A SIMULATION STUDY ON THE COMPARISON OF METHODS FOR THE ANALYSIS OF LONGITUDINAL COUNT DATA

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

BY

GÜL İNAN

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

JULY 2009

Approval of the thesis:

A SIMULATION STUDY ON THE COMPARISON OF METHODS FOR THE ANALYSIS OF LONGITUDINAL COUNT DATA

submitted by **GÜL İNAN** 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:

Assoc. Prof. Dr. İnci Batmaz
Department of Statistics, METU _____

Dr. Özlem İlk
Department of Statistics, METU _____

Assist. Prof. Dr. Recai Yücel
Department of Epidemiology and Biostatistics,
School of Public Health, University at Albany, SUNY _____

Dr. Berna Burçak Başbuğ Erkan
Department of Statistics, METU _____

Dr. Ceylan Yozgatlıgil
Department of Statistics, METU _____

Date: 30.07.2009

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: Gül İNAN

Signature:

ABSTRACT

A SIMULATION STUDY ON THE COMPARISON OF METHODS FOR THE ANALYSIS OF LONGITUDINAL COUNT DATA

İnan, Gül

M.Sc., Department of Statistics

Supervisor: Dr. Özlem İlk

July 2009, 76 pages

The longitudinal feature of measurements and counting process of responses motivate the regression models for longitudinal count data (LCD) to take into account the phenomenons such as within-subject association and overdispersion. One common problem in longitudinal studies is the missing data problem, which adds additional difficulties into the analysis. The missingness can be handled with missing data techniques. However, the amount of missingness in the data and the missingness mechanism that the data have affect the performance of missing data techniques. In this thesis, among the regression models for LCD, the Log-Log-Gamma marginalized multilevel model (Log-Log-Gamma MMM) and the random-intercept model are focused on. The performance of the models is compared via a simulation study under three missing data mechanisms (missing completely at random, missing at random conditional on observed data, and missing not random), two types of missingness percentage (10% and 20%), and four missing data techniques (complete case analysis, subject, occasion and conditional mean imputation). The simulation study shows that while the mean absolute error and mean square error values of Log-Log-Gamma MMM are larger in amount compared to the random-intercept model, both regression models yield parallel results. The

simulation study results justify that the amount of missingness in the data and that the missingness mechanism that the data have, strictly influence the performance of missing data techniques under both regression models. Furthermore, while generally occasion mean imputation displays the worst performance, conditional mean imputation shows a superior performance over occasion and subject mean imputation and gives parallel results with complete case analysis.

Key words: Longitudinal count data, gamma distributed random effects, drop-out or intermittent missing data, missing data mechanisms, missing data techniques.

ÖZ

UZUNLAMASINA KESİKLİ VERİ ANALİZİ İÇİN YÖNTEMLERİN KARŞILAŞTIRILMASI ÜZERİNE BİR BENZETİM ÇALIŞMASI

İnan, Gül

Yüksek Lisans, İstatistik Bölümü

Tez Yöneticisi: Dr. Özlem İlk

Temmuz 2009, 76 sayfa

Ölçümlerin, uzunlamasına özelliği ve bağımlı değişkenin sayım süreci, uzunlamasına kesikli veri analizi için geliştirilen regresyon modellerin birey içi ölçümler arası bağımlılık ve aşırı yayılım gibi olguları dikkate almasını sağlamaktadır. Uzunlamasına çalışmalarda, analizlere fazladan zorluk katan, ortak bir sorun, kayıp veri problemidir. Bu problem kayıp veri çözüm teknikleri vasıtasıyla üstesinden gelinebilir. Fakat verideki kayıp değer miktarı ve de verinin sahip olduğu kayıp veri mekanizması, çözüm tekniklerinin başarısını etkilemektedir. Bu tez çalışmasında, uzunlamasına kesikli veri için geliştirilmiş olan regresyon modelleri arasından, Log-Log-Gama marjinalleştirilmiş çok düzeyli model ile rasgele sabit terimli model üzerinde durulmuştur. Modellerin başarıları bir benzetim çalışması üzerinden, üç kayıp veri mekanizması (tamamıyla rasgele kayıp, gözlenmiş veriye bağlı rasgele kayıp, rasgele olmayan kayıp), iki tür kayıp yüzdesi (% 10 ve % 20), ve dört farklı kayıp veri çözüm tekniği (tüm durum analizi, yerine zaman ortalaması yükleme, yerine grup ortalaması yükleme, yerine regresyon yöntemiyle yükleme) altında karşılaştırılmıştır. Benzetim çalışması, Log-Log-Gama marjinalleştirilmiş çok düzeyli modelinin rasgele sabit terimli modelinden daha büyük ortalama mutlak hata ve ortalama karesel hata ürettiğini gösterirken, her iki regresyon modelinde paralel sonuçlar gözlenmektedir. Benzetim çalışması sonuçları, her iki regresyon

