_{i1}(\beta)V_iD_{i1}(\beta) & D'_{i1}(\beta)V_iD_{i2}(\beta) & \cdots & D'_{i1}(\beta)V_iD_{im}(\beta) \\ D'_{i2}(\beta)V_iD_{i1}(\beta) & D'_{i2}(\beta)V_iD_{i2}(\beta) & \cdots & D'_{i2}(\beta)V_iD_{im}(\beta) \\ \vdots & \vdots & \ddots & \\ D'_{im}(\beta)V_iD_{i1}(\beta) & D'_{im}(\beta)V_iD_{i2}(\beta) & \cdots & D'_{im}(\beta)V_iD_{im}(\beta) \end{bmatrix} \tag{3.6.3}$$

Pan (2001) provided a partial justification showing that asymptotically, $\text{Cov}(\text{vec}(\overline{W}_n)) \leq \text{Cov}(\text{vec}(\overline{C}_n))$. Thus, our proposed modification of QIF is

$$Q_n^*(\beta) = n\,\overline{g}'_n(\beta)\,\overline{W}_n^{-1}\,\overline{g}_n(\beta).$$

In the next section we study the performance of this modified QIF with respect to the original QIF in small samples.

III.7 Performance of modified QIF

Here we compare the performance of the modified QIF (mQIF) method discussed in Section III.6 with respect to the maximum likelihood estimate of the Markov chain model using small-sample simulations. We used models (3.5.1) and (3.5.2) and AR(1) and equicorrelated correlation structures. The MSEs were calculated generating 1000 samples of size $n = 30$ for both the mQIF and MC estimators. The relative efficiency is calculated by the ratio

$$RE = \frac{\text{MSE of MC estimate}}{\text{MSE of mQIF estimate}}.$$

III.7.1 Small-sample performance with AR(1) structure

Tables 3.14 and 3.15 shows the MSEs and relative efficiency of mQIF from model (3.5.1) with $t = 4$ and $t = 8$, respectively. Comparing these efficiency values with Tables 3.9 and 3.10, we can see that the mQIF estimator is an improvement over QIF estimator. For small values of ρ, mQIF is comparable with MC, specially the RE of $\widehat{\beta}_2$ is close to one when ρ is near zero. However, for large values of *rho*, even the mQIF is less efficient than MC, though it performs better than the QIF estimator.

Table 3.14: MSE and RE of mQIF and MC with one covariate, $t = 4$

	MSE($\widehat{\beta}_0$)			MSE($\widehat{\beta}_1$)		
ρ	mQIF	MC	RE	mQIF	MC	RE
0.00	0.041	0.037	0.913	0.044	0.044	0.995
0.05	0.040	0.036	0.906	0.052	0.052	0.996
0.10	0.057	0.052	0.908	0.042	0.042	0.993
0.15	0.053	0.047	0.887	0.044	0.044	0.989
0.20	0.056	0.049	0.875	0.043	0.042	0.976
0.25	0.060	0.054	0.899	0.035	0.033	0.937
0.30	0.059	0.054	0.918	0.039	0.036	0.919
0.35	0.072	0.062	0.862	0.034	0.028	0.839
0.40	0.079	0.064	0.808	0.037	0.028	0.755
0.45	0.088	0.077	0.869	0.034	0.022	0.651

NOTE: Range of ρ is $(-0.250, 0.465)$ and $\beta = (0.5, -0.5)'$.

Table 3.15: MSE and RE of mQIF and MC with one covariate, $t = 8$

	MSE($\widehat{\beta}_0$)			MSE($\widehat{\beta}_1$)		
ρ	mQIF	MC	RE	mQIF	MC	RE
0.00	0.019	0.018	0.934	0.022	0.022	1.002
0.05	0.021	0.020	0.964	0.020	0.020	1.010
0.10	0.027	0.025	0.915	0.022	0.022	1.001
0.15	0.024	0.023	0.962	0.023	0.022	0.986
0.20	0.029	0.027	0.925	0.021	0.021	0.989
0.25	0.033	0.030	0.909	0.018	0.017	0.922
0.30	0.037	0.033	0.909	0.019	0.015	0.803
0.35	0.040	0.034	0.849	0.017	0.011	0.685

NOTE: Range of ρ is $(-0.283, 0.385)$ and $\beta = (0.5, -0.5)'$.

Table 3.16: MSE and RE of mQIF with both covariates, $t = 4$

α	MSE($\widehat{\beta}_0$)	MSE($\widehat{\beta}_1$)	MSE($\widehat{\beta}_2$)
0.00	0.284 (0.861)	0.042 (0.835)	0.049 (0.938)
0.05	0.305 (0.816)	0.044 (0.799)	0.048 (0.983)
0.10	0.302 (0.846)	0.041 (0.792)	0.050 (0.961)
0.15	0.275 (0.834)	0.040 (0.828)	0.047 (0.937)
0.20	0.324 (0.785)	0.048 (0.775)	0.050 (0.898)
0.25	0.327 (0.784)	0.041 (0.770)	0.048 (0.807)
0.30	0.346 (0.788)	0.044 (0.787)	0.050 (0.779)
0.35	0.333 (0.792)	0.046 (0.712)	0.044 (0.593)

NOTE: Range of ρ is $(-0.162, 0.358)$ and $\beta = (0.1, 0.2, -0.5)'$. RE are given in parentheses.

Next, we compare mQIF estimator with simulating observations from model 3.5.2 that includes a discrete covariate. The results of the simulations are in Tables 3.16 and 3.17. When we compare these numbers with the ones in Tables 3.11 and 3.12, we can see that even in this case mQIF does better than the QIF estimator. So the mQIF estimator out performs QIF estimator in all cases.

Finally, we reexamined the cases where QIF did very poorly and checked the efficiency of mQIF in those cases. The results are in Table 3.18. Although mQIF is much less efficient compared to the MC estimator, but it performs better than QIF. We can see that the efficiency of $\widehat{\beta}_2$ is twice as high as the corresponding QIF estimator.

