A SIMULATION STUDY ON MARGINALIZED TRANSITION RANDOM EFFECTS MODELS FOR MULTIVARIATE LONGITUDINAL BINARY DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZERRİN YALÇINÖZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

MAY 2008

Approval of the thesis:

A SIMULATION STUDY ON MARGINALIZED TRANSITION RANDOM EFFECTS MODELS FOR MULTIVARIATE LONGITUDINAL BINARY DATA

submitted by **Zerrin Yalçınöz** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan ÖZGEN
Dean, Graduate School of **Natural and Applied Sciences** ____________________

Prof. Dr. Ali UZUN
Head of Department, **Statistics Department** ____________________

Dr. Özlem İLK
Supervisor, **Statistics Department, METU** ____________________

Examining Committee Members:

Prof. Dr. Gerhard Wilhelm WEBER
Institute of Applied Mathematics, METU ____________________

Dr. Özlem İLK
Department of Statistics, METU ____________________

Assoc. Prof. İnci BATMAZ
Department of Statistics, METU ____________________

Dr. Ceylan YOZGATLIGİL
Department of Statistics, METU ____________________

Dr. Zeynep KALAYLIOĞLU
Department of Statistics, METU ____________________

Date: 28.05.2008

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: ZERRİN YALÇINÖZ

Signature:

ABSTRACT

A SIMULATION STUDY ON MARGINALIZED TRANSITION RANDOM EFFECTS MODELS FOR MULTIVARIATE LONGITUDINAL BINARY DATA

YALÇINÖZ, Zerrin
M.S., Department of Statistics
Supervisor: Dr. Özlem İLK

May 2008, 45 Pages

In this thesis, a simulation study is held and a statistical model is fitted to the simulated data. This data is assumed to be the satisfaction of the customers who withdraw their salary from a particular bank. It is a longitudinal data which has bivariate and binary response. It is assumed to be collected from 200 individuals at four different time points. In such data sets, two types of dependence -the dependence within subject measurements and the dependence between responses- are important and these are considered in the model. The model is Marginalized Transition Random Effects Models, which has three levels. The first level measures the effect of covariates on responses, the second level accounts for temporal changes, and the third level measures the difference between individuals. Markov Chain Monte Carlo methods are used for the model fit. In the simulation study, the changes between the estimated values and true parameters are searched under two conditions, when the model is correctly specified or not. Results suggest that the better convergence is obtained with the full model. The third level which observes the individual changes is more sensitive to the model misspecification than the other levels of the model.

Keywords: Binary response, Bivariate response, Gibbs sampling, Hybrid markov chain steps, Model specification, Panel data, Simulation.

ÖZ

ÇOK DEĞİŞKENLİ VE İKİ SONUÇLU PANEL VERİSİ İÇİN MARJİNAL GEÇİŞLİ RASTGELE ETKİLER MODELLERİ ÜZERİNE BİR BENZETİM ÇALIŞMASI

YALÇINÖZ, Zerrin
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi: Dr. Özlem İLK

Mayıs 2008, 45 sayfa

Bu araştırmada, bir benzetim çalışması yapılmış ve bu çalışmadan elde edilen veriye istatistiksel bir model uydurulmuştur. Benzetim çalışmasından elde edilen verinin, maaşını belirli bir bankadan çeken müşterilerin müşteri memnuniyeti olduğu varsayılmıştır. Verimiz, iki sonuçlu ve iki bağımlı değişkenli panel veridir. Verinin 200 bireyden dört değişik zaman dilimi için toplandığı varsayılmıştır. Bu şekildeki veri kümelerinde iki tür bağımlılık –birey içi ölçümleri arası bağımlılık, bağımlı değişkenler arası bağımlılık- önemlidir ve önerilen modelle bu tür bağımlılıklar gözönüne alınmıştır. Model, üç düzeyli Marjinal Geçişli Rastgele Etkiler Modelleridir. Bu modelin birinci düzeyinde bağımsız değişkenlerin bağımlı değişkenler üzerindeki etkisi ölçülürken, ikinci düzeyinde zamana bağlı değişimler dikkate alınır; üçüncü düzeyinde ise bireyler arasındaki değişiklikler ölçülür. Model uyumu için Markov Zinciri Monte Carlo yöntemi kullanılmıştır. Bu benzetim çalışmasında, beklenen değerler ve gerçek değerler arasındaki farklar, model doğru şekilde belirlendiğinde ve belirlenmediğinde araştırılmıştır. Araştırma sonuçlarına dayanarak, doğru model ile daha iyi bir yakınsama sağlandığı söylenebilir. Diğer düzeylere kıyasla, modelin bireyler arası farkları ölçen üçüncü düzeyinde modelin yanlış belirlenmesi olayına daha duyarlı tepkiler alınmıştır.

Anahtar Kelimeler: İki sonuçlu bağımlı değişken, İki değişkenli bağımlı değişken, Gibbs örneklemesi, Karışık markov zinciri adımları, Model belirleme, Panel veri, Benzetim.

To My Parents

ACKNOWLEDGEMENTS

First of all, I would like to give special thanks to my thesis supervisor, Dr. Özlem İlk for her professional support, guidance and encouragements which are invaluable for me during my research. At writing stage of thesis, I had difficult times and Dr. Özlem İlk stands by me. At the same time, I am working at the public sector and it was too tiring to work on thesis when I am working. Therefore, I again give my special thanks to Özlem İlk to tolerate this situation, to her sensitivity and patience.

I am also grateful to my friend Gül İnan who is a research assistant in Statistic department at METU. I want to express my deep appreciation for her help about my departmental works, for her lovely conversation and also for her friendship.

I owe my loving thanks to my fiancé Cevdet İlker Kocatepe, for his technical support about computer jobs and also for his encouragements. He always let me feel that he would always stand by me. I shared my distress and happiness about thesis works firstly with him.

But most of all I would like to thank my mother and father for their sacrificing and endless love which can be replaced by nothing in the world. I also would like to thank my cousin Zeynep Çiğdem Gökçınar for her support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Panel data, also known as longitudinal data, is the collection of the repeated measurements from the same subject observed over time (Diggle et al., 2002). It is used for observing the changes over time within individuals. It occurs frequently in fields such as social and medical sciences. For instance, the change in CD4+, a marker for HIV, is used for detecting infected people. As another example, following-up the same people over years, one can observe the human aging process. Many examples of longitudinal data are available in Chapter 2 of Verbeke and Molenberghs (2000) and Ilk (2004).

This type of data is usually complex to deal with. In longitudinal data, the measurements are not independent; on the contrary, they are collected from the same individuals at different time periods. Therefore, the change over time is observed. Moreover, if more than one response is measured, then the correlation between responses may occur. In this thesis, a simulation study is held on a statistical model which takes these two types of dependencies -the dependence within subject measurements and the dependence between responses- into account. The aim of this thesis is to assess the performance of the estimation procedure against model misspecification via a simulation study. Model misspecification is investigated by ignoring the effects of some independent variables, which have the highest effects on the model.

There is a wealth of literature on modeling longitudinal data. A comprehensive discussion of the linear models for longitudinal data can be found in Verbeke and Molenberghs (2000) and Diggle et al. (2002). Due to flexibility of normal distribution assumption, models for continuous responses are frequently proposed. However, modeling longitudinal data with binary response is usually more challenging. Ribaudo and Thompson (2002) built a three level model for multivariate longitudinal binary data in the context of quality of life data. They introduced dependence by random effects and directly modeled conditional covariate effects. In addition, Reboussin and Anthony (2001) studied on marginal models for multivariate longitudinal binary data.

In this thesis, Marginalized Transition Random Effects Models (MTREM) for multivariate longitudinal binary data is used to model the simulated data. This model is introduced by Ilk and Daniels (2007). MTREM is an extension of the two models that is developed by Heagerty (1999, 2002). Both of these models, introduced for univariate binary data, have two levels. On the first level, a marginal logistic regression model is used to explain the average response. The second level consists of a random effects model in 1999 paper and transition model in 2002 model. MTREM has three levels. The first level of the model is marginal mean model that explains the mean response, the second level of the model is transition model that explains within-subject time dependence for each response, and the third level of the model is random effects model for the multivariate response structure at each time.

In line with the aim of this study, Markov Chain Monte Carlo (MCMC) method is used for estimation. MCMC is a general method based on drawing random values from approximate distribution, and then, correcting these values to better approximate the target posterior distribution. This method enables the statistician to examine data using complex but realistic statistical models such as MTREM (Martinez and Martinez, 2002). Bayesians and sometimes also frequentists need to integrate over possibly high dimensional probability distributions to make inference about model parameters or to make predictions. Bayesians need to integrate over the posterior distribution of model parameters given the data, and frequentists may need to integrate over the distribution of observables given parameter values. The dependence within subject measurements and high dimensional integrations during calculations are not problem with the use of MCMC approaches. Hope (1968) showed that results from a Monte Carlo simulation are unbiased, under the assumption that the programming is correct. In this study, Gibbs sampling with Hybrid steps are used to sample from the posterior distributions of the parameters in MCMC. Gibbs sampling was developed by Geman and Geman (1984) and it is used to generate the sample from desired distribution. It is one of the most popular MCMC techniques (Ross, 2002). In Gibbs Sampling method, first of all a starting point is selected. Each component of the parameter vector, then, is updated by sampling from the full conditional distribution of each parameter. Unlike Gibbs Sampling, in Hybrid Markov Chain (Neal, 1996), full conditionals do not have to be in a known parametric form. Therefore, when they are in a hard form to sample, Hybrid Markov Chain can be used within a Gibbs sampler. Hybrid Markov Chain makes use of gradient information, which facilitates the convergence. In this method, acceptance probability is suggested to be around 90%. The tuning parameters are selected to attain this probability.

Within this context, Bayesian inference is used to fit the mentioned model to the simulated data and to make conclusions about this method, with the help of the distribution and the parameters of the model used. One advantage of Bayesian inference is that it does not require large sample theory. Nevertheless, one should be careful about the choice of the prior distribution (Diggle et al., 2002). As it is stated earlier, MCMC method is used in this study, which has an intensive effect on Bayesian statistics. For complex problems like in this study, Bayesian MCMC approach provides computational advantages over the other approaches. Readers who are interested in the details of MCMC may refer to Gelman et al. (2004).

To sum up, in this thesis, a simulated bivariate longitudinal binary data is modeled by MTREM to investigate the properties of this complex model. An application of this model on a real life dataset can be found in Ilk and Daniels (2007). Although in this paper, the model fit for the second and the third level of MTREM is assessed via Deviance Information Criterian (DIC) (Spiegelhalter et al., 2002), and posterior predictive checks (Gelman et al., 2004), model fit related to first level has not been investigated previously. The effect of any misspecification on this first level is studied throughout this thesis.

The rest of the chapters are introduced as follows: In Chapter 2, Section 2.1, the statistical model is introduced. In this model, three levels are used to account for three different effects. Marginalized mean model, transition model and random effects model. In Section 2.2, data description and generation process are introduced. Our data is assumed to be about the satisfaction of the customers who withdraw their salary from a particular bank. There are 15 covariates to explain the responses. It is assumed that, data are collected at four different time points for each individual. In Section 2.3, there is some information about MCMC, Newton-Raphson Method and Gibbs Sampling. The implementation of our simulation study and the details about program used in simulation are explained in Section 2.4. The results are presented in Chapter 3.

CHAPTER 2

METHODOLOGY

This thesis aims at studying the properties of a statistical model developed for a complex structured data set through a simulation study. In this chapter, statistical model, description of simulated data, generation process of the data and estimation techniques are discussed.

2.1 STATISTICAL MODEL

The data sets which consist of repeated measurements in time on a collection of individuals, that is longitudinal data, require special methods for handling the correlations between the observations on a given individual (Gelman et al., 2004). Marginalized Transition Random Effects Models (MTREM), which is the collection of three models, is developed to handle such complex data sets. By these three models, the dependence of the response on the explanatory variables; the autocorrelations among the responses; and the correlations among responses at a fixed time point are modeled.

The first level of the model is Marginal Mean Model, which is used to explain the mean response. Repeated measurements are collected from same individuals, so that, the values collected from them are not independent. This correlation should be taken care of in analysis step. With the marginal models approach, the mean and covariance is modeled separately (Diggle et al., 2002).