modelinde de veri içindeki kayıp değer miktarı ve kayıp veri mekanizmasının, kayıp veri çözüm tekniğinin başarısını ciddi bir şekilde etkilediğini doğrulamaktadır. Ayrıca, genellikle yerine zaman ortalaması yükleme tekniği en kötü başarıyı gösterirken, yerine regresyon yöntemiyle yükleme tekniği, yerine zaman ortalaması yükleme ve yerine grup ortalaması yükleme tekniklerine nazaran üstün bir başarı sergilemektedir ve tüm durum analiziyle benzer sonuçlar doğurmaktadır.

Anahtar Kelimeler: Uzunlamasına kesikli veri, gama dağılımlı rasgele etkiler, drop-out veya kesintili kayıp veri, kayıp veri mekanizmaları, kayıp veri çözüm teknikleri.

To my family,

for their constant support

and

unconditional love…

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my thesis supervisor Dr. Özlem İLK who accepted me as her M.Sc. student without hesitation and give me the opportunity to work with her. It was a great fortune for me to take the advantage of her invaluable knowledge and extensive academic experience. I am deeply indebted to her, for patience, help, suggestions and encouragements in all the process of research and writing of this thesis. Her guidance, support, and feedbacks have turned this study to an immeasurable learning experience for me.

I would also like to acknowledge my appreciation to my examining committee, Assoc. Prof. Dr. İnci Batmaz, Assist. Prof. Dr. Recai Yücel, Dr. Berna Burçak Başbuğ Erkan, and Dr. Ceylan Yozgatlıgil, for spending their valuable time in reviewing my thesis. Their valuable comments and suggestions have improved the quality of my thesis.

I also owe special thanks to The Scientific and Technological Research Council of Turkey (TÜBİTAK) for funding my graduate studies.

I would like to convey my heartfelt thanks to Tuğba Özaktaş. Thesis research and writing was both a painful and enjoyable journey. She was always there to listen and to give me advice. I greatly appreciate her gentle friendship, everlasting patience, and loving kindness.

To survive in graduate school, it was important to be a member of friendship group which can double your joy and divide your grief. Very special thanks go to my classmates Emrah Eren, Gözde Şener, and Zerrin Yalçınöz for everything. I am grateful to Zerrin for sharing her experience of thesis writing as well.

My sincere thanks go to Fatih Yalabuk. Whenever I had problems with my

computer, he offered me a lot of friendly help, and provided me an efficient computer environment to complete my thesis.

A big THANK YOU to my students as well, for their understanding, support, and interest in my thesis progress.

Last, but not least, I want to thank to my mummy for her unconditional support and encouragement to pursue my academic career, to my sister Gonca, for listening to my complaints and frustrations, for overcoming my bitterness, to my brother Başar, for his words of encouragement, and especially for Istanbul adventures. I owe my all achievements to them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

Abbreviations	Explanations
AR(1)	Autoregressive, order 1
CDF	Cumulative distribution function
CS	Compound symmetry
GEE	Generalized estimating equation
GLM	General linear model
GLMM	Generalized linear mix model
LCD	Longitudinal count data
LOCF	Last observation carried forward
Log-Log-Gamma MMM	Log-Log-Gamma marginalized multilevel model
MA(1)	Moving average, order 1
MAE	Mean absolute error
MAR	Missing at random
MCAR	Missing completely at random
MLE	Maximum likelihood estimation
MQL	Marginal quasi-likelihood
MSE	Mean square error
MVN	Multivariate normal distribution
NLMM	Nonlinear mixed model
NMAR	Not missing at random
NOCB	Next observation carried backward
PIT	Probability integral transformation
PQL	Penalized quasi-likelihood
REML	Restricted maximum likelihood
UN	Unstructured

CHAPTER 1

INTRODUCTION

In longitudinal studies, measurements from the same subjects over a sequence of time periods are taken so that changes in measurements over time periods can be observed. When the response variable of a longitudinal dataset represents the counts of a total number of a defined event occurring in a given time interval, this type of longitudinal data is known as the longitudinal count data (LCD). Many longitudinal count data examples can be considered from different disciplines. Examples from econometrics can be the annual number of tourist arrivals in each of the Mediterranean countries over several years, the total number of patents acquired yearly by each firm over many years, and the total number of medals won by nations in the Olympic Games over several periods. An example from political science can be the total number of homeless protests across cities of a country over several years. Examples from clinical research include the number of epileptic seizures of each patient per two-weeks over an eight-week treatment period, and the number of panic attacks for each patient in a week over a one-month psychological intervention program.