III.7.2 Small-sample performance with equicorrelated structure

Now, we compare the performance of mQIF estimator with QIF estimator through simulated small sample data with equicorrelated structure. We kept the two covariates the same in models 3.5.1 and 3.5.2. The observations are generated from the multivariate model with equicorrelated structure, and MSEs are calculated for mQIF

Table 3.17: MSE and RE of QIF with both covariates, $t = 8$

α	MSE($\widehat{\beta}_0$)	MSE($\widehat{\beta}_1$)	MSE($\widehat{\beta}_2$)
0.00	0.114 (0.852)	0.006 (0.821)	0.030 (0.966)
0.05	0.134 (0.894)	0.007 (0.833)	0.030 (0.952)
0.10	0.137 (0.854)	0.007 (0.792)	0.030 (0.968)
0.15	0.141 (0.818)	0.008 (0.704)	0.030 (0.915)
0.20	0.139 (0.853)	0.007 (0.794)	0.025 (0.909)
0.25	0.159 (0.836)	0.008 (0.771)	0.030 (0.704)
0.30	0.157 (0.756)	0.008 (0.715)	0.023 (0.723)
0.35	0.192 (0.784)	0.010 (0.709)	0.024 (0.514)

NOTE: Range of ρ is $(-0.104, 0.351)$ and $\beta = (0.1, 0.2, -0.5)'$. RE are given in parentheses.

Table 3.18: MSE and RE of QIF with both covariates, $t = 8$

α	MSE($\widehat{\beta}_0$)	MSE($\widehat{\beta}_1$)	MSE($\widehat{\beta}_2$)
0.00	0.253 (0.600)	0.041 (0.445)	0.067 (0.816)
0.05	0.267 (0.646)	0.042 (0.481)	0.060 (0.903)
0.10	0.317 (0.711)	0.040 (0.617)	0.071 (0.800)
0.15	0.330 (0.647)	0.045 (0.537)	0.070 (0.774)
0.20	0.307 (0.677)	0.051 (0.409)	0.075 (0.782)
0.25	0.317 (0.652)	0.062 (0.429)	0.074 (0.641)
0.30	0.342 (0.621)	0.065 (0.374)	0.067 (0.571)
0.35	0.314 (0.530)	0.059 (0.341)	0.074 (0.593)

NOTE: Range of ρ is $(-0.008, 0.359)$ and $\beta = (0.5, 0.5, -0.5)'$. RE are given in parentheses.

Table 3.19: MSE and RE of QIF and mQIF with one covariate, $t = 4$

	MSE($\widehat{\beta}_0$)			MSE($\widehat{\beta}_1$)		
ρ	QIF	mQIF	RE	QIF	mQIF	RE
0.00	0.046	0.039	0.848	0.176	0.159	0.905
0.05	0.056	0.044	0.779	0.204	0.176	0.863
0.10	0.072	0.050	0.700	0.188	0.174	0.925
0.15	0.068	0.053	0.782	0.217	0.188	0.865
0.20	0.078	0.062	0.790	0.217	0.202	0.931
0.25	0.084	0.066	0.783	0.228	0.206	0.901
0.30	0.092	0.066	0.721	0.230	0.215	0.933
0.35	0.100	0.074	0.733	0.240	0.221	0.921
0.40	0.111	0.084	0.761	0.263	0.249	0.946

NOTE: Upper bound of ρ is 0.428 and $\beta = (0.5, -0.5)'$.

and QIF estimators. The efficiency is calculated as

$$RE = \frac{\text{MSE of mQIF estimate}}{\text{MSE of QIF estimate}}.$$

The simulation results can be found in the following tables 3.19, 3.20, 3.21 and 3.22. The message is clear, mQIF is a better estimator than QIF estimator.

III.8 Misspecification of correlation structure

Qu at el. (2000) showed that if the repeated measurements are continuous and normally distributed, under the correct assumption of working correlation structure, QIF performs as good as GEE. And if the working correlation structure is misspecified, the QIF approach is more efficient than GEE. In this section we examine the performance of the MC method under misspecification as opposed to the mQIF estimator under correct specification of the correlation structure for binary data.

We generated binary samples of size $n = 30$ with repeated measurements equal to eight ($t = 8$), and equicorrelated structure, from models 3.5.1 and 3.5.2. Here modified QIF estimates are obtained using the basic matrix for equicorrelated structure,

Table 3.20: MSE and RE of QIF and mQIF with one covariate, $t = 8$

	MSE($\widehat{\beta}_0$)			MSE($\widehat{\beta}_1$)		
ρ	QIF	mQIF	RE	QIF	mQIF	RE
0	0.025	0.022	0.84	0.026	0.021	0.806
0.05	0.033	0.029	0.86	0.029	0.023	0.802
0.10	0.039	0.032	0.84	0.028	0.022	0.774
0.15	0.053	0.044	0.83	0.023	0.019	0.837
0.20	0.068	0.051	0.74	0.026	0.020	0.777
0.25	0.071	0.051	0.73	0.022	0.017	0.779
0.30	0.083	0.065	0.78	0.021	0.017	0.819

NOTE: Upper bound of ρ is 0.327 and $\beta = (0.5, -0.5)'$.

Table 3.21: MSE and RE of mQIF with both covariates, $t = 4$

α	MSE($\hat{\beta}_0$)	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)
0.00	0.608 (0.822)	0.106 (0.733)	0.140 (0.910)
0.05	0.512 (0.830)	0.098 (0.713)	0.152 (0.896)
0.10	0.518 (0.802)	0.093 (0.668)	0.154 (0.968)
0.15	0.511 (0.840)	0.097 (0.700)	0.166 (0.960)
0.20	0.525 (0.844)	0.101 (0.708)	0.162 (0.945)
0.25	0.453 (0.859)	0.089 (0.665)	0.181 (0.894)
0.30	0.491 (0.890)	0.094 (0.738)	0.167 (0.935)
0.35	0.473 (0.843)	0.099 (0.615)	0.169 (0.963)

NOTE: Upper bound of ρ is 0.365 and $\beta = (0.1, 0.2, -0.5)'$.
RE are given in parentheses.

Table 3.22: MSE and RE of mQIF with both covariates, $t = 8$

α	MSE($\hat{\beta}_0$)	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)
0.00	0.100 (0.757)	0.005 (0.764)	0.025 (0.668)
0.05	0.111 (0.764)	0.005 (0.715)	0.029 (0.691)
0.10	0.111 (0.762)	0.005 (0.759)	0.025 (0.667)
0.15	0.109 (0.811)	0.005 (0.744)	0.027 (0.680)
0.20	0.110 (0.751)	0.005 (0.672)	0.024 (0.653)

NOTE: Upper bound of ρ is 0.232 and $\beta = (0.1, 0.2, -0.5)'$.
RE are given in parentheses.

which means that there is no model misspecification for mQIF estimator. Since the correlation structure under first order Markov chain model is AR(1), it is a misspecified model for these simulated data. The mean square errors are calculated repeating the process 1000 times. The relative efficiency is calculated as

$$RE = \frac{\text{MSE of MC estimate}}{\text{MSE of mQIF estimate}}.$$

The results are displayed in Tables 3.23 and 3.24. We can see from the tables the relative efficiencies are close to 1, which indicates that the two approaches are almost equivalent for different value of ρ. The results show that MC estimator is quite robust and efficient as the mQIF estimator.