Suppose Y_{itj} denote the j^{th} response type on the i^{th} individual at the time point t. In this study, j=1,2 ; i= 1,2,...,200 and t=1,2,3,4. The first level of the model is given in equation (2.1.1).

$$\text{Logit } P(Y_{itj}=1 \mid X_{itj}) = X_{itj} * \beta . \qquad (2.1.1)$$

Here, X_{itj} denotes the covariates and β denotes the coefficients corresponding to the covariates.

The second level is the Transition Model, which captures the longitudinal dependence within each of the j responses by a transition model of order p. It takes the past outcomes into consideration. In this type of model, the conditional expectation $E(Y_{itj}|Y_{i,t-1,j}, ..., Y_{i1j}, X_{itj})$ is dealed with and also the dependence of Y on X and of repeated Y's within themselves are accepted (Diggle et al., 2002). The second level of the model is illustrated in equation (2.1.2).

$$\text{Logit } P(Y_{itj}=1| y_{i,t-1,j}; \ldots; y_{i,t-p,j}; X_{itj}) = \Delta_{itj}+ \sum^{p}_{m=1} \gamma_{itj,m}\, y_{i,t-m,j} . \qquad (2.1.2)$$

Here, Δ_{itj} is the intercept in the logistic regression on the conditional probabilities and $\gamma_{itj,m}$ is the log odds ratio measuring the association between any pair of successive observations. Since the model which is used in this thesis is the first order model, the lag is one, i.e. p=1. At the same time, the transition parameters, γ_{itj}, is written as $\gamma_{itj}=\alpha_t*C_{itj}$, where C_{itj} is a vector of subset of covariates. By this specification, the transition parameter is allowed to differ by subject-specific covariates and also by responses and time (Ilk, 2004).

The third level is Random Effects Model, which models the correlation among the j responses at each time, conditional on the covariates and the previous responses. This method is generally used when inferences are to be made about individuals. The random effects model for a binary data is illustrated in equation (2.1.3):

$$\text{Logit } P(Y_{itj}=1| y_{i,t-1,j} ; ...; y_{i,t-p,j} ; X_{itj} ; b_{it}) = \Delta^{*}_{itj}+ \lambda_j b_{it} . \qquad (2.1.3)$$

Here, Δ^{*}_{itj} is the intercept and b_{it} is the random effects coefficient specific to subject i at time t. It is assumed that $b_{it} \sim N(0, \sigma_t^2)$. The parameter λ_j is related to the correlations between responses at a given time. If $\lambda_j=1$ for all j, then equal correlation among responses is assumed, conditional on the previous time responses. For identifiability, λ_1 is taken as 1.

These three levels of the model are connected to each other through the following two equations.

$\mathrm{P}(\mathrm{Y}_{itj} = 1 \mid \mathrm{X}_{\mathrm{itj}}) = \sum_{y_{i,t-1,j}} \mathrm{P}(\mathrm{Y}_{itj} = 1 | \mathrm{y}_{i,t-1,j}, \mathrm{X}_{itj})\, \mathrm{P}(\mathrm{Y}_{i,t-1,j})$ which can be rewritten as

$$\frac{\exp(X_{itj}\beta)}{1+\exp(X_{itj}\beta)} = \sum_{y_{i,t-1,j}} \frac{\exp(\Delta_{itj} + \gamma_{itj} y_{i,t-1,j})}{1+\exp(\Delta_{itj} + \gamma_{itj} y_{i,t-1,j})} \frac{\exp((X_{i,t-1,j}\beta) y_{i,t-1,j})}{1+\exp(X_{i,t-1,j}\beta)} . \tag{2.1.4}$$

and

$\mathrm{P}(\mathrm{Y}_{itj} = 1 | \mathrm{y}_{i,t-1,j}, \mathrm{X}_{itj}) = \int \mathrm{P}(\mathrm{Y}_{itj} = 1 | \mathrm{y}_{i,t-1,j}, \mathrm{X}_{itj}, \mathrm{b}_{it})\, \mathrm{dF}(\mathrm{b}_{it})$ which can be rewritten as

$$\frac{\exp(\Delta_{itj} + \gamma_{itj} y_{i,t-1,j})}{1+\exp(\Delta_{itj} + \gamma_{itj} y_{i,t-1,j})} = \int \frac{\exp(\Delta^*_{itj} + \lambda_j \sigma_t z_i)}{1+\exp(\Delta^*_{itj} + \lambda_j \sigma_t z_i)} \phi(z_i) dz_i \ , \tag{2.1.5}$$

where ϕ corresponds to standard normal density. To approximate the 1-dimensional integral above, Gauss–Hermite quadrature is used and to solve the equations Newton–Raphson methods are used (Ilk, 2005).

There is no history data available for the initial time point. Therefore, the second level of the model which regresses on the responses at the previous time can not be used. The specified simpler model for t=1 is as below in equations (2.1.6) and (2.1.7):

$$\mathrm{LogitP}(\mathrm{Y}_{i1j} = 1 \mid \mathrm{X}_{\mathrm{i1j}}) = \mathrm{X}_{i1j}\ \beta^* \ , \tag{2.1.6}$$

$$\mathrm{LogitP}(\mathrm{Y}_{i1j} = 1 | \mathrm{X}_{i1j}, \mathrm{b}_{i1}) = \Delta^*_{i1j} + \lambda^*_j\ \mathrm{b}_{i1} \ , \tag{2.1.7}$$

where $\mathrm{b}_{i1} \sim \mathrm{N}(0,\ \sigma_1^2)$, and $\lambda^*_1 = 1$. Note that, different marginal covariate parameters, β^*, are used in these equations instead of the original model parameters, β. In longitudinal data, more variability is expected at baseline, and marginal covariate effects are usually different from the effects at the other time points (Ilk and Daniels, 2007).

The details about the estimation methods are given in Section 2.3 and details about the model, such as likelihood function and the computational algorithm, can be found in Ilk (2004) and Ilk (2005).

2.2 DATA GENERATION PROCESS

This chapter aims at describing the data used in this study. To begin with, for the simulation study, it is assumed that the satisfaction of the customers who withdraw their salary from a particular bank is investigated. This assumption satisfies this study become more realistic and interesting. Suppose the binary and bivariate responses are to be, Y_1, the satisfaction of the customers from employees/staff of the bank (satisfaction=1, no satisfaction=0); and Y_2, the satisfaction of the customers from the substructure of the bank (like internet service, ATM, phone service) (satisfaction=1, no satisfaction=0). The covariates that are thought to be related to these responses, together with their descriptions are illustrated in Table 2.2.1. Moreover, it is assumed that the data are collected from individuals at four different time points, and that, it is collected for every six months. These time points are required to be equally spacing, because the second level of the model is an autoregressive model.

Note that, some variables such as X3 and X10 are time dependent; whereas variables such as X1 are time independent. The variable X2, age, takes the same value for the first and the second time. Age at the third and fourth time is calculated by adding one to the age at the first time point.

Besides that, for this simulation study, X1, X4, X5, X6, X8, X9 and X11 are generated from binomial distributions with different p values. These p values are chosen depending on the assumed occurrence proportions. For example, the p value is chosen to be 0.5 for the “X1=gender of the customer”. This means that the assumption is to have approximately equal number of male and female customers in our study. However, it is decided to choose 0.3 for “X8= Do the customers use other services like phone and internet (yes=1, no=0)”, which means that the expectation is to observe more customers to answer “no”. On the other hand, X2 is generated from normal distribution with minimum value 18; X3 from normal distribution with mean one and standard deviation 12; X7 from a normal distribution conditional on X3.

Similarly, X10 is generated from normal distribution with mean 95 and standard deviation 35. In addition, some restrictions are put during generation, because of several reasons: The age of the customers and average age of employees are expected to be greater than or equal to 18; the average waiting time and average procedure number to be nonnegative; the number of the active account number to be a nonnegative integer. Moreover, positive relationship between X3 and X7, and no change in X1 over the years are expected. Additionally, the interaction effects of X8*X2 and X8*X9 are added to the model. 200 observations are made at each four time periods. The R codes of the generation of the covariates for the first time point is shown in Appendix A. An example code for the covariates which are time dependent is also shown in Appendix A.

Table 2.2.1 Description of the Variables

X1	Gender of the customer (male=0, female=1)
X2	Age of the customer
X3	Average waiting time during the last six months (time to wait for the transition to be accomplished on phone or on branch/minute)
X4	The transition done in bank (credit card) (yes=1, no=0)
X5	The transition done in bank (investment procedure) (yes=1, no=0)
X6	The procedure which is done in bank (billing procedure) (yes=1, no=0)
X7	Average age of the employees/staff
X8	Do the customers use other services like phone or internet (yes=1, no=0)
X9	During the procedure -the customer have a problem related to the bank or employee=0 -the transition is done accurately=1
X10	Average number of transition done on the branch in a day
X11	Number of active accounts for the customer
X12	Response type (=1 for the satisfaction from the employees/staff, =0 for the satisfaction from the substructure of the bank)

On the other hand, dependent variables Y_{i1j} and Y_{itj} are generated from Bernoulli distribution. The probability functions of the dependent variables are seen in the equations (2.2.1) and (2.2.2).

$$P(Y_{i1j}=1) = \frac{\exp(\Delta^*_{i1j} + \lambda^*{}_j b_{i1})}{1+\exp(\Delta^*_{i1j} + \lambda^*{}_j b_{i1})} \tag{2.2.1}$$

$$P(Y_{itj}=1) = \frac{\exp(\Delta^*_{itj} + \lambda_j b_{it})}{1+\exp(\Delta^*_{itj} + \lambda_j b_{it})} \tag{2.2.2}$$

for i=1, ... ,200; j=1,2; t=2,3,4.

Here, the parameter $\Delta^*{}_{itj}$, is the intercept term. In the simulation of dependent variables, the parameters on the first and second levels of our model are taken into account. Therefore, it is provided that the responses of consecutive time points being dependent. Moreover, it is considered that $b_{it} = \sigma_t z_i$, where z_i ~ N(0,1) and the standard deviation is taken into consideration in simulation. The detailed information about the parameters is illustrated in Section 2.4.

The data generation process is repeated for 20 times, and each of them constitutes one sample. The model fit is done for all of these 20 samples. Before fitting the model, summary statistics such as the serial correlations between responses at consecutive time points and correlations between different response types at a fixed time points are checked for all 20 samples to verify that the simulated data sets are realistic. These serial correlation values are shown in Table 2.2.2 and Table 2.2.3. Here, Y_{itj} denotes the jth response type on the ith individual at the time point t. For instance, Cor(Y_{i1j},Y_{i2j}) column provides the correlation coefficients between responses at the first and the second time points. Note that, temporal correlations in the simulated data range between 0.016 and 0.305. Larger correlations ranging between 0.404 and 0.740 due to multiple responses are observed. In general, these values are expected to be nonnegative and moderate in size; and this is what is achieved by obtaining nonnegative and moderate values.

As it is known, many digits in the variables should be used in calculations in order to prevent round off errors. Assuming it will be enough, simulated data is rounded to three digits. Besides, standardized values of covariates are used, because the variables have substantially different magnitudes ranging from 1 to 1000 units. With the help of this standardization, it is possible to compare the variables with each other without the issue of unit difference. This step is also beneficial for computational issues. The estimation procedure requires the calculation of exponential functions of linear combinations of covariates. When covariates take large values, use of raw data may lead to overflow problems. Standardization of covariates enables us to overcome such problems.