Like longitudinal data analysis for continuous response variables and for binary response variables, the starting point for the statistical analysis of longitudinal count data is the generalized linear models (GLMs) (Nelder and Wedderburn, 1972). Specifically, similar to cross-sectional count data, longitudinal one is modeled by Poisson regression as well (Cameron and Trivedi, 1998, Chapter 9). However, the analysis of longitudinal count data and the estimation of regression parameters require more special methods due to the longitudinal feature of measurements and counting process of responses. The most important feature of longitudinal data

that motivates the statistical analysis is the association of measurements within a subject. Since repeated measurements on the same subject are taken over several time periods, the observations obtained from the same subject are expected to be correlated. The statistical distribution of the counts is traditionally assumed to be Poisson distribution (Diggle et al., 2002, Chapter 8). It is well-known that mean equals to the variance (equi-dispersion) for the Poisson distribution. When the variability of counts is greater than its expected value under the Poisson model, then this phenomenon is called overdispersion. More specifically, extra-Poisson variation occurs (Barron, 1992). As an alternative, the negative-binomial distribution is the most commonly used model for overdispersed count data (Diggle et al., 2002, Chapter 8). Apart from these, other characteristics of longitudinal data causing additional difficulties to the statistical analysis are that the subjects may not have the same number of repeated measurements, that subjects may be measured at uncommon set of time intervals, and/or that measurements within a subject may be taken at non-equidistant time intervals. These irregularly or unequally spaced longitudinal data will result in an unbalanced longitudinal data design. More information on longitudinal data design is available in Ilk (2008, Chapter 2). Last but not least, when one or more measurements from subjects are missing, the problem which is named incomplete longitudinal data occurs. In order to accommodate the longitudinal features of measurements and counting process of responses, methods accommodating these problems should be developed.

In this thesis, a simulation study is carried out to assess the performance of the regression estimates produced from a Log-Log-Gamma marginalized multilevel model developed (Log-Log-Gamma MMM) by Griswold and Zeger (2004) for longitudinal count data. This model is an expansion of the marginalized latent variable models, proposed by Heagerty and Zeger (2000). Marginalized latent variable models are likelihood-based methods that combine a marginal regression model for the mean response with making use of the flexible dependence specifications of GLMMs to model the within-subject association. A comparison of

this model is made with a random-intercept model which is the simplest case of GLMMs where the model includes only one random effect in the linear predictor, in addition to fixed effects.

For the present study, longitudinal count datasets on 100 subjects are generated for each model and the simulation study is repeated 120 times. First of all, statistical analysis of each dataset is performed according to its own statistical model and results are recorded. Afterwards, observations in the generated datasets for each model are deleted according to scenarios of three different missing data mechanism (MCAR, MAR conditional on observed data, and NMAR). Deletion is limited to only observations from response variable and percentage of missingness in the samples is adjusted as 10% or 20%. A complete-case analysis is applied and results are recorded. That is, subjects having missing observations are discarded from the study and the subjects having complete observations are retained in the study. Later, the so-called missing observations are filled in by three different single imputation techniques. Specifically, subject mean imputation, occasion mean imputation and conditional mean imputation are carried out. In conditional mean imputation, a Markov model of order 1, that's first-order autoregressive, AR(1), model is preferred. Finally, the statistical evaluation of 120 samples for each model, under each 24 conditions (3 missing data mechanisms $\times$ 2 types of missingness percentage $\times$ 4 missing data techniques) is compared via mean absolute error (MAE) and mean square error (MSE) values.

The development of subsequent chapters of this thesis is organized as follows: Chapter 2 gives background information on regression model classes for longitudinal count data and missing data problem. Chapter 3 introduces the marginalized latent variable models and then focuses on the Log-Log-Gamma marginalized multilevel model, and its competitor regression model, random-intercept model for longitudinal count data. Later sections of this chapter provide detailed information on the three missing data mechanisms, and missing data

techniques i.e. complete case analysis and single imputation methods. Chapter 4 is devoted to simulation studies and put emphasis on the data generation scenarios, true values, missing data generation scenarios and estimation of the models. Based on simulation results, the evaluation of the performance of the models are discussed in Chapter 5. Finally, concluding remarks and suggestions for future work are presented in Chapter 6.

CHAPTER 2

HISTORICAL BACKGROUND

This Chapter aims to give brief information about historical background of regression models for longitudinal count data and missing data problem through examples from literature.

2.1 Literature Review of Regression Models for Longitudinal Count Data

Diggle et al. (2002, Chapter 7) classify the extension of GLMs for longitudinal data into three different regression model classes, which is also valid for longitudinal count data. These are:

i) Marginal or Population-Average models,
ii) Random-Effects or Subject-Specific models,
iii) Transition or Response Conditional models.