III.9 Real data examples

In this section, we apply QIF, mQIF, Markov chain and GEE methods of estimation on real data given in Chapter I.

Example 3.1. Six city data.

We applied the MC method first to the six city data and calculated the feasible bounds on the correlation, and found the range to be (-0.136, 0.927), which is pretty wide. For this data the GEE estimate of the correlation is 0.40 which is within the feasible range. Recall that QIF and mQIF methods do not give an estimate of the

Table 3.23: MSE and RE of QIF and mQIF with one covariate, $t = 8$

	$\text{MSE}(\widehat{\beta}_0)$			$\text{MSE}(\widehat{\beta}_1)$		
ρ	MC	mQIF	RE	MC	mQIF	RE
0	0.021	0.021	0.979	0.022	0.022	0.993
0.05	0.023	0.024	0.955	0.021	0.020	1.003
0.10	0.035	0.036	0.986	0.022	0.022	1.007
0.15	0.040	0.041	0.969	0.021	0.022	0.956
0.20	0.047	0.048	0.983	0.019	0.019	0.975
0.25	0.058	0.059	0.985	0.018	0.021	0.860
0.30	0.066	0.065	1.016	0.015	0.017	0.863

NOTE: Upper bound of ρ is 0.313 and $\beta = (0.5, -0.5)'$.

Table 3.24: MSE and RE of mQIF with both covariates, $t = 8$

α	$\text{MSE}(\hat{\beta}_0)$	$\text{MSE}(\hat{\beta}_1)$	$\text{MSE}(\hat{\beta}_2)$
0.00	0.094 (0.951)	0.004 (0.943)	0.033 (0.894)
0.05	0.096 (0.951)	0.004 (0.923)	0.030 (0.814)
0.10	0.111 (0.948)	0.004 (0.918)	0.033 (0.879)
0.15	0.101 (0.985)	0.004 (0.953)	0.031 (0.884)
0.20	0.104 (0.946)	0.004 (0.953)	0.031 (0.830)

NOTE: Upper bound of ρ is 0.216 and $\beta = (0.1, 0.2, -0.5)'$.
RE are given in parentheses.

correlation parameter. The results for the six city data using the four methods of estimation are shown in Table 3.25. The estimates and standard errors are very similar for the four methods. They all show that the maternal smoking habit is not a significant factor for children's respiratory illness.

Table 3.25: Analysis of parameter estimates for Six-city data

(i) Parameter	Est.	SE	p-value	(ii) Parameter	Est.	SE	p-value
Intercept	-1.917	0.120	<0.001	Intercept	-1.918	0.116	<0.001
Age	-0.147	0.059	0.0122	Age	-0.147	0.056	0.0084
Smoking	0.287	0.190	0.1316*	Smoking	0.300	0.196	0.1262*
Age×Smoking	0.078	0.090	0.3840	Age×Smoking	0.076	0.094	0.4176
(iii) Parameter	Est.	SE	p-value	(iv)Parameter	Est.	SE	p-value
Intercept	-1.921	0.110	<0.001	Intercept	-1.920	0.120	<0.001
Age	-0.152	0.070	0.0292	Age	-0.147	0.059	0.0134
Smoking	0.295	0.171	0.0849*	Smoking	0.295	0.190	0.1201*
Age×Smoking	0.112	0.058	0.4394	Age×Smoking	0.082	0.091	0.3688
Est.corr	Feasible range			Est.corr			
0.384	[-0.136,0.927]			0.400			

NOTE: (i) QIF (ii) mQIF (iii) MC and (iv) GEE

Example 3.2. Hamilton's depression study

We consider next the analysis of the Ham-D data. In the analysis we included only those patients with all 8 measurements so that the data is balanced. The analysis using the four methods is displayed in Table 3.26. For this data, the range of the correlation parameter calculated fitting the MC method is (-0.043, 0.632). It is also wide but the interval is not as broad as the six-city data. We can see from the table that the estimates of correlation parameter fall within the range for both MC and GEE methods, and both the methods show that treatment is significant. But for QIF and mQIF, the p-values are larger than 0.05 indicating that treatment is not a

significant factor, which we believe is an erroneous conclusion.

Table 3.26: Analysis of parameter estimates for Ham-D data

(i) Parameter	Est.	SE	p-value	(ii) Parameter	Est.	SE	p-value
Intercept	-6.919	0.382	<0.001	Intercept	-6.789	0.376	<0.001
Treatment	0.276	0.262	0.2839*	Treatment	0.280	0.261	0.2839*
Time	1.781	0.096	<0.001	Time	1.732	0.097	<0.001
(iii) Parameter	Est.	SE	p-value	(iv) Parameter	Est.	SE	p-value
Intercept	-4.164	0.276	<0.001	Intercept	-4.298	0.232	<0.001
Treatment	0.432	0.204	0.0343*	Treatment	0.487	0.213	0.0222*
Time	0.917	0.054	<0.001	Time	0.949	0.057	<0.001
Est.corr	Fréchet bounds			Est.corr			
0.499	[-0.043,0.632]			0.441			

NOTE: (i) QIF, (ii) mQIF, (iii) MC and (iv) GEE

CHAPTER IV

HIGH ORDER MARKOV CHAIN MODELS

IV.1 Introduction

We have discussed in Chapter II, a first order Markov chain model for dependent binary observations. A potential approach for constructing higher order Markov chain models is through the use of copulas, which are multivariate distributions with uniform margins. In this chapter we give an introduction to copulas and show how can they be used to construct multivariate binary distributions with specified marginal means and dependence structure.

IV.2 Introduction of copulas

Constructing multivariate distributions using copulas has become popular in recent years. The motivation for the development of copula methods is rooted in the attempt of constructing multivariate distributions with given non-normal marginal distributions. The copula method is based on the idea that the distribution function could be used to convert the marginal distribution to a uniform distribution, and then a joint distribution with specified marginal distribution can be obtained using a multivariate distribution with uniform marginals. There are many families of copulas that differ in the type of dependence they exhibit. In this section we give a brief description of copulas.

Definition 4.1. *A t-dimensional copula is a function C from $[0,1]^t \to [0,1]$ with the following properties:*

1. $C(1, 1, \cdots, 1, u_i, 1, \cdots, 1) = u_i$ *for all* $i = 1, 2, \cdots, t$.
2. $C(u_1, \cdots, u_t) = 0$ *if at least one* $u_i = 0$ *for* $i = 1, 2, \cdots, t$.
3. *For all* $u_{i_1} < u_{i_2}$, *where* $i = 1, 2, \cdots, t$,

$$\sum_{j_1=1}^{2}\sum_{j_2=1}^{2}\cdots\sum_{j_t=1}^{2}(-1)^{j_1+j_2+\cdots+j_t}C(u_{1_{j_1}}, u_{2_{j_2}}, \cdots, u_{t_{j_t}}) \geq 0.$$

The following theorem due to Sklar (1959) is fundamental to the development of copulas. Sklar's theorem elucidates the role that copulas play in the relationship between multivariate distribution functions and their univariate margins.