Table2.2.2 Correlations between Responses at Consecutive Time Points

	Cor(Y_{i1j},Y_{i2j})	**Cor(Y_{i2j},Y_{i3j})**	**Cor(Y_{i3j},Y_{i4j})**
sample1	0.140	0.075	0.067
sample2	0.171	0.147	0.185
sample3	0.107	0.204	**0.305**
sample4	0.113	0.181	0.195
sample5	0.050	0.133	0.278
sample6	0.084	0.225	0.190
sample7	0.078	0.225	0.247
sample8	0.155	0.186	0.135
sample9	0.039	0.217	0.107
sample10	0.085	0.152	0.212
sample11	0.140	0.242	0.211
sample12	0.216	0.202	0.173
sample13	0.151	0.211	0.141
sample14	0.116	0.194	0.187
sample15	0.076	0.141	0.205
sample16	0.052	0.169	0.163
sample17	0.167	0.247	0.166
sample18	0.098	0.186	0.243
sample19	0.107	0.179	0.217
sample20	**0.016**	0.173	0.202

Table2.2.3 Correlations between Different Response Types at a Fixed Time Point

	Cor(Y_{i11},Y_{i12})	**Cor(Y_{i21},Y_{i22})**	**Cor(Y_{i31},Y_{i32})**	**Cor(Y_{i41},Y_{i42})**
sample1	0.579	0.531	0.555	0.481
sample2	0.625	0.569	0.526	0.526
sample3	0.649	0.514	0.584	0.529
sample4	0.729	0.553	0.483	0.477
sample5	0.631	0.552	0.461	0.430
sample6	0.558	0.530	0.509	0.524
sample7	0.620	0.509	0.433	0.503
sample8	**0.740**	0.478	0.435	0.509
sample9	0.721	0.509	0.428	0.486
sample10	0.640	0.472	0.485	0.466
sample11	0.665	0.520	0.539	0.416
sample12	0.680	0.449	0.450	0.452
sample13	0.610	0.435	0.434	0.475
sample14	0.724	0.510	0.491	0.470
sample15	0.690	0.558	0.517	0.457
sample16	0.665	0.471	0.506	0.483
sample17	0.648	0.503	0.516	0.448
sample18	0.650	0.533	0.515	**0.404**
sample19	0.652	0.576	0.441	0.420
sample20	0.721	0.511	0.516	0.476

2.3 ESTIMATION USING MCMC METHOD

In this section the basic information about priors, Markov Chain Monte Carlo Method (MCMC), Bayesian Statistics, Gibbs Sampling and Newton-Raphson Method are provided.

The prior distributions for the parameters β, β^* and α are specified from multivariate normal distributions with means of zero and large variances, $\sigma^2{}_{\beta}I$, $\sigma^2{}_{\beta*}I$ and $\sigma^2{}_{\alpha}I$, respectively. In this thesis, all these variances are taken to be 100. Note that, β values are assumed to be independent with each other and this is a realistic assumption. This is going to be seen from the scatter plots in the results section. Moreover, the prior distributions for the parameters λ and λ^* are specified from normal distributions with mean one and variance two. This means that, the parameters centered at the value one corresponding to equal correlation among the responses at a given time. The variance of two reflects the weakness of the correlation among the j responses; however, the results are not very sensitive to the specification of the variance. The prior distributions for parameters $\sigma^2{}_t$ are proportional to $1/(1+\sigma^2{}_t)^2$, which is the indication of positive probability at $\sigma^2{}_t = 0$, no multivariate dependence, and is on a similar scale to λ_j.

Estimation is handled by MCMC methods (Brooks, 1998). The MCMC methods could handle complex problems, such as allowing inference from a multi-level model like MTREM. MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps.

In this thesis, performing the integration of the parameters in equation (2.1.5) has some difficulties, and simpler method is used in such a complex situation to make the analysis feasible. Monte Carlo integration using MCMC is a solution to solve this complex problem. To explain the method, let think x as a vector of random variables, with distribution denoted by Π(x). Now, the goal is to obtain the expectation

$$E[f(x)] = \frac{\int f(x)\,\Pi(x)\,dx}{\int \Pi(x)dx} \quad . \qquad (2.3.1)$$

With MCMC method the distribution of x have to be known. The denominator of the equation 2.3.1 can be unknown. The fundamental part of the MCMC methodology is that the problem can be reduced to find integrals. Monte Carlo integration estimates the integral in equation (2.3.1) by obtaining samples x_t , t=1, 2, …, n from the distribution Π(x). Here, the important point is that, the samples are not independent. That does not limit MCMC methods' use in finding integrals using approximations (Martinez and Martinez, 2002).

With Hybrid Markov Chain in Gibbs Sampling, the first derivatives of full conditionals are calculated without integrating out b_{it} in likelihood, unlike in classical approach. Gibbs Sampling is an algorithm to generate samples from the joint probability distributions of two or more random variables. The purpose is to approximate the joint distribution, or to compute an integral such as an expected value. Gibbs sampling is a special case of the Metropolis-Hastings algorithm, and thus an example of a MCMC algorithm.

Gibbs sampling is applicable when the joint distribution is not known explicitly, but the conditional distribution of each parameter is known. The goal of the Gibbs sampling algorithm is to generate a value from the distribution of each parameter in turn, conditional on the current values of the other parameters (Gelman, 1996).

Hybrid Monte Carlo (HMC) tries to avoid random walk behavior by introducing a vector and implementing Hamiltonian dynamics where the potential function is the target density. Some samples are discarded after sampling. The end result of Hybrid MCMC is that moves across the sample space are in larger steps and are therefore less correlated and converge to the target distribution more rapidly.

To find a root of a complicated function algebraically is usually difficult. Using some basic concepts of calculus and Newton-Raphson method, it can be easier to evaluate the roots of complicated functions numerically. In numerical analysis, Newton's method is perhaps the best known method for finding successively better approximations to the roots of a real-valued function. With this iterative process it is approximated to one root, considering the function, its derivative, and an initial x-value. In this thesis, Newton's method is used to solve the equations 2.1.4 and 2.1.5. Newton's method can often converge quickly, especially if the iteration begins sufficiently near the desired root. Just how near sufficiently near needs to be, and just how remarkably quickly can be, depends on the problem.

2.4 SIMULATION STUDY

For data generation and checking the results, R is used, which is a language similar to S. R provides a wide variety of statistical and graphical techniques, and is highly extensible. It is available as Free Software which can be downloaded from the web page http://www.r-project.org. The readers who are interested in the R codes which are studied on, can have a look at Appendices A, B and C. On the other hand, for model fit, Fortran is used, which is a programming language especially suited for numeric computation and scientific computing. Fortran codes are available upon request.

It is known that in general, the number of cases, N, should be at least six to ten times bigger than the number of independent variables (Neter et al., 1996). In the model we have 15 variables, including the intercept and interactions. The number of observations, 200, is selected so that the general rule that is mentioned above is satisfied. Measurements are assumed to be taken at four time points. After the simulation of data, MCMC methods are applied to obtain the posterior distribution of the parameters. Chain is run for 5100 iterations; the first 100 iterations are thought as burn-in period and discarded. This is done because, although the simulation reaches to the approximate convergence eventually, the early iterations are influenced by starting values (Gelman et al., 2004). It is suggested that the size of burn-in period should be between 1% and 2% of iteration number (Martinez and Martinez, 2002). The use of Hybrid Markov Chain facilitates the convergence. Therefore, 100 iterations of burn-in period are satisfactory. The number of the burn-in period can also be decided by starting from different points and then drawing the trace plots on the same graph. When plotted lines first match with each other, this point can be taken as the end of the burn-in period. This means that, in the case of using different starting values, after some iteration the chain rapidly finds its way. Starting values have no effect on the chain if the model is true. As an example, the trace plot of the parameter beta0 using different starting values can be seen in Figure 2.4.1. Three different starting values 1.680, 1.880 and 2.680 are used for the parameter beta0 for which the true value is 0.68. Note that, although starting values for these three chains are quite different from each other, they take values very close to 0.68 even at iteration one. This quick convergence is due to the use of Hybrid Markov Chain. After 100 iterations the three lines are inseparable. Therefore, the first 100 iteration is decided to be used as burn-in period.

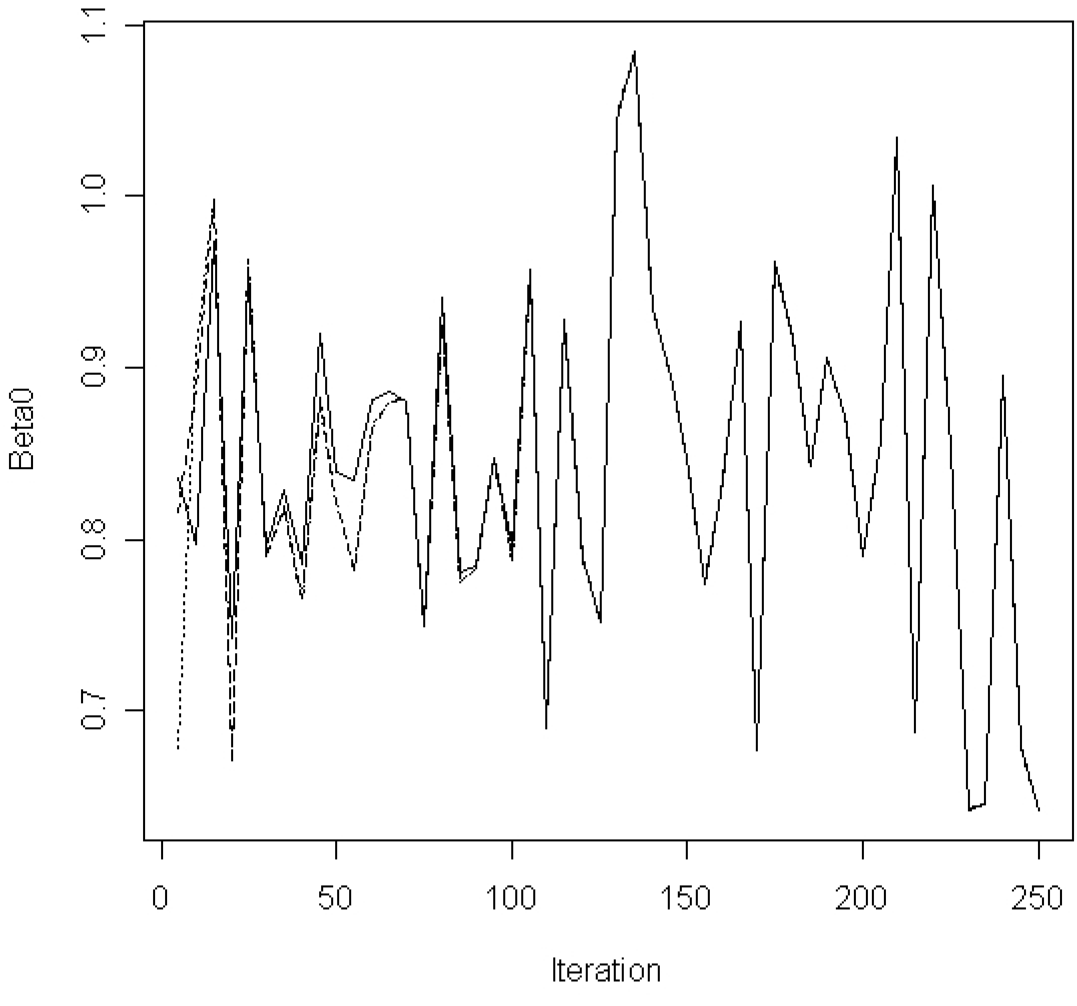

Figure 2.4.1 The Convergence Plot of the Parameter beta0 with Three Different Starting Values

On the first level of the model, 30 parameters, β and β^*'s, are estimated for this particular data set. On the second level, three α's are estimated and 200*3*2=1200 Δ intercepts (one for each 200 individuals, 3 time points, and 2 response types) are calculated. On the third level, 200*4=800 random effects coefficients are estimated together with the calculation of 200*4*2=1600 intercepts (Δ^*). Each of these intercepts correspond to one of the n=200 individuals at time t (t=1,2,3,4) for the jth response type (j=1,2). Note that, smaller number of intercept terms is calculated for the second level of the model compared to the third level. That is because the second level is not applicable for the first time point. These calculations are repeated for 5100 iterations. Therefore, the process is quite computationally intensive and takes a long time.

In this thesis, the simulation study is made under two different conditions. In the first one, the three levels of the models are fitted to the data with 15 variables, which constitutes our full model. In the second condition, two variables which have important effects on the model are removed from the model. These two variables are the ones whose coefficient have the highest true value for the main model (beta11=0.832, beta12=1). The results are compared under these two conditions in terms of parameter estimations for all the levels of the model. These results are shown in Chapter 3.

To interpret the MCMC outputs, both visual and exact tests were applied using the Bayesian Output Analysis (BOA). BOA is a program running under R/S-PLUS for carrying out convergence diagnostics and for statistical and graphical analysis of Monte Carlo sampling output. Autocorrelation, Density, Running Mean and Trace graphs helped us visually for checking convergence. On the other hand, Heidelberger and Welch Test (1983) as well as Raftery and Lewis Test (1992b) helped us on the convergence checking with the exact results.