In general these three regression model classes view the association problem between the repeated measurements of a subject from different perspectives and this leads the models to differ in the interpretation of the regression parameters.

Firstly, marginal models directly specify a regression model for the mean response, using a known link function**.** The mean responses are related to the covariates as follows:

$$\begin{aligned} &E(Y_{ij} \mid X_{ij}) = \mu_{ij} \text{ and} \\ &g(\mu_{ij}) = X_{ij}\beta, \end{aligned} \tag{2.1}$$

where g is a common link function. The within-subject association, the association between the repeated measurements of a subject, is modeled separately, possibly

using additional association parameters**.** Here, the main interest is on the mean response and the regression model for the mean response depends only on covariates, for that reason it is named as marginal. The regression parameters, β's, in (2.1) interpreted as the averages of population as in cross-sectional analysis.

When the responses are discrete, i.e., binary or count, the complete joint distribution of longitudinal responses requires the specification of two-way associations between the responses (Fitzmaurice and Molenberghs, 2008, Chapter 1). However, building models for these associations that are consistent in an interpretable manner with the model for the mean response is difficult, in the context of marginal models (Lipsitz and Fitzmaurice, 2008, Chapter 3). Consequently, it is hard to estimate regression parameters of the marginal models by likelihood-based methods (Fitzmaurice and Molenberghs, 2008, Chapter 1). When the assumption on the distribution of repeated responses is avoided, an estimation method that is called the generalized estimating equation (GEE) is considered. It is developed by Liang and Zeger (1986) and it is the multivariate extension of the quasi-likelihood estimation method (Wedderburn, 1974) by including additional nuisance parameters in the formulation of covariance matrix of responses. GEE provides as efficient estimates as maximum likelihood estimation (MLE) and consistent and asymptotically normal estimates provided that the mean response model is correctly specified. However, since GEE deprive us of using likelihood-based methods; this method is not used in this thesis. Further information on GEE is available in Liang and Zeger (1986), Molenberghs and Verbeke (2005, Chapter 8) and Lipsitz and Fitzmaurice (2008, Chapter 3). In the framework of marginal analysis with GEE, it is possible to find several examples in the literature. For instance, Diggle et al. (2002, Chapter 7) and Molenberghs and Verbeke (2005, Chapter 19) fit a Poisson model to the data from a clinical trial of 59 epileptics which was first introduced by Leppik et al. (1985). While Diggle et al. propose a parametric model for the correlation coefficient, Molenberghs and Verbeke propose first-order autoregressive, AR(1) correlation

structure for the within-subject association using SAS GENMOD procedure.

The second method considering the extension of GLMs to longitudinal data setting is the random-effects models. The random-effects models assume that there is a natural heterogeneity between the subjects due to unmeasured covariates (Diggle et al., 2002, Chapter 7). In this sense, regression parameters randomly varying from one subject to other subject are included into the regression modeling of the mean response.

There are several ways of introducing randomness into the regression parameters of the mean response model. Thall and Vail (1990) suggest a family of covariance models that take within-subject association and overdispersion into account by introducing an interaction effect of subject-specific and time-specific random coefficients into the mean response model. While regression parameters are estimated by GEE, estimation of additional parameters is carried out by the method of moments. However, their approach cannot be used to model special type of autocorrelation structures such as first-order autoregressive, AR(1) and first-order moving average, MA(1). Following this, Jowaheer and Sutradhar (2002) propose a random-effects model for overdispersed longitudinal count data, which follow a Gaussian-type autocorrelation structure. Estimation of regression parameters, association and overdispersion is done via GEE. Although GEE is dominant in marginal models, it is obvious that random-effects models can be fitted by GEE as well.

However, among the random-effects models, generalized linear mixed effects models is the most frequently used one for discrete repeated measurements (Molenberghs and Verbeke, 2005, Chapter 14). Generalized linear mixed effects models are also called as generalized linear mixed models (GLMMs) after highly cited paper of Breslow and Clayton (1993). In social sciences, these models are known as hierarchical, multilevel, or random-coefficient models as well.

In GLMMs, the model for the mean response depends both on covariates and random effects which enter linearly into the linear predictor via a known link function. Similarly, mean responses are related to covariates and vector of random effects as follows:

$$\begin{aligned} &E(Y_{ij} \mid X_{ij}, b_i) = \mu_{ij} \quad \text{and} \\ &g(\mu_{ij}) = X_{ij}\beta + Z_{ij}b_i, \end{aligned} \tag{2.2}$$

where b_i is a vector of random effects for subject i, $(b_{0i}, b_{1i}, ..., b_{q-1\,i})$, and g is a common link function. Within-subject association is assumed to be resulted from unobservable variables which are common to each repeated measurement from the same subject. GLMMs assume the subject-specific random effects, b_i, to have a multivariate normal distribution with zero mean and a covariance matrix, V.