Theorem 4.1. Sklar's Theorem. *Let H be a joint cumulative distribution function with margins $F_1, F_2, \cdots, F_t$. Then there exists a t-dimensional copula C such for all $y_1, y_2, \cdots, y_t \in \Re$,*

$$H(y_1, y_2, \cdots, y_t) = C(F_1(y_1), F_2(y_2), \cdots, F_t(y_t)).$$

If $F_1, F_2, \cdots, F_t$ are continuous distribution functions, then C is unique; otherwise C is uniquely determined on $range(F_1) \times range(F_1) \times \cdots \times range(F_t)$.

Lemma 4.1. Fréchet-Hoeffding Bounds. *If C is a t-dimensional copula, then for all $u_i \in [0,1]$, $i = 1, 2, \cdots, t$,*

$$C_L(u_1, u_2, \cdots, u_t) \leq C(u_1, u_2, \cdots, u_t) \leq C_U(u_1, u_2, \cdots, u_t),$$

where C_L and C_U are defined as

$$\begin{aligned} C_L(u_1, u_2, \cdots, u_t) &= \max(0, u_1 + u_2 + \cdots + u_t - (t-1)), \\ C_U(u_1, u_2, \cdots, u_t) &= \min(u_1, u_2, \cdots, u_t). \end{aligned}$$

C_U is a copula for any t, but C_L is a copula only when $t = 2$.

IV.3 Examples of bivariate copulas

In this section, we present some well known bivariate copulas.

Example 4.1. (Independence Copula). *The function given by*

$$C(u, v) = uv$$

is the bivariate Independence copula. Its density is simply $c(u, v) = 1$.

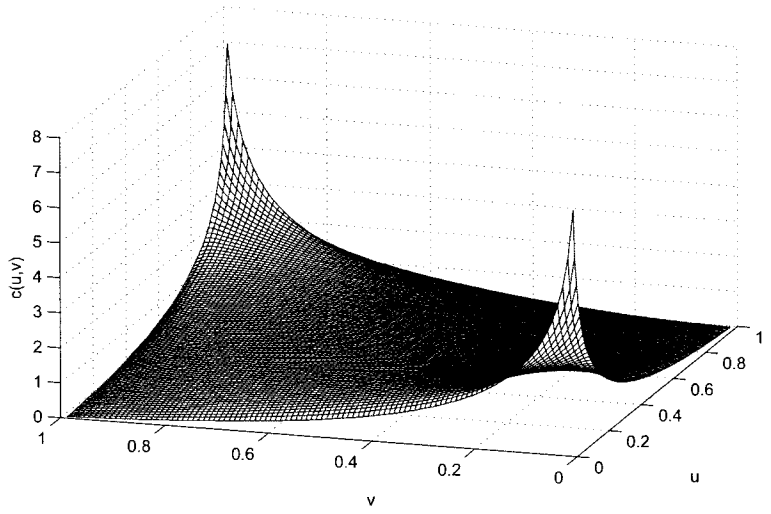


Figure 4.1: Density function of bivariate Gaussian copula

Example 4.2. (Bivariate Gaussian copula). *For* $0 \leq \delta \leq 1$, $C(u, v; \delta) = \Phi_\delta(\Phi^{-1}(u), \Phi^{-1}(v))$, *where* Φ *is the standard normal distribution function, and* Φ_δ *is the bivariate standard normal distribution function with correlation* δ. *Let* $x = \Phi^{-1}(u)$ *and* $y = \Phi^{-1}(v)$, *the density function is given by*

$$c(u, v; \delta) = (1 - \delta^2)^{-1/2} \exp\{-\frac{1}{2}(1 - \delta^2)^{-1}[x^2 + y^2 - 2\delta xy]\} \exp[\frac{1}{2}(x^2 + y^2)].$$

Example 4.3. (Frank Copula). *For* $0 \leq \delta \leq \infty$, $\eta = 1 - e^{-\eta}$,

$$C(u, v; \delta) = -\eta^{-1} \log([\eta - (1 - e^{-\delta u})(1 - e^{-\delta v})]/\eta).$$

The density is

$$c(u, v; \delta) = \delta\eta e^{-\delta(u+v)}/[\eta - (1 - e^{-\delta u})(1 - e^{-\delta v})]^2.$$

Example 4.4. (Gumbel Copula). *Let* $\tilde{u} = -\log(u)$, $\tilde{v} = -\log(v)$, *for* $0 \leq \delta \leq \infty$,

$$C(u, v; \delta) = \exp\{-(\tilde{u}^\delta + \tilde{v}^\delta)^{1/\delta}\}.$$

The density is

$$c(u, v; \delta) = C(u, v; \delta)(uv)^{-1} \frac{(\tilde{u}\tilde{v})^{\delta-1}}{(\tilde{u}^\delta + \tilde{v}^\delta)^{2-1/\delta}} [(\tilde{u}^\delta + \tilde{v}^\delta)^{1/\delta} + \delta - 1].$$