The Heidelberger and Welch convergence diagnostic is appropriate for the analysis of individual chains. If there is evidence of non-stationarity, the test is repeated after discarding the first 10% of the iterations. This process continues until the resulting chain passes the test or more than 50% of the iterations have been discarded. BOA reports the number of iterations that were kept, the number of iterations that were discarded. Failure of the chain to pass this test indicates that a longer run of the MCMC sampler is needed in order to achieve convergence (Smith, 2003).

The Raftery and Lewis convergence diagnostic is also appropriate for the analysis of individual chains. This diagnostic test for convergence to the stationary distribution and estimates the run-lengths needed to accurately estimate quartiles of functions of the parameters. The user may specify the quartile of interest, the desired degree of accuracy in estimating this quartile, and the probability of attaining the indicated degree of accuracy. BOA computes the “lower bound” – the number of iterations needed to estimate the specified quartile to the desired accuracy using independent samples. If sufficient MCMC iterations are available, BOA lists the lower bound, the total number of iterations needed for each parameter, the number of initial iterations to discard as the burn-in set. The dependence factor measures the multiplicative increase in the number of iterations needed to reach convergence due to within-chain correlation. Dependence factors greater than 5.0 often indicate convergence failure and a need to reparameterize the model (Raftery and Lewis, 1992a). The detailed information about these tests is available in Cowles and Carlin (1996, pp.885-890). The code for BOA can be seen in Appendix D.

CHAPTER 3

RESULTS

The simulation results can be seen in Table 3.1. In this table, true parameter values, the mean values, bias and Mean Square Error (MSE) values of the parameters are shown. Here, betas0, betas1, ..., betas14, lambda2s, log.sigma2.1 are the parameters which are used for the first time point of the model. The parameter alpha2.1 connects the data at the second time point to the first time point. The other parameters are used for the other three time points of the data.

As can be seen from Table 3.1, the mean values of the parameters obtained from the full model are very close to the true values. This result shows us that the estimation technique used (MCMC) performs well in terms of estimation. After discarding two parameters from the model the new mean values are obtained. As expected, the mean values which are observed with the full model are closer to the true values compared with the mean values obtained by the reduced model. The bias, that is the difference between the true parameter and obtained mean value, is usually bigger for the misspecified model.

Together with the mean values, MSE values are also calculated. These values are the aspect of the good-fitting model, a kind of proof. In the calculation of MSE values, differences of all 20 sample mean values of all parameters to the actual values are used. The formula for calculating MSE is shown in equation (3.1). In the formula, $\overline{\theta_k}$ represents the sample mean of parameter k which is sampled from the posterior distribution with 5100 iterations. However, in the calculation of MSE values, the first 100 iterations are discarded as well as in the calculation of the mean values of the parameters.

$$\text{MSE}= \sum_{k=1}^{20} (\overline{\theta_k} - \theta_{\text{true}})^2/20 \qquad (3.1)$$

Generally small MSE values are observed. There are some bigger values among the MSE values, e.g. lambda2 and lambda2s under full model. For other parameters, MSE values are in general smaller for the full model than the misspecified one as it is expected.

In Table 3.1, the parameters of the three levels of the model are illustrated. The parameters betas0, betas1, …, betas14 and , beta0, beta1, …, beta14 are the parameters that are used in the first level of the model; alpha2.1, alpha3.1, alpha4.1 are the parameters that are used in the second level; and lambda2, lambda2s, log.sigma2.1, …, log.sigma2.4 are the parameters that are used in the third level of the model. As stated earlier, the first level of the model observes the changes in subgroups; the second level of the model observes the change in observations compared to previous time observations; and the third level of the model observes individual changes.

According to the values that are introduced in Table 3.1, there is not much change in MSE values for the parameters in the first level between the full model and the misspecified model. Nevertheless, the parameters in the second level of the model, alpha2.1, alpha3.1, and alpha4.1, have increase in the MSE values in the case of misspecification. The parameters in the second level of the model are more sensitive to misspecification compared to the parameters in the first level of the model. Another finding is that, the third level of the model is more sensitive to the misspecification than the first and the second levels of the model. As can be seen in Table 3.1, the close mean values are obtained with full model for parameters of the third level of model, such as log.sigma2.1, log.sigma2.2, log.sigma2.3 and log.sigma2.4. However, after misspecification, negative mean values are observed for the positive true values for these parameters, which mean very different values are obtained. Also, the MSE values for these parameters increase very much with the misspecification. Therefore, the interpretations about the individuals are more sensitive to the model misspecification.

It is important to note that smaller differences are seen between full and misspecified models in parameters related to first time point (betas0, …, betas14, alpha2.1, lambda2s, log.sigma2.1) compared to the parameters related to the later time points. This situation is because of the parameters' effect that is discarded. The discarded parameter, betas12, has no effect on the model. The true value of this parameter is '0' as seen in Table 3.1. Therefore, not much change is observed in MSE values of full model and reduced model. This result is even more fascinating considering smaller number of observations is used for estimating parameters at the first time point. This implies that, unless the number of parameters ignored is severe, the misspecification on the first level does not have a crucial impact on estimation procedure.

Table 3.1 Comparison of Estimated Values with the True Values (Mean Values and MSE Values)

5000 iter	true values	Mean (for full model)	Bias (for full model)	MSE (for full model)	Mean (for reduced model)	Bias (for reduced model)	MSE (for reduced model)
alpha2.1	0.74	0.746	-0.006	0.152	0.460	0.28	0.176
alpha3.1	1.03	1.161	-0.131	0.115	1.508	-0.478	0.314
alpha4.1	1.14	1.065	0.075	0.130	1.517	-0.377	0.248
beta0	0.68	0.817	-0.137	0.024	0.611	0.069	0.008
beta1	0.0505	0.079	-0.0285	0.022	0.032	0.0185	0.007
beta2	-0.627	-0.610	-0.017	0.043	-0.495	-0.132	0.027
beta3	-0.805	-0.738	-0.067	0.163	-0.652	-0.153	0.033
beta4	0.572	0.627	-0.055	0.018	0.452	0.12	0.018
beta5	0.789	0.775	0.014	0.147	0.660	0.129	0.029
beta6	0.351	0.353	-0.002	0.006	0.257	0.094	0.012
beta7	-0.495	-0.529	0.034	0.028	-0.402	-0.093	0.019
beta8	0.535	0.524	0.011	0.141	0.465	0.07	0.080
beta9	0.632	0.626	0.006	0.119	0.521	0.111	0.019
beta10	0.387	0.437	-0.05	0.009	0.325	0.062	0.010
beta11	0.832	0.868	-0.036	0.013			
beta12	1	1.005	-0.005	0.030			
beta13	-0.675	-0.716	0.041	0.083	-0.550	-0.125	0.080
beta14	0.478	0.461	0.017	0.036	0.363	0.115	0.040
betas0	0.041	0.117	-0.076	0.015	0.097	-0.056	0.008
betas1	-0.0765	0.001	-0.0775	0.064	-0.038	-0.0385	0.026
betas2	-0.83	-0.769	-0.061	0.204	-0.769	-0.061	0.048
betas3	-0.58	-0.593	0.013	0.163	-0.596	0.016	0.063
betas4	0.83	0.714	0.116	0.176	0.757	0.073	0.025
betas5	0.799	0.788	0.011	0.143	0.763	0.036	0.033
betas6	0.38	0.352	0.028	0.040	0.319	0.061	0.030
betas7	-0.093	0.028	-0.121	0.053	0.003	-0.096	0.043
betas8	0.637	0.603	0.034	0.772	0.752	-0.115	0.568
betas9	0.74	0.740	0	0.156	0.685	0.055	0.039
betas10	0.215	0.261	-0.046	0.045	0.208	0.007	0.021
betas11	0.872	1.004	-0.132	0.079			
betas12	0	-0.055	0.055	0.017			
betas13	-0.675	-0.773	0.098	0.702	-0.800	0.125	0.550
betas14	0.455	0.549	-0.094	0.086	0.483	-0.028	0.083
lambda2	1.22	2.306	-1.086	1.193	1.479	-0.259	0.122
lambda2s	1.07	2.082	-1.012	1.025	2.026	-0.956	1.038
log.sigma2.1	1.474	1.314	0.16	0.147	1.734	-0.26	0.259
log.sigma2.2	1.386	0.843	0.543	0.505	-0.507	1.893	3.836
log.sigma2.3	1.209	0.216	0.993	1.110	-0.399	1.608	2.752
log.sigma2.4	0.889	0.427	0.462	0.351	-0.276	1.165	1.573

Figures 3.1 and 3.2 are the scatter plots that show the correlations between some of the parameters in the full model and the misspecified model respectively. These graphs illustrate that the correlations between the parameters are low. It is a good result, since it provides the computational time to be small. These low correlations are satisfied by the standardization of the covariates.

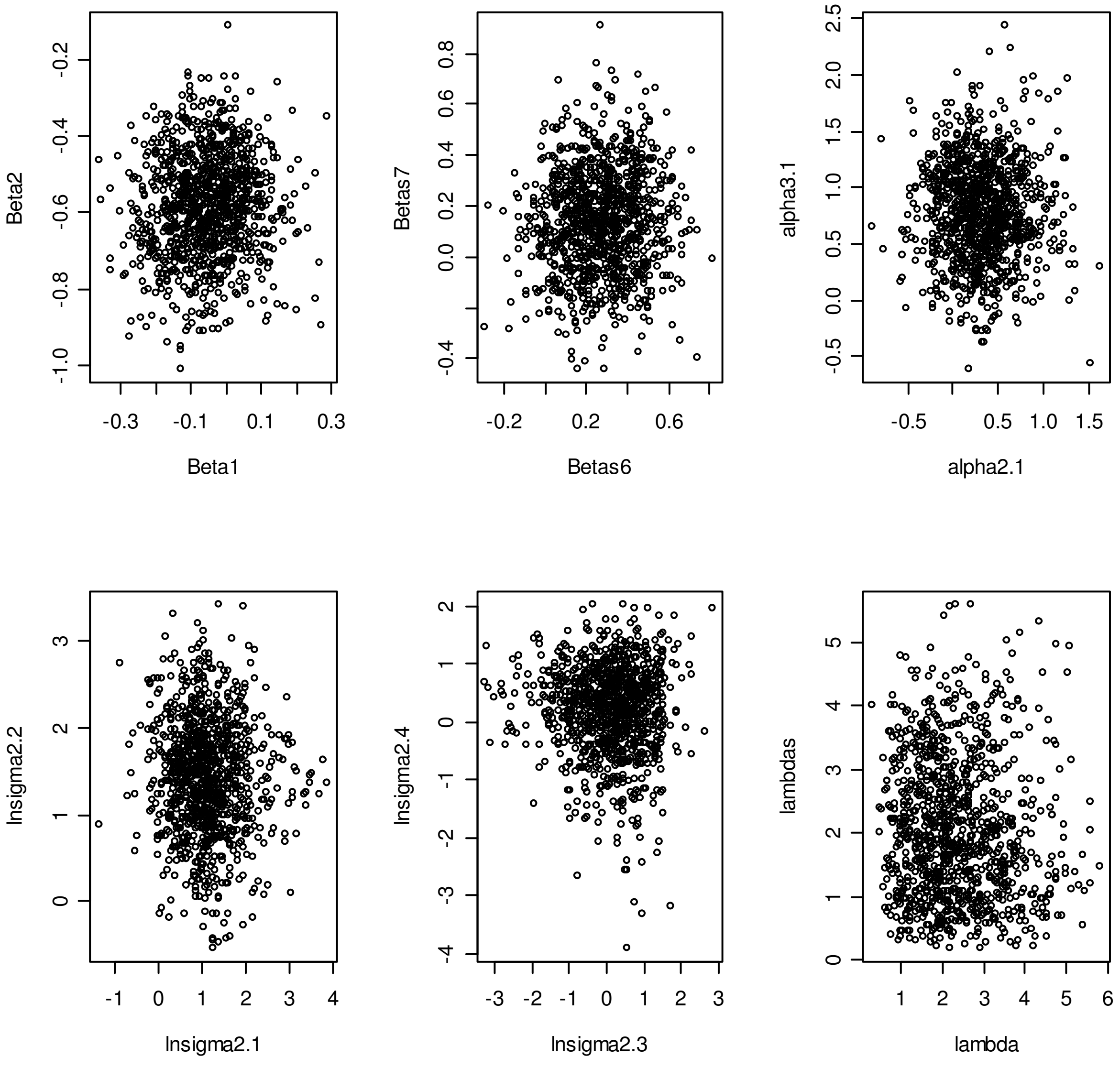

Figure 3.1 Scatter Plots of Some Parameters for a Randomly Chosen Sample with Full Model After Burn-in Period is Discarded