In GLMMs, the aim is to make inference on individual subjects rather than the population average; for that reason the fixed effects regression parameters, β's, in (2.2) are influenced by the subject-specific interpretations whereas in marginal models target inference is the population.

In GLMMs, maximum likelihood estimation of regression parameters, β's, in (2.2) requires the maximization of the likelihood of the data. Maximization is achieved by integration over the distribution of random effects, b_i's. However, distribution specification of random effects and high-dimensional integration of them together with a possibly nonlinear link function may cause computational difficulties in the evaluation of the likelihood and analytic solutions cannot be provided. Molenberghs and Verbeke (2005, Chapter 14) divide the approaches toward maximum likelihood estimation in GLMMs into three categories according to the frequency of usage and to the availability in statistical software. These are the approaches based on the approximation of i) the integrand, ii) data, and iii) integral itself.

While Laplace–type approximations fall in the first category, penalized quasi-

likelihood (PQL) and marginal quasi-likelihood (MQL) fall in the second category. The numerical integration methods such as Gaussian quadrature and adaptive Gaussian quadrature fall in the latter category. Molenberghs and Verbeke (2005, Chapter 5) suggests that serious attention should be paid to statistical software available and to the approximations, which these statistical software are based on, since different methods produce considerably different parameter estimates. Detailed information on approximation techniques in GLMMs is available in Molenberghs and Verbeke (2005, Chapter 14). Furthermore, the implementation of GLMMs for longitudinal count data can be done via glmm function under the repeated library in R, glmmPQL function under the MASS library in R, SAS GLIMMIX procedure, and SAS NLMIXED procedure. In this thesis, SAS NLMIXED procedure with the numerical integration method Gaussian quadrature, and SAS GLIMMIX procedure with PQL are used for the regression models discussed in Chapter 3. The reasons and detailed information on the SAS procedures and approximation and numerical integration techniques are presented in Section 4.4.

The last method considering the extension of GLMs to longitudinal data setting is the transition models. In transition models, the mean response is regressed on the covariates and a subset of past responses of the same subject via a known link function. These past responses can be considered as additional explanatory variables. Mean responses are related to the covariates and past responses as follows:

$$\begin{aligned} &E(Y_{ij} \mid X_{ij}, H_{ij}) = \mu_{ij} \quad \text{and} \\ &g(\mu_{ij}) = X_{ij}\beta + \sum_{r=1}^{p} \alpha_r f_r(H_{ij}), \end{aligned} \tag{2.3}$$

where g is a common link function with $H_{ij} = (Y_{i1},\ldots,Y_{ij-1})$ and $f_r(H_{ij})$ is a function of past responses. The association between the repeated measurements of a subject is considered to be as a result of the effect of past responses on the present response. A specific class of (2.3) is the Markov models of order p (Feller, 1968). The order of

the Markov models, p, is considered to be the number of past responses influencing the current response. Application of Markov models to longitudinal data is difficult in the event of unequally spaced time intervals between repeated measurements of a subject, and any missing measurement in the data. The interpretation of the fixed effects regression parameters, β's, in (2.3) depends on other responses for the same subject and the order p.

2.2 Literature Review of Missing Data

Since longitudinal studies obtain measurements from all subjects over a sequence of time periods, it is highly possible to encounter missing values, i.e., the intended measurements on some subjects cannot be obtained, or are not available for some reasons. For example, in clinical trials, subjects are followed over a number of scheduled visits, and incomplete longitudinal data can occur due to missed visits, withdrawal from the study, loss to follow-up, or death.

In longitudinal studies, two kinds of missing data pattern exist: drop-out and intermittent missingness. When the subject is withdrawn from the longitudinal study before its intended completion, no data can be provided thereafter, and this process is defined as drop-out. Drop-out truncates the longitudinal process and leads to a monotone missing data pattern for the measurements (Daniels and Hogan, 2008, Chapter 5). In other words, monotone missing data pattern occur, for example, when a measurement for a subject is missing at a scheduled visit, and data for subsequent measurements are also missing thereafter. On the other hand, in intermittent missing data pattern, a measurement for a subject is missing at one or more scheduled visits, but after that, data for subsequent measurements are provided. Intermittent missingness creates gaps in the longitudinal process and leads to a non-monotone missing data pattern for the measurements. Tables 2.1 and 2.2 below contain some possible monotone and non-monotone missing data patterns in measurements where O = Observed and M = Missing. In longitudinal

studies, it is common to encounter a mixture of drop-outs and intermittent missing data patterns.