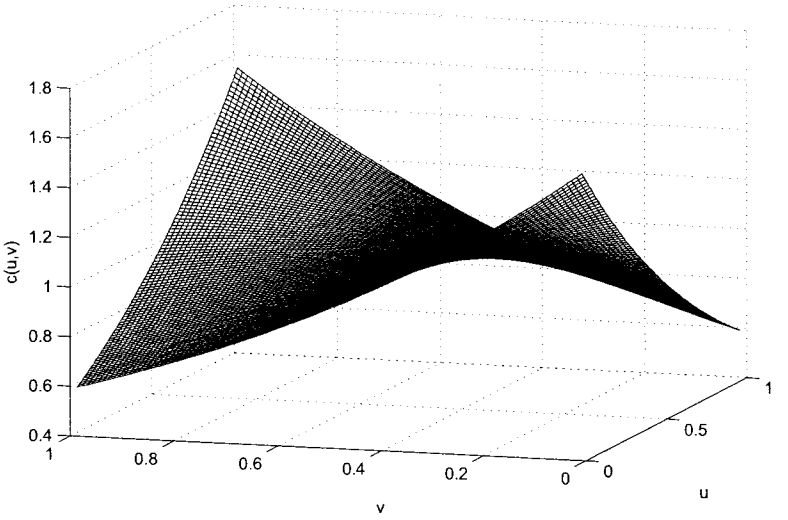


Figure 4.2: Density function of Frank copula

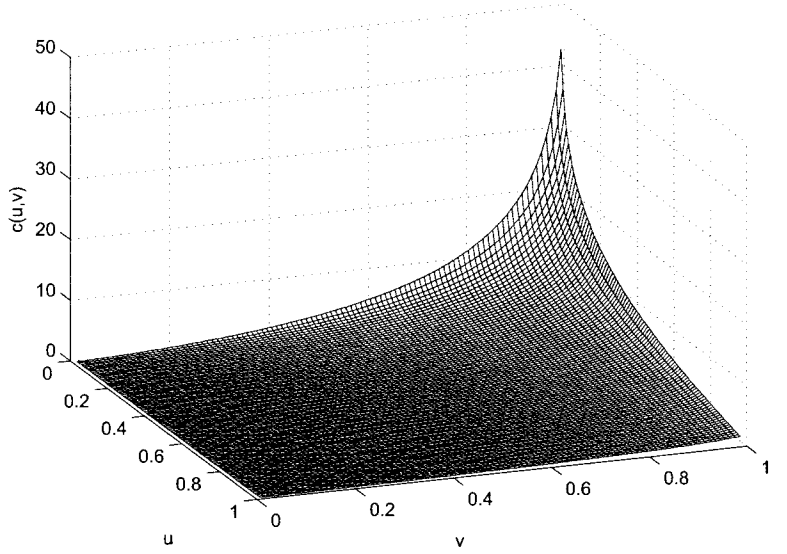


Figure 4.3: Density function of Gumbel copula

In a multivariate copula model, the cumulative distribution function of a random vector $Y = (Y_1, \cdots, Y_t)$ is given by

$$F(y) = C(F_1(y_1), \cdots, F_t(y_t)), \tag{4.3.1}$$

where C is a t-dimensional copula and F_i is marginal cumulative distribution function of Y_i. If Y is continuous, then its probability density function is

$$f(y) = \prod_{i=1}^{t} f_i(y_i) c(F_1(y_1), \cdots, F_t(y_t)), \tag{4.3.2}$$

where f_i is marginal probability density function of Y_i and

$$c(F_1(y_1), \cdots, F_t(y_t)) = \frac{\partial^t C(u_1, \cdots, u_t)}{\partial u_1 \cdots \partial u_{t-1}} \tag{4.3.3}$$

is the density of copula C.

For discrete margins the multivariate probability mass function is given by

$$P(Y = y) = \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_t=1}^{2} (-1)^{\sum_{k=1}^{t} j_k} C(u_{1j_1}(y_1), \cdots, u_{1j_t}(y_t)), \tag{4.3.4}$$

where $u_{i1}(y_i) = F_i(y_{i-})$ and $u_{i2}(y_i) = F_i(y_i)$. Here $F_i(y_{i-})$ is the left-hand limit of F_i at y_i, which is equal to $F_i(y_i - 1)$ when the support of F_i is the set of integers.

In particular, suppose Y_i is a binary variable with $P(Y_i = 1) = p_i$. The distribution function of Y_i is

$$F_i(y_i) = \begin{cases} 0 & y_i < 0 \\ 1 - p_i & 0 \le y_i < 1 \\ 1 & y_i \ge 1 \end{cases}$$

If $t = 3$, according to (4.3.4), the joint probability function is of the form

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \;=\; & C(u_{12}, u_{22}, u_{32}) - C(u_{12}, u_{22}, u_{32}) - C(u_{12}, u_{22}, u_{32}) \\ & - C(u_{12}, u_{22}, u_{32}) + C(u_{12}, u_{22}, u_{32}) + C(u_{12}, u_{22}, u_{32}) \\ & + C(u_{12}, u_{22}, u_{32}) - C(u_{12}, u_{22}, u_{32}). \end{aligned} \tag{4.3.5}$$

The eight probabilities can be written in a table form.

Table 4.1: Trivariate joint probability using copula

y_1	y_2	y_3	$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$
1	1	1	$p_1 + p_2 + p_3 - 2 + C(1, 1-p_2, 1-p_3) + C(1-p_1, 1, 1-p_3)$ $+C(1-p_1, 1-p_2, 1) - C(1-p_1, 1-p_2, 1-p_3)$
1	1	0	$1 - p_3 - C(1, 1-p_2, 1-p_3) - C(1-p_1, 1, 1-p_3)$ $+C(1-p_1, 1-p_2, 1-p_3)$
1	0	1	$1 - p_2 - C(1, 1-p_2, 1-p_3) - C(1-p_1, 1-p_2, 1)$ $+C(1-p_1, 1-p_2, 1-p_3)$
0	1	1	$1 - p_1 - C(1-p_1, 1, 1-p_3) - C(1-p_1, 1-p_2, 1)$ $+C(1-p_1, 1-p_2, 1-p_3)$
1	0	0	$C(1, 1-p_2, 1-p_3) - C(1-p_1, 1-p_2, 1-p_3)$
0	1	0	$C(1-p_1, 1, 1-p_3) - C(1-p_1, 1-p_2, 1-p_3)$
0	0	1	$C(1-p_1, 1-p_2, 1) - C(1-p_1, 1-p_2, 1-p_3)$
0	0	0	$C(1-p_1, 1-p_2, 1-p_3)$

IV.4 Markov chain based on copulas

A first order Markov chain with given univariate binary margins can be constructed from a bivariate copula. This is a generalization of the normal AR(1) time series, since the normal AR(1) time series can be obtained using the bivariate normal copula and univariate normal margins. A Markov chain of second order, can be constructed from a trivariate copula which has the property that the (1,2) and (2,3) bivariate margins are the same. This generalizes the normal AR(2) time series. Extensions to Markov chains of higher order require multivariate copulas with constraints on the lower dimensional margins.

The description of a stationary Markov chain with discrete state space based on a bivariate copula $C(u, v)$ is as follows. Suppose that $\{Y_t, t = 1, 2, \ldots\}$ takes non-negative integers values. Let $F(y_t)$ and $f(y_t)$ be the cdf and pmf of Y_t, respectively.

The transition distribution function from Y_{t-1} to Y_t is given by

$$\begin{aligned} H(y_t|y_{t-1}) &= P(Y_t \leq y_t|Y_{t-1} = y_{t-1}) \\ &= [C(F(y_{t-1}), F(y_t)) - C(F(y_{t-1} - 1), F(y_t))]/f(y_{t-1}). \end{aligned}$$

Escarela at el. (2009) have discussed a fully parametric first order autoregressive model for longitudinal binary data using bivariate Gaussian copula.