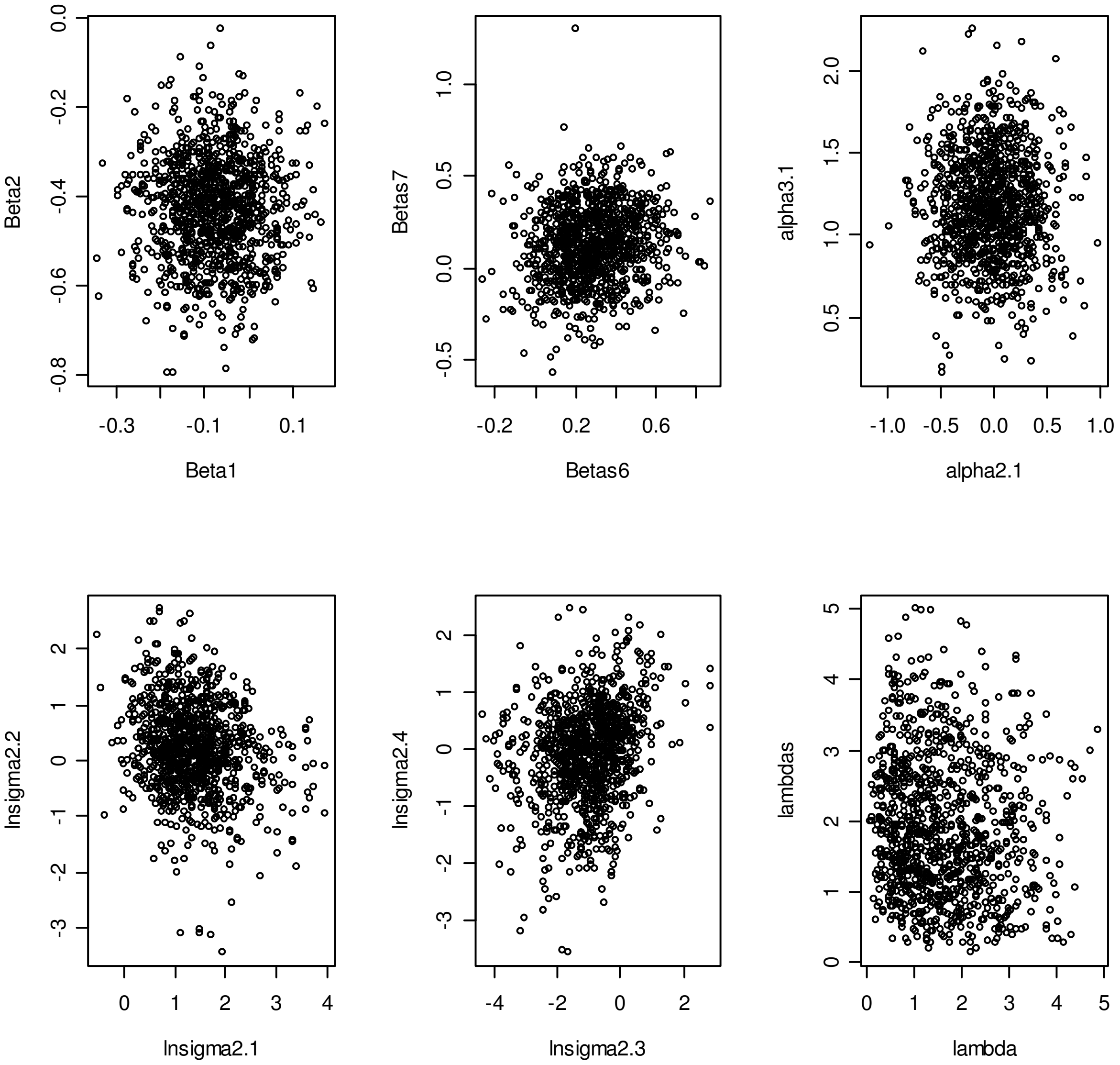

Figure 3.2 Scatter Plots of Some Parameters for a Randomly Chosen Sample After Burn-in Period is Discarded in the Case of Model Misspecification

In the autocorrelation graphs in Figures 3.3 and 3.4 for most parameters, there are no significant lags. For some parameters, such as ‘lambda2’, ‘lambda2s’ and ‘log.sigma2.2’, there seems to be large autocorrelations. Here, the important point is that, every fifth iterations are taken into account during all analysis. If wider lag intervals are taken, the decrease of autocorrelation can be observed more significantly. However, to avoid the loss of more information, this is not applied. Also, when the autocorrelation graphs of the full model and the misspecified model are compared, it is obvious that, there seems to be larger correlations with the misspecified model.

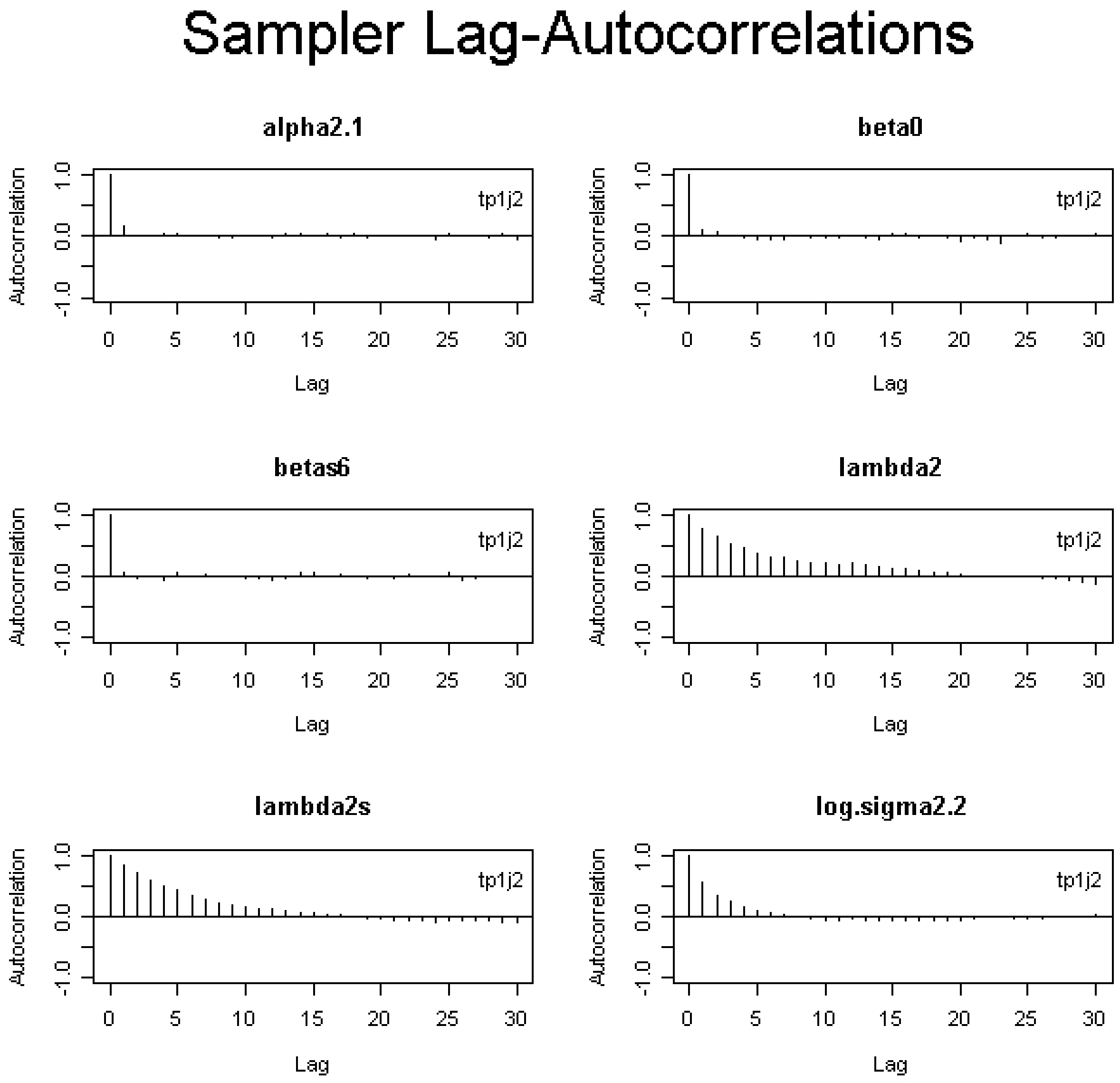

Figure 3.3 Autocorrelation Graph for a Randomly Chosen Sample with Full Model After Burn-in Period is Discarded

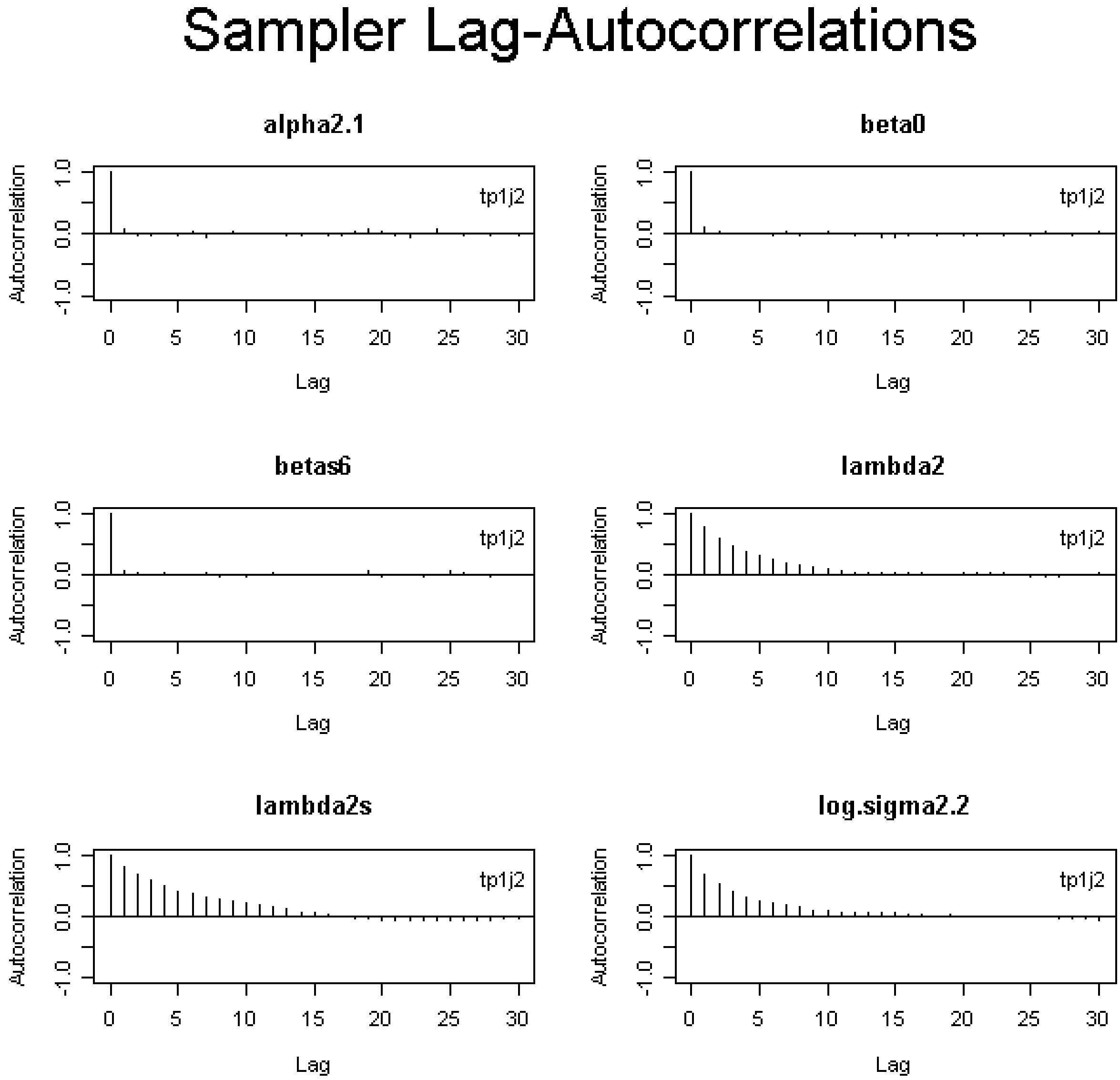

Figure 3.4 Autocorrelation Graph for a Randomly Chosen Sample After Burn-in Period is Discarded in the case of Model Misspecification

Starting values different from the actual values are used for parameters in the simulation. The aim is to see whether the convergence is provided in such a situation. If the model is true, the starting values do not affect the convergence. A rapid chain quickly finds its way even from extreme starting values (Geman and Geman, 1984). The convergence is satisfied in following iterations. It is obvious from Figure 3.5 that, approximately after the first 100 iterations, the values converge to the true values. However, by omitting two parameters and obtaining the misspecified model, convergence to the actual values is slower. The convergence is better for the full model. For instance, in Figure 3.5, the parameter ‘log.sigma2.2’ converges to the true value 1.39. Better convergence is obtained for the parameter ‘log.sigma2.2’ compared to the graph in Figure 3.6.