Table 2.1 Monotone Missing Data Patterns

Subject	Y_1	Y_2	Y_3	.	.	.	Yn_i
1	O	O	O	.	.	.	O
2	O	O	O	.	.	.	M
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	O	O	O	O	M	.	M
.	O	O	O	M	M	.	M
.	O	O	M	M	M	.	M
N	O	M	M	M	M	.	M

Table 2.2 Non-Monotone Missing Data Patterns

Subject	Y_1	Y_2	Y_3	.	.	.	Yn_i
1	O	O	O	.	.	.	O
2	O	M	O	.	.	.	O
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	O	M	M	.	.	.	O
.	O	O	O	M	M	.	O
.	O	O	M	M	M	.	O
N	O	M	M	M	M	.	O

In the event of missing data, it is important to reveal the relationship between the response variable and the reasons causing missingness (Fitzmaurice and Verbeke, 2008, Chapter 2).

According to terminology by Rubin (1976) and Little and Rubin (2002, Chapter 1) there are three types of reasons, often referred to as missing data mechanisms, which cause data to be missing. If the missingness is independent of both observed and unobserved measurements, this mechanism is said to be missing completely at random (MCAR). If the missingness is independent of unobserved measurements, this mechanism is said to be missing at random (MAR) conditional on observed data. Both of these types of missing data mechanisms are also called ignorable since it does not affect the inference on the population parameters of interest in the statistical model (Bennett, 2001). When the missingness is neither MCAR nor MAR conditional on observed data, it is said to be missing not at random (MNAR). In a similar fashion, this type of missing data mechanism is also called non-ignorable since it affects the inference on the population parameters of interest in the statistical model (Bennett, 2001).

Molenberghs and Fitzmaurice (2008, Chapter 17) state when the missingness is unrelated to the response variable of interest, loss of information in the data

would not complicate the statistical analysis. However, when it is related to the response variable of interest, loss of information in the data would address more serious questions and require more attention during the statistical analysis. For that reason, in order to yield valid statistical inferences from incomplete longitudinal data**,** the statistical analysis should be performed according to the type of missingness mechanism that the data have.

Bennett (2001) classifies general methods dealing with missing data into five main categories, which is also applicable to incomplete longitudinal data. These are: i) methods that ignore missing values, ii) single imputation methods, iii) other imputation methods (multiple imputation and markov-chain imputation), iv) likelihood-based methods, and v) indicator methods. He compares and contrasts these methods in terms of producing bias in the estimates under the three missing data mechanisms (MCAR, MAR conditional on observed data, and NMAR), handling with variability, and availability in statistical software.

In addition to missing data mechanism, the amount of missingness in the data is also an important determinant in the event of selecting a statistical method to analyze the missing data (Roth, 1994).

Joseph and Molenberghs (2009) give a comprehensive review of missing data patterns, mechanisms, models specifying the joint distribution of the data and missingness mechanism (shared-parameter, pattern-mixture, and selection models) inferential paradigms, ignorability, and response types in longitudinal studies. The statistical literature on missing data problem in longitudinal studies with continuous response or with discrete response is mostly on dropouts and there is much less attention to intermittent missing data. Correspondingly, in the framework of incomplete longitudinal count data, Li et al. (2007) propose a random-effects Markov transition model for Poisson-distributed repeated measures when the data contain non-ignorable missing values, by making use of shared-parameter models. Kaciroti et al. (2008) suggest a Bayesian pattern-mixture model to model

longitudinal data from Poisson responses with potential non-ignorable drop-outs. An ignorability index is developed to capture the effect of a non-ignorable missing data mechanism on the statistical analysis and, in turn, it can be used for sensitivity analyses.

CHAPTER 3

METHODOLOGY

In this chapter, we firstly introduce features of the regression models for longitudinal count data used in this thesis, that's Log-Log-Gamma marginalized multilevel model, and random-intercept model. In oncoming part of the chapter, the detailed descriptions of missing data mechanisms, MCAR, MAR conditional on observed data, and NMAR, which an incomplete data may have, are illustrated with notations and examples. We conclude the chapter with missing data techniques to handle the incomplete data.