More generally, stationary Markov chains of order $m-1$ can be constructed from an m-variate copula C that satisfies the following conditions: (i) the bivariate margins C_{ij} are such that $C_{i,i+l} = C_{1,1+l}$, $l = 1, \cdots, m-2, i = 2, \cdots, m-l$; (ii) the higher-dimensional margins are such that $C_{i_1,\cdots,i_k} = C_{1,i_2-i_1+1,\cdots,i_k-i_1+1}$ for $1 \leq i_1 < \cdots < i_k \leq m$, $3 \leq k \leq m-1$; and (iii) C is differentiable in its first $m-1$ arguments. For second order Markov chains, these conditions simply become $C_{12} = C_{23}$.

If $F_{1,\cdots,m} = C(F, \cdots, F)$ is an m-variate cdf, such that C is a copula with the above properties, then the transition cdf of the stationary Markov chain is

$$H(y_t|y_{t-m+1}, \cdots, y_{t-1}) = \frac{a(F(y_{t-m+1}), \cdots, F(y_t))}{b(F(y_{t-m+1}), \cdots, F(y_t))}, \tag{4.4.1}$$

where

$$a(F(y_{t-m+1}), \cdots, F(y_t)) = \frac{\partial^{m-1} C}{\partial u_1 \cdots \partial u_{m-1}}(u), \tag{4.4.2}$$

and

$$b(F(y_{t-m+1}), \cdots, F(y_t)) = \frac{\partial^{m-1} C_{1\cdots m-1}}{\partial u_1 \cdots \partial u_{m-1}}(u_1, \cdots, u_{m-1}), \tag{4.4.3}$$

with $C_{1\cdots m-1}$ be an $(m-1)$-dimensional marginal of C.

IV.4.1 A feasible family of copula

Here, we discuss a family of copula that can be used to construct higher order Markov chain. Let ψ be a Laplace transformation and $l_i = (v_i + t - 1)^{-1}$, where v_i's are constants. Consider the copula

$$C(\mathbf{u}) = \psi\left(-\sum_{i<j} \log K_{ij}(e^{-l_i\psi^{-1}(u_i)}, e^{-l_j\psi^{-1}(u_j)}) + \sum_{i=1}^{t} v_i l_i \psi^{-1}(u_i)\right), \tag{4.4.4}$$

where $K_{ij}, 1 \leq i < j \leq t$ are bivariate copulas. An interpretation is that the Laplace transformation ψ leads to a minimal level of pairwise dependence, the copulas K_{ij} add some individual pairwise dependence beyond the global dependence, and the parameters v_i lead to bivariate and multivariate asymmetry.

A useful special case of (4.4.4) is the following. Let $t = 3$, $K_{1,3}(u,v) = uv$, $v_1 = v_3 = -1, v_2 = 0$, $K_{12}(u,v) = K_{23}(u,v) = K(u,v)$. Then (4.4.4) becomes

$$\begin{aligned} C(\mathbf{u}) &= \psi\Big(-\log K(e^{-\psi^{-1}(u_1)}, e^{-0.5\psi^{-1}(u_2)}) \\ &\quad -\log K(e^{-\psi^{-1}(u_3)}, e^{-0.5\psi^{-1}(u_2)})\Big). \end{aligned} \tag{4.4.5}$$

Here the (1,2) and (3,2) bivariate margins of (4.4.5) are the same and are more concordant than the (1,3) margin. Hence this model would be appropriate for generating a second order Markov chain. With $j = 1, 3$, the bivariate margins of (1,2) and (3,2) are

$$\psi\Big(-\log K(e^{-\psi^{-1}(u_j)}, e^{-0.5\psi^{-1}(u_2)}) + 0.5\psi^{-1}(u_2)\Big). \tag{4.4.6}$$

We can choose K as the bivariate Gumbel copula with parameter $\delta \geq 1$, and ψ be $\psi(s) = \exp(-s^{1/\theta})$ with parameter $\theta \geq 1$. For this ψ, we have $\psi^{-1}(s) = (-\log(s))^{\theta}$. Then (4.4.5) can be written as

$$C(\mathbf{u}) = \exp\left\{-[(z_1^{\theta})^{\delta} + (0.5z_2^{\theta})^{\delta}]^{1/\delta} - [(z_3^{\theta})^{\delta} + (0.5z_2^{\theta})^{\delta}]^{1/\delta}\right\}^{1/\theta}, \tag{4.4.7}$$

where $z_i = -\log(u_i)$, $i = 1, 2, 3$. The bivariate margins are

$$C_{i2}(u_i, u_2) = \exp\left\{-[(z_i^{\theta})^{\delta} + (0.5z_2^{\theta})^{\delta}]^{1/\delta} + 0.5z_2^{\theta}\right\}^{1/\theta}. \tag{4.4.8}$$

Further research along these lines will be pursued in the future.

CHAPTER V

DISCUSSION

There is a vast literature on the analysis of longitudinal and clustered binary data. Much of it is focused on marginal models, due to the difficulty in constructing proper likelihood models. Our first goal in this dissertation is to study multivariate binary distributions namely the first order Markov chain model and the multivariate probit model. The first order Markov chain model results in first order autoregressive correlation structure and is appropriate for analyzing longitudinal data. The multivariate probit model is useful to analyze both longitudinal and clustered binary data. For both models we studied maximum likelihood estimates and their asymptotic variances computed via Fisher information matrices. Large and small sample simulations show that the estimates are comparable in terms of efficiency and no one model is uniformly superior over the other model. Other multivariate binary models include the quadratic exponential model. We checked numerically that the quadratic exponential family generates the same probability distribution as the Markov chain model for the autoregressive correlation structure.

As mentioned earlier marginal models do not specify the complete distribution and are motivated by quasi-likelihood ideas. A popular parameter estimation procedure in marginal models is the generalized estimation equations method. Though popular this method has several drawbacks. We have used simulations to show that the generalized estimation equations method is less efficient when compared to the maximum likelihood estimates for the Markov chain model. The efficiency is poor when there is a high correlation in the data. An alternative to the generalized estimating equations is the quadratic inference function (QIF) approach due to Qu et al. (2000). This method bypasses estimating the correlation parameter. We have shown that this method is also less efficient than the maximum likelihood estimates for the Markov chain model. We also introduced a modified quadratic inference method (mQIF) and showed using simulations this method has better efficiency when compared to the original QIF method.

Finally, we discussed copula models. These models are potential pathways for constructing high order Markov chain models.

REFERENCES

Albert, P. S., and Waclawiw, M. A. (1998), "A two-state Markov chain for heterogeneous transtional data: A quasi-likelihood approach," *Statistics in Medicine*, 17, 1481-1493.