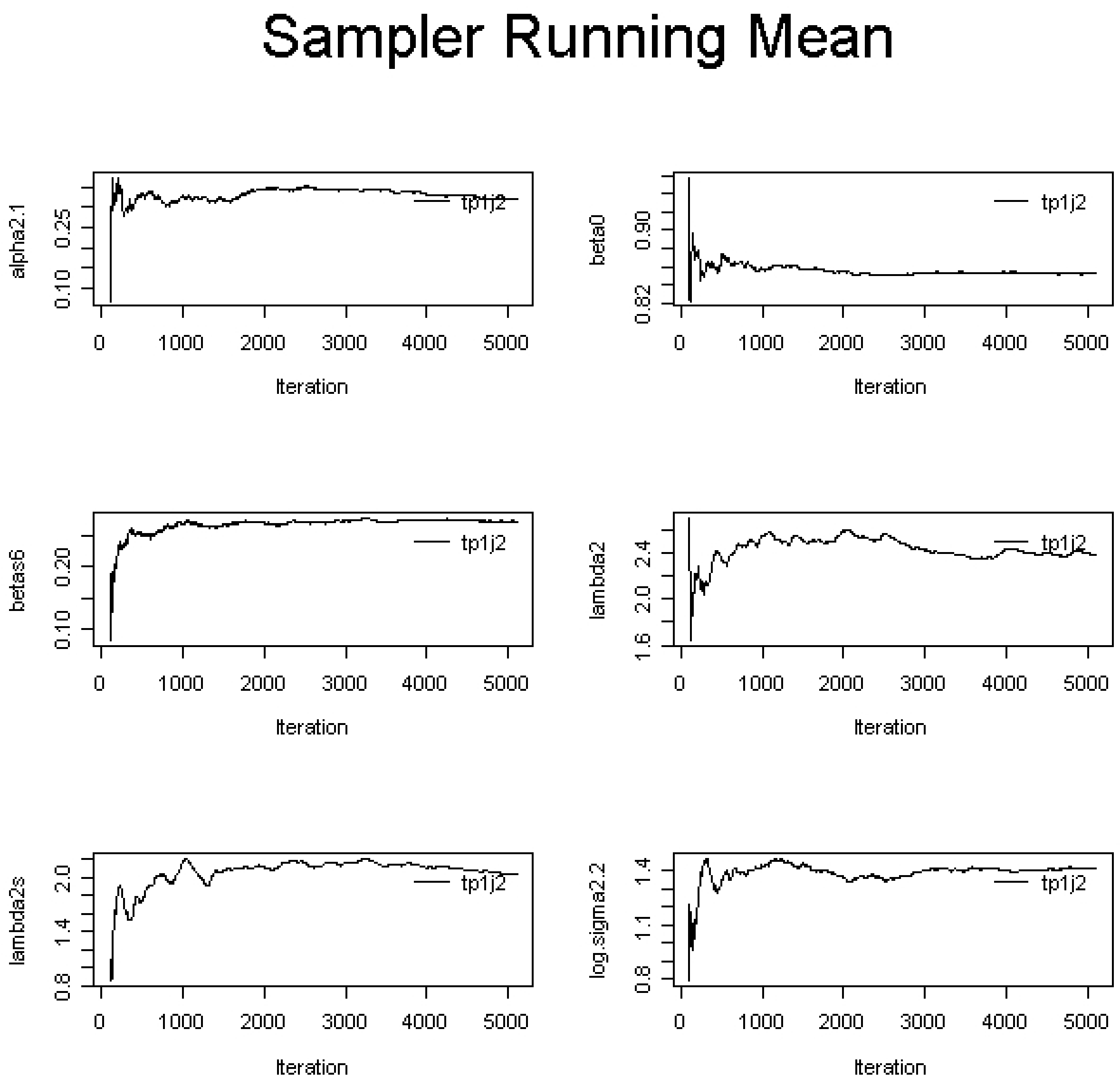

Figure 3.5 Running Mean Graph for a Randomly Chosen Sample with Full Model After Burn-in Period is Discarded

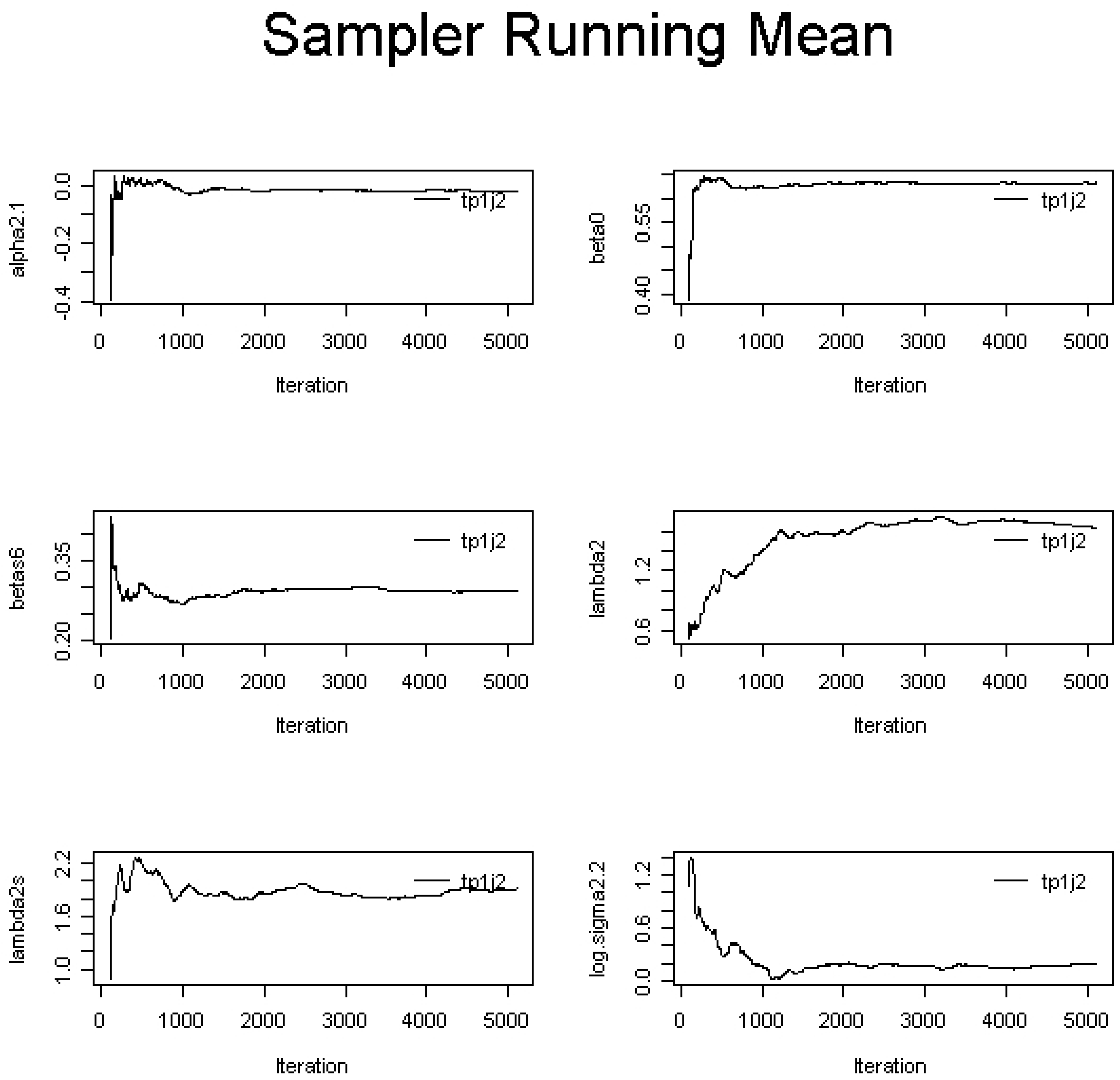

Figure 3.6 Running Mean Graph for a Randomly Chosen Sample After Burn-in Period is Discarded in the Case of Model Misspecification

Besides, the trace graph, which is shown in Figure 3.7, is quite stable. That means there is good convergence. For instance, in the trace graph of variable 'alpha2.1', the plotted values fluctuate around the actual value, 0.74 (true values are shown in Table 3.1). The same thing is valid for the other parameters. However, when the graphs in Figure 3.7 (the full model) are compared with the graphs in Figure 3.8 (the misspecified model), the convergence seems to be better in the full model. As it is expected, better convergence and results are obtained with the full model.

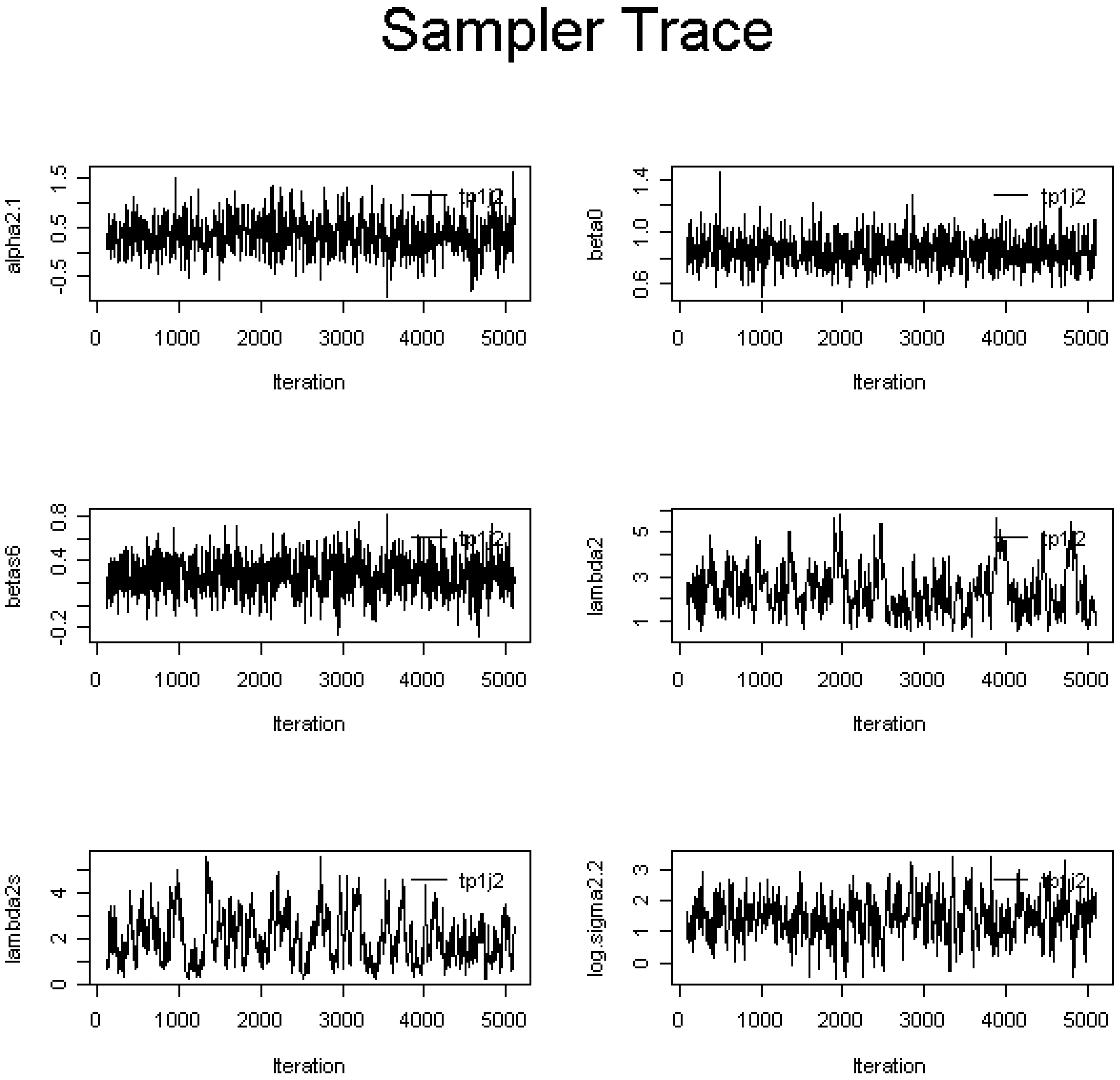

Figure 3.7 Trace Graph for a Randomly Chosen Sample with Full Model After Burn-in Period is Discarded