3.1 Regression Models

In longitudinal studies, main interest is to model the relationship between the covariates and the response, as well as the change of response over time. In Chapter 2, several approaches are introduced which account for the within-subject association for longitudinal data in order to properly assess the regression parameter estimates in the model. Marginal models construct two separate regression models for the mean response, and association between repeated measurements of a subject. The regression parameters describe the effects of covariates on the population averaged mean response and their interpretation is independent of specification of within-subject association model (Fitzmaurice and Molenberghs, 2008, Chapter 1). Regression parameters of marginal models can be estimated without specifying joint distribution of responses which leads to GEE to be developed. However, avoiding defining the complete joint distributions deprive us of using likelihood-based methods. In this sense, GLMMs are developed such that a regression model is built for the mean response conditional on both

covariates and random effects. Random effects, which are shared by the measurements within a subject, are introduced into the regression model to represent within-subject association. Contrary to marginal models, GLMMs model the mean response and the within-subject association through a single equation and random effects are viewed as the potential source of within-subject association. The regression parameters of GLMMs describe the effects of covariates on an individual's mean response by controlling for the random-effects. However, interpretations being dependent on random effects and sensitive to within-subject association specifications and robustness of estimates being dependent on the distribution of the random effects reflect the disadvantages of GLMMs (Heagerty and Zeger, 2000).

3.1.1 Marginalized Latent Variable Models and Log-Log-Gamma Marginalized Multilevel Model

Marginalized latent variable models are proposed by Heagerty and Zeger (2000). These marginalized multilevel models combine the features of GLMMs and marginal models, with an aim to compensate the distinctions of these two models. While marginalized latent variable models take likelihood-based inference capabilities and flexible within-subject association specifications from GLMMs, they take the interpretation and robustness of regression parameters from marginal models (Griswold and Zeger, 2004).

The formulation of marginalized latent variable models define a GLM for the model of the mean response, and a nonlinear mixed model (NLMM), which is assumed to be nonlinear in the random effects, for the dependence between measurements of a subject (within-subject association) and specify a statistical distribution for the random effects as follows:

i) Marginal Mean Model: $g(\mu_{ij}^{M}) = X_{ij}\beta^{M}$

ii) Association Model: $g(\mu_{ij}^C) = \Delta_{ij} + Z_{ij}b_{ij}$

iii) Random Effects Distribution: $b_i \square F_b(0, D)$

iv) Conditional Response Distribution: $Y_{ij}^C = (Y_{ij} \mid b_{ij}) \square F_{Y^C}(\mu_{ij}^C, \psi)$

where Y_{ij} represents the j^{th} measurement of the i^{th} subject, ($j=1,2,\ldots,n_i$, $i= 1,2,\ldots,N$), g is a common link function for both marginal mean model and association model, $\mu_{ij}^M = E(Y_{ij} \mid X_{ij})$ and $\mu_{ij}^C = E(Y_{ij} \mid X_{ij}, b_{ij})$, $\beta^M = (\beta_0^M, \beta_1^M, \ldots, \beta_{p-1}^M)'$ refers to the $p \times 1$ vector of marginal regression parameters, $b_i = (b_{0i}, b_{1i}, \ldots, b_{q-1\,i})'$ refers to the $q \times 1$ vector of subject-specific random effects with a $q \times q$ covariance matrix D and a distribution $F_b(.)$. While b_{0i} stands for the random intercept, $b_{1i}, b_{2i}, \ldots,$ and $b_{q-1\,i}$ represent the random slopes in the linear predictor of the association model. Given b_{ij}, Y_{ij}^C 's independently come from a distribution, which is a member of exponential family, with mean μ_{ij}^C and a dispersion parameter ψ. X_{ij} is a $N \times p$ matrix of fixed effects covariates, and Z_{ij} is a $N \times q$ matrix of random effects covariates. Z_{ij} is usually taken as a subset of X_{ij}.

In probability theory, any conditional expectation can be written in terms of marginal expectation, such that $E_b\{\mu_{ij}^C\} = E_b\{E(Y_{ij} \mid X_{ij}, b_{ij})\} = E(Y_{ij} \mid X_{ij}) = \mu_{ij}^M$. This implies that the integration of conditional mean, μ_{ij}^C, over the distribution of random effects results in marginal mean, μ_{ij}^M.

$$\mu_{ij}^M = \int_b \mu_{ij}^C \, dF(b) = \int_b g^{-1}(\Delta_{ij} + Z_{ij}b_{ij}) \, dF(b), \qquad (3.1)$$

where g^{-1} is the inverse-link function. Thus, the parameter Δ_{ij} in (3.1) makes a connection between marginal mean model and association model. The parameter Δ_{ij} depends on both $X_{ij}\beta^M$ marginal linear predictor and the distribution of b_i.