Anderson, T. W., and Goodman, A. (1957), "Statistical inference about Markov chains," *The Annals of Mathematical Statistics*, 28, 89-110.

Azzalini, A. (1994), "Logistic regression for autocorrelated data with application to repeated measures," *Biometrika*, 81, 767-75.

Chaganty, N. R., and Joe, H. (2004), "Efficiency of generalised estimating equations for binary resoonses," *Journal of the Royal Statistical Society, Ser. B*, 66, 851-60.

Chaganty, N. R., and Joe, H. (2006), "Range of correlation matrices for dependent Bernoulli random variables," *Biometrika*, 93, 197-206.

Crowder, M.J., and Hand, D.J. (1990), *Analysis of Repeated Measures*, Chapman and Hall, London.

Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer-Verlagnn, New York.

Devore, J. L. (1976), "A note on the estimation of parameters in a bernoulli model with dependence". *The Annals of Statistics*, Vol.4, No.5, 990-992.

Drezner Z. (1990), "Approximations to the multivariate normal intergral," *Communications in Statistics: Simulation and Computation*, Vol 19, No 2, 527-534.

Emrich, L. J., and Piedmonte, M. R. (1991), "A method for generating high-dimensional multivatiate binary variates," *The American Statistician*, Vol.45, No.4, 302-304.

Fitzmaurice, G. M., and Laird, N. M. (1993), "A likelihood-based method for analysing longitudinal binary responses," *Biometrika*, 80, 141-151.

Frees, E. W. (2004). *Longitudinal and panel data.* Cambridge University Press.

Hansen, L. P. (1982), "Large sample properties of generalized method of moment estimators," *Econometrica*, 50, 1029-1054.

Joe, H. (1995), "Approximations to multivariate normal rectangle probabilites based on conditional expectations," *Journal of the American Statistical Association*, 90, 957-964.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapmam & Hall, London.

Kanter, M. (1975), "Autoregression for discrete processes mod 2," *Journal of Applied Probability*, 12, 371-375.

Klotz, J.(1973), "Statistical inference in Bernoulli trials with dependence," *The Annals of Statistics*, Vol.1, No.2, 373-379.

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

Muenz, L. R., and Rubinstein, L. V. (1985), "Markov models for covatiate dependence of binary sequences," *Biometrics*, 41, 91-101.

Nunèz-Anton, V., and Woodworth, G. G. (1994), "Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors," *Biometrika*, 77, 549-555.

Pan, W. (2001), "On the robust variance estimator of generalized estimating equations," *Biometrika* 88, 3, 901-906.

Qaqish, B. F. (2003), "A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations," *Biometrika* 90, 455-63.

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving generalized estimating equations using quadratic inference functions," *Biometrika*, 87, 4, 823-836.

Qu, A., and Song, P.X.-k. (2004), "Assessing robustness of generalised estimating equations and quadratic inference functions," *Biometrika*, 91, 447-459.

Qu, A., Lee, J. J., and Lindsay, B. G. (2008), "Model diagnostic tests for selecting informative correlation structure in correlated data," *Biometrika*, 95, 4, 891-905.

Sabo, R. T., and Chaganty, N. R. (2010), "What can go wrong when ignoring correlation bounds in the use of generalized estimating equations," *Statistics in Medicine*, Vol 29, Issue 24, 2501-2507.

Song, P.X.-k., Jiang, Z., Park, E., and Qu, A. (2009), "Quadratic inference functions in marginal models for longitudinal data," *Statistics in Medcine*, 28, 3683-3696.