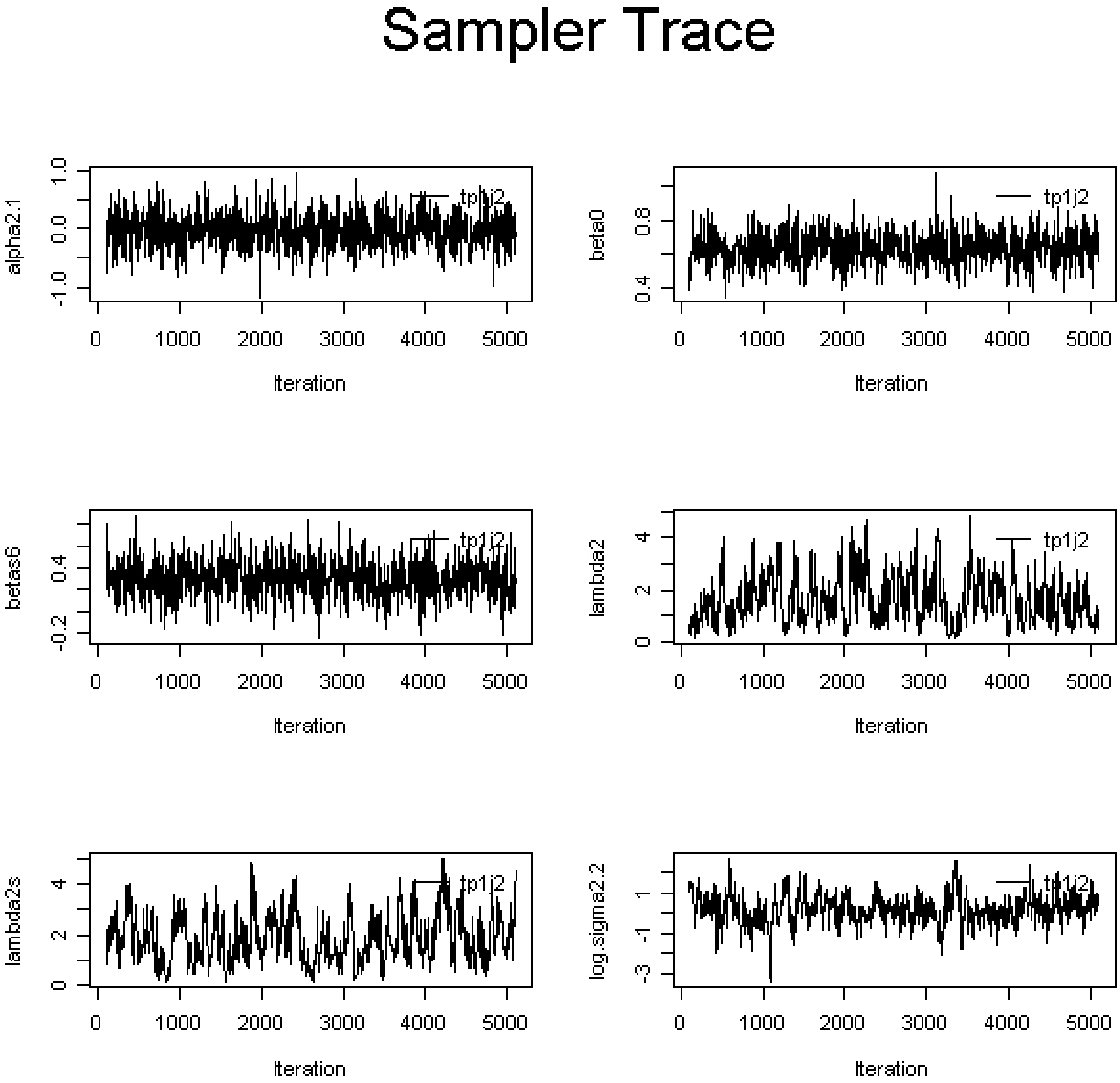

Figure 3.8 Trace Graph for a Randomly Chosen Sample After Burn-in Period is Discarded in the Case of Model Misspecification

The Heidelberger & Welch Test and Raftery & Lewis Test results for the full model and misspecified model can be seen in the Table 3.2 and Table 3.3. As it can be seen from these tables, all parameters pass the stationarity test and all iterations are kept in the analysis. Note that, since only one out of every five iterations is saved to avoid autocorrelation between iterations, 1000 iterations in this table correspond to the previously mentioned 5000 iterations. In addition, reasonable numbers are obtained for the Raftery & Lewis Test for both full and misspecified model. As stated earlier, obtaining smaller values than five is a requirement for Raftery & Lewis Test.

Table 3.2 Convergence Test Results for the Full Model

	Heidelberger and Welch Test			Raftery and Lewis Convergence Diagnostic
parameters	**StationarityTest**	**Keep**	**Discard**	**Dependence Factor**
alpha2.1	Passed	1000	0	1.184
alpha3.1	Passed	1000	0	1.184
alpha4.1	Passed	1000	0	1.079
beta0	Passed	1000	0	1.184
beta1	Passed	1000	0	1.079
beta2	Passed	1000	0	1.079
beta3	Passed	1000	0	1.079
beta4	Passed	1000	0	1.000
beta5	Passed	1000	0	1.184
beta6	Passed	1000	0	1.000
beta7	Passed	1000	0	1.079
beta8	Passed	1000	0	2.789
beta9	Passed	1000	0	1.079
beta10	Passed	1000	0	1.079
beta11	Passed	1000	0	1.184
beta12	Passed	1000	0	1.079
beta13	Passed	1000	0	3.053
beta14	Passed	1000	0	1.000
betas0	Passed	1000	0	1.079
betas1	Passed	1000	0	1.079
betas2	Passed	1000	0	1.079
betas3	Passed	1000	0	1.079
betas4	Passed	1000	0	1.184
betas5	Passed	1000	0	1.000
betas6	Passed	1000	0	1.079
betas7	Passed	1000	0	1.000
betas8	Passed	1000	0	1.421
betas9	Passed	1000	0	1.184
betas10	Passed	1000	0	1.184
betas11	Passed	1000	0	1.079
betas12	Passed	1000	0	1.079
betas13	Passed	1000	0	4.158
betas14	Passed	1000	0	1.079
lambda2	Passed	1000	0	2.474
lambda2s	Passed	1000	0	3.237
log.sigma2.1	Passed	1000	0	2.579
log.sigma2.2	Passed	1000	0	3.895
log.sigma2.3	Passed	1000	0	3.263
log.sigma2.4	Passed	1000	0	3.421

Table 3.3 Convergence Test Results for the Misspecified Model

	Heidelberger and Welch Test			Raftery and Lewis Convergence Diagnostic
parameters	**StationarityTest**	**Keep**	**Discard**	**Dependence Factor**
alpha2.1	Passed	1000	0	1.184
alpha3.1	Passed	1000	0	1.184
alpha4.1	Passed	1000	0	1.079
beta0	Passed	1000	0	1.158
beta1	Passed	1000	0	1.000
beta2	Passed	1000	0	1.079
beta3	Passed	1000	0	1.184
beta4	Passed	1000	0	1.079
beta5	Passed	1000	0	1.184
beta6	Passed	1000	0	1.184
beta7	Passed	1000	0	1.079
beta8	Passed	1000	0	3.421
beta9	Passed	1000	0	1.079
beta10	Passed	1000	0	1.079
beta11				
beta12				
beta13	Passed	1000	0	1.289
beta14	Passed	1000	0	1.079
betas0	Passed	1000	0	1.079
betas1	Passed	1000	0	1.184
betas2	Passed	1000	0	1.184
betas3	Passed	1000	0	1.289
betas4	Passed	1000	0	1.079
betas5	Passed	1000	0	1.079
betas6	Passed	1000	0	1.079
betas7	Passed	1000	0	1.289
betas8	Passed	1000	0	1.184
betas9	Passed	1000	0	1.211
betas10	Passed	1000	0	1.079
betas11				
betas12				
betas13	Passed	1000	0	3.316
betas14	Passed	1000	0	1.184
lambda2	Passed	1000	0	2.684
lambda2s	Passed	1000	0	4.211
log.sigma2.1	Passed	1000	0	2.684
log.sigma2.2	Passed	1000	0	4.895
log.sigma2.3	Passed	1000	0	3.553
log.sigma2.4	Passed	1000	0	1.421

The average computation time over 20 samples for model fit with 15 variables is 28899.636 seconds, which is approximately 8.03 hours. This time is 26555.619 seconds, which is 7.38 hours, for the samples with 13 variables (misspecified model).

CHAPTER 4

CONCLUSION

In this thesis a simulation study is held with the aim of assessing the sensitivity of the estimation procedure against model misspecification via a simulation study. The simulation study is done to model the response, satisfaction of the customers who withdraw their salary from a particular bank. To model the data, the three level Marginalized Transition Random Effects Model introduced by Ilk and Daniels (2007) is used. The estimation and convergence with full model and reduced model is investigated. The differences between true parameter values and estimated ones are observed with the full and misspecified model.

Exploratory and confirmatory analyses are held with the help of Bayesian Output Analysis. Autocorrelation graphs are used to see the correlation of the parameters on different iterations. In this study, there is not such a correlation problem. On the other hand, running mean and trace graphs are used to decide about convergence. The convergence is satisfied as seen in these graphs and also in the convergence tests (Heidelberger & Welch Test and also Raftery & Lewis Test).

The parameters which are used in the first level of the model have better convergence than the parameters which are used in the other levels of the model in both of the simulation studies. Again, as expected, the better convergence is obtained with the full model. The calculated mean values of the parameters are very close to the actual values in the full model case. On the other hand, in the misspecified model, these values go away from the actual values.

The smaller MSE values are obtained with full model, especially for some parameters rather than the misspecified model. Also, the parameters of the second level of the model are more sensitive to the misspecification than the parameters of the first level of the model. The changes in MSE values with misspecification are much more in the parameters of the second level of the model. The parameters of the third level of the model are more sensitive to the misspecification than the other levels of the model.

We should emphasize that model misspecification is not a preferred situation in the analysis of the data. However, since in real life the true model is unknown, in most situations it is not possible to attain it. The misspecification in the first level slightly affects the results of first and second level of the model. On the other hand, if it is crucial to measure the individual differences it is suggested to be extra careful about possible misspecification.

The expected convergence is satisfied even in the case of using different starting values from actual parameter values. With this simulation trick, the truth of the computational algorithm is ensured. The starting values converge to the actual values in the further paths of the simulation.

To run the simulation study, considerably much time is spent. Time factor is important in choosing the number of simulation repetitions. More repetitions or more observation number in each sample satisfies more realistic results. Therefore, increasing the number of repetition or the number of observation can be attempted. Also, the iteration number in each sample can be increased to have more realistic results.

Some limitations are used for the covariates, in their numbers or in their distributions. These limitations can also be changed in future studies or the time period can also be changed. For instance, the data can be calculated for more times than it is done in this thesis. Besides, observing the changes in the results according to the pattern in missing cases, that is, sensitivity analysis, can be investigated in future studies.