APPENDIX

SAS PROGRAMS

A.1 SAS Macro TMMLE

```
% let ROOT = % STR(file location);
FILENAME Hession "&ROOT \ Hession.sas";
options ls = 84 nodate;
% macro tmmle(data=, yvar=, xvar=, id=, fun=logit, outpar=,outmle=);
%macro con(x,y,id,func);
   %global y1;
   %global x1;
   %global id1;
   %global lfun;
   %let y1=y;
   %let x1=x;
   %let id1=id;
   %let lfun=func;
%mend con;

ods listing;
proc genmod data=&data descend;
   class &id;
   model &yvar=&xvar / link=&fun dist=BIN;
   repeated subject = &id / type=ar covb corrw;
   ods output GEEEmpPEst = GEEEmpPEst GEEWCorr=GEEWCorr;
run;
ods listing close;

ods listing;
proc iml;
%include Hession;
use GEEEmpPEst;
```

```
read all varestimate into ibeta;
use GEEWCorr;
read all varCol2 into cor;
use &data;
read all var&xvar into x;
read all var&yvar into y;
read all var&id into id;
k = ncol(x)+1;
%if &fun=logit %then func='logit';
%else func = 'probit';
%con(x,y,id,func);

Start U(u, v);
   temp = sqrt((1-u)*v/u/(1-v)) || sqrt(u*(1-v)/v/(1-u));
   return(min(temp));
Finish;

Start L(u, v);
   temp = (-sqrt((1-u)*(1-v)/u/v)) || (- sqrt(u*v/(1-v)/(1-u)));
   return(max(temp));
Finish;

start arbound(p);
   t = max(ncol(p), nrow(p));
   if t=1 then bdi=-1 1;
   else do;
      temp = j(t-1,2,0);
      do i=1 to (t-1);
         temp[i,1]=L(p[i],p[i+1]);
         temp[i,2]=U(p[i],p[i+1]);
      end;
      bdi = max(temp[,1])|| min(temp[,2]);
   end;
   return(bdi);
```

```
finish arbound;

start irhobound(x,beta,nrep) global(&lfun);
   xx= j(nrow(x),1)|| x;
   u = xx*beta;
   fun = &lfun;
   if fun = 'logit' then p = exp(u)/(1+exp(u));
   else p = probnorm(u);
   index = 0;
   n = max(nrow(nrep),ncol(nrep));
   bd = j(n,2,0);
   do i =1 to n;
      ti = nrep[i];
      if ti=1 then pi = p[index+ti];
      else pi = p[(index+1):(index+ti)];
      bd[i,]=arbound(pi);
      index =index+ti;
   end;
   L=max(bd[,1]);
   U=min(bd[,2]);
   irb = L//U;
   return(irb);
finish irhobound;

start betacond(m);
   temp = j(1,m,.)//j(1,m,.);
   return(temp);
finish betacond;


start Lhf(theta) global(&x1,&y1,&id1,&lfun);
   k = max(ncol(theta),nrow(theta));
   beta = theta[1:(k-1)];
   rho = theta[k];
   x = &x1;
```

```
y = &y1;
id = &id1;
fun = &lfun;
xi = j(nrow(x),1)——x;
n = nrow(y);
u = xi*beta;
if fun = 'logit' then p = exp(u)/(1+exp(u));
else p = probnorm(u);
q=1-p;
s = sqrt(p#(1-p));
nrep=1;
nsub=1;
do i = 2 to n;
   if id[i]=id[i-1] then nrep[nsub]=nrep[nsub]+1;
   else do;
      nrep=nrep//1;
      nsub=nsub+1;
   end;
end;

nnn = max(nrow(nrep),ncol(nrep));
bd = j(nnn,2,0);
indexi = 0;
do i =1 to nnn;
   ti = nrep[i];
   if ti=1 then pi = p[indexi+ti];
   else pi = p[(indexi+1):(indexi+ti)];
   bd[i,]=arbound(pi);
   indexi =indexi+ti;
end;
L=max(bd[,1]);
U=min(bd[,2]);
rhobd = L//U;
if(rhobd[1]<=rho & rho<=rhobd[2]) then rho=rho;
```

```
   else rho=(rhobd[2]+rhobd[1])/2;

      mle = 0;
   index =0;
   do i = 1 to nsub;
      fi = 0;
      ti = nrep[i];
      yi = y[(index+1):(index+ti)];
      pi = p[(index+1):(index+ti)];
      qi = q[(index+1):(index+ti)];
      si = s[(index+1):(index+ti)];
      fi = yi[1]*log(pi[1]) + (1-yi[1])*log(1-pi[1]);
      if ti^=1 then
      do;
         do j = 2 to ti;
            pij = pi[j]**yi[j]*qi[j]**(1-yi[j])+
            (-1)**(yi[j]+yi[j-1])*rho*si[j]*si[j-1]/(pi[j-1]**yi[j-1]*qi[j-1]**(1-yi[j-1]));
            fi = fi + log(pij);
         end;
      end;
      else fi=fi;
      mle = mle + fi;
      index = index + ti;
   end;
   return(mle);
finish Lhf;


start getname;
   parameter = j(k+1,1,char(20));
   parameter[1,1] = "intercept";
   do i=1 to k;
      parameter[i+1,1] = scan("&xvar", i);
   end;
   parameter[k+1,1] = "rho";
```

```
finish;


start outdata;
   Test = "MLE";
   DF = df;
   Statistic = stat;
   AIC = -2*stat+2*(k-1);
   BIC = -2*stat+(k-1)*log(nsub);


   %if %length(&outmle)>0 %then %do;
      create &outmle var Test Statistic DF AIC BIC;
      append;
   %end;

   %if %length(&outpar)>0 %then %do;
      estimate = ntheta;
      stderr = stderror;
      create &outpar var parameter estimate stderr Z pvalue;
      append;
   %end;
finish;


start main;
   n = nrow(y);
   nrep=1;
   nsub=1;
   df=k-1;
   do i = 2 to n;
      if id[i]=id[i-1] then nrep[nsub]=nrep[nsub]+1;
      else do;
         nrep=nrep//1;
         nsub=nsub+1;
      end;
```

```
end;

optn = {1,0};
rhobd = irhobound(x,ibeta,nrep);
irho=(rhobd[2]-rhobd[1])/2;
cond = betacond(k)——rhobd;
itheta=ibeta//irho;
do until(rhobd[1]¡=irho & irho¡=rhobd[2]);
   call NLPNRR(rc, ntheta, "Lhf", itheta, optn,cond);
   ibeta = ntheta[1:k];
   rhobd = irhobound(x,ibeta,nrep);
   irho = ntheta[k+1];
   itheta = ibeta//irho;
   cond = betacond(k)——rhobd;
end;
beta = ntheta[1:k];
rho = ntheta[k+1];
xx = j(nrow(x),1)——x;
u=xx*beta;
if func='logit' then do;
   p = exp(u)/(1+exp(u));
   dp = p#(1-p);
   d2p = (exp(u)-exp(2*u))/(1+exp(u))##3;
end;
else do;
   p = probnorm(u);
   dp = normpdf(u);
   d2p = devnormpdf(u);
end;
scnddev=0;
scnddevj=0;
index = 0;
do i = 1 to nsub;
   ti = nrep[i];
```

```
      yi = y[index+1:index+ti];
      xi = xx[index+1:index+ti,];
      pi = p[index+1:index+ti];
      dpi = dp[index+1:index+ti];
      d2pi = d2p[index+1:index+ti];
      scnddev = scnddev + devP2nd(yi,xi,pi,dpi,d2pi,rho,k);
      index = index + ti;
   end;
   hess = scnddev;
   var = -ginv(scnddev);
   print hess[label='Hession Matrix'];
   print var[label='Covariance Matrix'];
   stderror = sqrt(vecdiag(var));
   par = t(ntheta);
   Z = par/stderror;
   pvalue = 2*(1-probnorm(abs(Z)));
   print par stderror Z pvalue;
   stat = Lhf(ntheta);
   rangeofrho=irhobound(x,beta,nrep);
   print rangeofrho;

finish main;
run main;
run getname;
run outdata;
quit;
ods listing close;
%mend armle;
```

VITA

Weiming Yang
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529

Education

Ph.D.	Old Dominion University, Norfolk, VA. (December 2010)
	Major: Computational and Applied Mathematics (Statistics)
MS	Old Dominion University, Norfolk, VA. (Summer 2009)
	Major: Computational and Applied Mathematics (Statistics)
MS	University of Toledo, Toledo, OH. (Summer 2005)
	Major: Applied Mathematics
BS	Chengdu University of Technology, China. (Summer 2003)
	Major: Applied Mathematics

Experience

Research Assistant, ISTART Lab (08/2007 - 08/2008)
Old Dominion University, Norfolk, VA

Teaching Assistant (08/2006 - 08/2007 & 08/2008 - 12/2010)
Old Dominion University, Norfolk, VA

Publications

Yang, W. and Chaganty, N. R. , *"Maximum likelihood estimation of a transition model for binary longitudinal data"*, under preparation.

Yang, W. and Chaganty, N. R. , *"On the use of Quadratic Inference Function for binary longitudinal data"*, under preparation.

Case, R. and **Yang, W.**, *"A Study to Examine Differences Between In Person and Online Survey Data Collection Methodologies"*, Sport Management International Journal, Vol. 6, No. 2, 2009, pp. 5-20.

Yan. G, **Yang, W.**, Shaner, E. F. and Rawat, D. , *"Intrusion-Tolerant Location Information Services inIntelligent Vehicular Networks"*, Communications in Computer and Information Science, accepted.

Typeset using LaTeX.