REFERENCES

Brooks, S. P. (1998). Markov Chain Monte Carlo Method and Its Application. The Statistician, 47, 69-100.

Cowles, M. K. and B. P. Carlin (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. Journal of the American Statistical Association, vol.91, no.434, 883-904.

Diggle, P. J., P. Heagerty, K. Y. Liang, S. L. Zeger (2002). Analysis of Longitudinal Data. 2nd ed. Oxford University Press, 126-140, 214-216.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin (2004). Bayesian Data Analysis. 2nd ed. Chapman and Hall/CRC Press, 275-349, 415-440.

Gelman, A. (1996). Inference and Monitoring Convergence in Markov Chain Monte Carlo in Practice. W.R. Gilks, S. Richardson and D.T. Spiegelhalter, eds. London: Chapman and Hall, 131-143.

Geman, S. and D. Geman (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Trans.Pattn.Anal.Mach.Intel., 6, 721-741.

Heagerty, P. J. (2002). Marginalized Transition Models and Likelihood Inference for Longitudinal Categorical Data. Biometrics, 58, 342-351.

Heagerty, P. J. (1999). Marginally Specified Logistic-Normal Models for Longitudinal Binary Data. Biometrics, 55, 688-698.

Heidelberger, P. and P. Welch (1983). Simulation Run Length Control in the Presence of an Initial Transient. Operations Research, 31, 1109-1144.

Hope, A. C. A. (1968). A Simplified Monte Carlo Significance Test Procedure. Journal of the Royal Statistical Society, Series B, 30, 582-598.

Ilk, O. and M. J. Daniels (2007). Marginalized Transition Random Effects Models For Multivariate Longitudinal Binary Data. The Canadian Journal of Statistics -La revue canadienne de statistique, 35 (1), 105-123.

Ilk, O. (2005). Modeling Multivariate Repeated Binary Measurements. Proceedings of the 35th International Conference on Computers and Industrial Engineering. Vol.1, 997-1002.

Ilk, O. (2004). Exploratory Multivariate Longitudinal Data Analysis and Models for Multivariate Longitudinal Binary Data. PhD thesis, Iowa State University.

Martinez, W. L. and A. R. Martinez (2002). Computational Statistics Handbook with Matlab. Chapman and Hall/CRC, 191-227, 425-461.

Neal, R. M. (1996). Bayesian Learning for Neural Networks. Springer, New York.

Neter, J., M. H. Kutner, C. J. Nachtsheim, W. Wasserman (1996). Applied Linear Statistical Models. Irwin Publication, 4th ed., 437.

Raftery, A. E. and S. M. Lewis (1992a & 1992b). 'How Many Iterations in the Gibbs Sampler?' in Bayesian Statistics 4, J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith, eds., Oxford: Oxford University Press, 763-773.

Reboussin, B. A. and J. C. Anthony (2001). Latent Class Marginal Regression Models for Longitudinal Data: Modeling youthful drug involvement and its suspected influences. Statistics in Medicine, 20, 623-639.

Ribaudo, H. and S. G. Thompson (2002). The Analysis of Repeated Multivariate Binary Quality of Life Data: A Hierarchical Model Approach. Statistical Methods in Medical Research, 11, 69-83.

Ross, S. M. (2002). A Course in Simulation. New York: Macmillan, London; Collier Macmillan, c1990, 223-246.

Smith J. B. (2003). Bayesian Output Analysis program. http://www.public-health.uiowa.edu/boa. (Date retrieved: October, 2007).

Spiegelhalter, D. J., Best, N.G., Carlin, B. P., and Vander Linde, A. (2002), Bayesian Measures of Model Complexity and Fit, Journal of the Royal Statistical Society, Series B, 64 (4), 583-616.

Verbeke, G. and G. Molenberghs (2000). Linear Mixed Models For Longitudinal Data. New York: Springer.

APPENDIX A

R CODES FOR GENERATION OF COVARIATES

```
x0<-rep(1,400)

x1<-rep(rbinom(200,1,0.5),2)  #to generate 200 numbers from binomial distribution
                               with equal  probability(0.5)

#to generate 200 numbers from uniform distribution
d=0
u1=runif(220)
u2=runif(220)
y=-log(u1+d)
epart=exp(-(y-1)^2/2+(d-1)^2/2)
z=NULL
z=ifelse(u2<=epart,y,NA)
x2.na= 18+z*12    #to obtain the values with mean 18 and standard deviation 12
x2=x2.na[!is.na(x2.na)]
x2=x2[1:200]
x2<-ifelse(x2<18,18,x2)  #to omit the values which are less than 18
x2<-round(x2,0)
x2<-rep(x2,2)

d=0
set.seed(1233)  #seed number satisfies generating the same number in each time
u1=runif(220)
set.seed(1233)
u2=runif(220)
y=-log(u1+d)
epart=exp(-(y-1)^2/2+(d-1)^2/2)
z=NULL
z=ifelse(u2<=epart,y,NA)
x3.na=1+z*12   #to obtain the values with mean 1 and standard deviation 12
x3=x3.na[!is.na(x3.na)]
x3=x3[1:200]
x3<-ifelse(x3<0,1,x3)   #to omit the values which are less than 0
x3<-rep(x3,2)

x4<- rep(rbinom(200,1,0.6),2)   #to generate 200 numbers from binomial distribution
                                 with  probabilities 0.6 and 0.4
x5<-rep(rbinom(200,1,0.4),2)
x6<-rep(rbinom(200,1,0.6),2)
```

```
set.seed(1233)
x7=rnorm(200,35+0.5*(sqrt(36)/sqrt(var(x3)))*(x3-mean(x3)),sqrt(36*(1-0.5^2)))
x7<-ifelse(x7<18,18,x7)
x7=rep(x7,2)

x8<-rep(rbinom(200,1,0.3),2)
x9<-rep(rbinom(200,1,0.7),2)

d=0
u1=runif(220)
u2=runif(220)
y=-log(u1+d)
epart=exp(-(y-1)^2/2+(d-1)^2/2)
z=NULL
z=ifelse(u2<=epart,y,NA)
x10.na= 95+z*35
x10=x10.na[!is.na(x10.na)]
x10=x10[1:200]
x10<-rep(x10,2)

x11<-rbinom(200,3,0.5)
x11<-ifelse(x11==0,1,x11)
x11<-rep(x11,2)

x12<-c(rep(0,200),rep(1,200))

xc<-cbind(x0,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x2*x8,x8*x9) #introduce the matrix
xst<-xc
for(i in 2:dim(xc)[2]){xst[,i]<-(xc[,i]-mean(xc[,i]))/sqrt(var(xc[,i]))} #to standardize the data
write.table(round(xst,3),"F:/tez/generate x/sample1/sim-x-t1.txt",sep="\t",
row.names=F, col.names=F)

#to generate X3 values for time 2
d=0
set.seed(124)  #the seed is changed for obtaining different values than in time 1
u1=runif(220)
set.seed(124)
u2=runif(220)
y=-log(u1+d)
epart=exp(-(y-1)^2/2+(d-1)^2/2)
z=NULL
z=ifelse(u2<=epart,y,NA)
x3t2.na= 1+z*12  #to obtain the values with mean 1 and standard deviation 12
x3t2=x3t2.na[!is.na(x3t2.na)]
x3t2=x3t2[1:200]
x3t2<-ifelse(x3t2<0,1,x3t2)  #to omit the values which are less than 0
x3t2<-rep(x3t2,2)
```

APPENDIX B

R CODES FOR CALCULATING SERIAL CORRELATIONS

```
# location of data
setwd("C:/Documents and Settings/Administrator/Desktop/tez/sample1")

# to introduce the columns of the matrix
y=matrix(scan("y"),ncol=1)
i<-rep(rep(1:200,2),4)
t<-c(rep(1,400),rep(2,400),rep(3,400),rep(4,400))
j<-rep(c(rep(1,200),rep(2,200)),4)
yc<-cbind(i,t,j,y)

# calculation of correlation between responses at different time points
cor(yc[yc[,2]==2,4],yc[yc[,2]==1,4],method=”spearman”)
cor(yc[yc[,2]==3,4],yc[yc[,2]==2,4],method=”spearman”)
cor(yc[yc[,2]==4,4],yc[yc[,2]==3,4],method=”spearman”)

# calculation of correlation between different response types at a fixed time point
cor(yc[yc[,2]==1&yc[,3]==1,4],yc[yc[,2]==1&yc[,3]==2,4],method="spearman")
cor(yc[yc[,2]==2&yc[,3]==1,4],yc[yc[,2]==2&yc[,3]==2,4],method="spearman")
cor(yc[yc[,2]==3&yc[,3]==1,4],yc[yc[,2]==3&yc[,3]==2,4],method="spearman")
cor(yc[yc[,2]==4&yc[,3]==1,4],yc[yc[,2]==4&yc[,3]==2,4],method="spearman")
```

APPENDIX C

R CODES FOR CALCULATING ACCEPTANCE PROBABILITIES

```
# location of data
setwd("C:/Documents and Settings/Administrator/Desktop/tez/sample1")

b<-matrix(scan("bp1j2.out"),ncol=26,byrow=TRUE)
b<-b[-c(1:20),]

sum(b[,2]>b[,3])/nrow(b)   #acceptance probability of random effect term in the
                            3.level of the model at time 1
sum(b[,10]>b[,11])/nrow(b)  #acceptance probability of random effect term in the
                              3.level of the  model at times 2,3 and 4.

rest2<-matrix(scan("tp1j2.out"),ncol=54,byrow=TRUE)
rest2<-rest2[-c(1:20),]

sum(rest2[,2]>rest2[,3])/nrow(rest2)     #acceptance probability of beta
sum(rest2[,19]>rest2[,20])/nrow(rest2)   #acceptance probability of betas
sum(rest2[,36]>rest2[,37])/nrow(rest2)   #acceptance probability of alpha
sum(rest2[,45]>rest2[,46])/nrow(rest2)   #acceptance probability of lambda
sum(rest2[,48]>rest2[,49])/nrow(rest2)   #acceptance probability of lambdas
```

APPENDIX D

BOA CODES

```
#the location of the data
setwd("C:/Documents and Settings/Administrator/Desktop/tez/sample1")

#to read the files and required columns of the files
b<-matrix(scan("bp1j2.out"),ncol=26,byrow=TRUE)  #to read random effects
                                                   coefficients
b<-b[-c(1:20),]   #to discard the first 20*5=100 iterations as burn-in
rest2<-matrix(scan("tp1j2.out"),ncol=54,byrow=TRUE)   #to read the coefficients
                                                         for other  parameters
rest2<-rest2[-c(1:20),]

#to omit the columns that that is only used for calculating acceptance probabilities
rest2<-rest2[,-c(2:3,19,20,36,37,45,46,48,49,51:54)]
colnames(rest2)<c("iter","b1","b2","b3","b4","b5","b6","b7","b8","b9","b10","b11",
"b12","b13","b14","b15","bs1","bs2","bs3","bs4","bs5","bs6","bs7","bs8","bs9","bs1
0","bs11","bs12","bs13","bs14","bs15","alpha2;1","alpha3;1","alpha4;1","log.sigma
2.1","log.sigma2.2","log.sigma2.3","log.sigma2.4","lambda2","lambda2s")
write.table(rest2,"tp1j2.txt",sep="\t",row.names=F)

install.packages("boa")    # only on the first time to install the package automatically
library(boa)               #to attach the necessary library
boa.menu()
1 # to choose File
3 # to import data
7 # options
1 #  to set working directory
C:/Documents and Settings/Administrator/Desktop/tez/sample1
4 # flat ASCII file
tp1j2
1 # back
1 # back

#### graphics
4 # plot
3 # descriptive
3 # autocorrelations
4 # density
5 # running mean
6  # trace
```

```
#### confirmatory analysis
3 # analysis
3 # descriptive
6 # Summary statistics (gives the mean, SD, CI, and median values)
1 # back
5 # options
7 # accuracy
0.05
4 # convergence
4 # geweke (there is evidence against convergence when p-value is less than 0.05)
5 # Heidelberger and welch (if the stationary test fails, chain needs to be run longer
  # for convergence purposes. If the halfwidth test fails, chain might be run longer to
    increase the accuracy in estimating posterior estimate)
6 # Raftery and lewis (dependence factor greater than 5 implies convergence
    problem)

### if the menu unexpectedly terminates, type:
boa.menu(recover = TRUE)
```