Different choices for i – iv result in different models. Griswold and Zeger (2004)

expand the marginalized latent variable models of Heagerty and Zeger (2000) for count data. In practice, count data do not hold the equi-dispersion (equality of mean and variance) assumption of Poisson distribution, and it exhibits most often overdispersion rather than underdispersion (Cameron and Trivedi, 1998, Chapter 3). Overdispersion generally arises from omitted covariates in the regression model. Neglecting overdispersion in the data causes misestimation of the regression parameters and related estimates such as standard errors and, in turn, misleads inferences regarding the regression parameters (Yang et al., 2007). To handle overdispersed count data, a number of approaches are proposed. Among these, one standard approach is to assign a gamma distribution for random effects and a Poisson distribution for the conditional response. Since the gamma distribution is conjugate of the Poisson distribution, this Poisson-gamma mixture leads to negative-binomial distribution which accommodates overdispersion well (Greenwood and Yule, 1920; Barron, 1992; Cameron and Trivedi, 1998; Jowaheer and Sutradhar, 2002; Yang et al., 2007). Griswold and Zeger (2004) follow the same logic and assume a gamma distribution for random effects and a Poisson distribution for the conditional response distribution, so that the marginal distribution of responses becomes negative-binomial distribution.

Like in cross-sectional Poisson and negative binomial regression, a log link function is commonly assumed for marginal mean and association models. Afterwards, the marginalized multilevel model of Griswold and Zeger (2004) for count data is named as Log-Log-Gamma MMM where log stands for common link function of marginal mean and association models, and Gamma refers to the distribution of random effects. The formulation of the Log-Log-Gamma MMM is as follows:

i) Marginal Mean Model: $\log(\mu_{ij}^{M}) = X_{ij}\beta^{M}$

ii) Association Model: $\log(\mu_{ij}^{C}) = \Delta_{ij} + \log(b_{ij})$

iii) Random Effects Distribution: $b_{ij} \square \text{Gamma}(\theta_1, \theta_2)$

iv) Conditional Response Distribution: $Y_{ij}^{C} = (Y_{ij} \mid b_{ij}) \square \text{Poisson}(\mu_{ij}^{C})$

By the relationship between marginal and conditional mean, we have:

$$\mu_{ij}^{M} = \int_{b} \mu_{ij}^{C}\, dF(b),$$

$$\Rightarrow \exp(X_{ij}\beta^{M}) = \int_{b} \exp\{\Delta_{ij} + \log(b_{ij})\}\, dF(b),$$

$$\Rightarrow X_{ij}\beta^{M} = \log\left(\int_{b} \exp\{\Delta_{ij} + \log(b_{ij})\}\, dF(b)\right), \qquad (3.2)$$

$$\Rightarrow (X_{ij}'X_{ij})\beta^{M} = X_{ij}'\, \log\left(\int_{b} \exp\{\Delta_{ij} + \log(b_{ij})\}\, dF(b)\right),$$

$$\Rightarrow \beta^{M} = (X_{ij}'X_{ij})^{-1}X_{ij}'\, \log\left(\int_{b} \exp\{\Delta_{ij} + \log(b_{ij})\}\, dF(b)\right).$$

The parameter Δ_{ij} in (3.2) can be solved analytically. The log-link function and Poisson-gamma mixing distribution, together with the connection between marginal mean and conditional mean model, lead to $\Delta_{ij} = X_{ij}\beta^{M} - \log(\nu_{ij})$ where $E(b_{ij}) = \theta_1 \times \theta_2 = \nu_{ij}$ (Griswold and Zeger, 2004). Hence, the conditional mean, μ_{ij}^{C}, can be written in terms of the marginal regression parameters, β^{M}, such that

$$\mu_{ij}^{C} = \exp\{\Delta_{ij} + \log(b_{ij})\} = \exp\{X_{ij}\beta^{M} - \log(\nu_{ij}) + \log(b_{ij})\}. \qquad (3.3)$$

Since (3.3) includes the marginal regression parameters, β^{M}, the estimation of β^{M} can be performed by fitting the conditional model, μ_{ij}^{C}, via standard NLMM techniques. The regression parameters, β^{M}, describe the effects of covariates on the population averaged mean response, averaging over the random effects. Contrary to GLMMs, random effects in Log-Log-Gamma MMM follow a non-Gaussian distribution, that's Gamma distribution, and are allowed to enter the model nonlinearly.

In addition to fixed effects, we will allow only an intercept coefficient, b_{0i}, to randomly vary from subject to subject in the model. For that reason, to compare the efficiency of the regression estimates produced from this model, we will use the random-intercept model from GLMMs as a competitor regression model.

3.1.2 Random-Intercept Model

As stated in Chapter 2, generalized linear mixed models are the extension of generalized linear models for longitudinal data. They include multivariate normally distributed random effects, in addition to fixed effects, in the linear predictor (Rabe-Hesketh, and Skrondal, 2008). The basic aim of GLMMs is to introduce subject-specific random effects into the linear predictor to represent the natural heterogeneity between subjects. In the context of GLMMs, in the linear predictor of the conditional mean model, in addition to fixed effects, it is possible to assume either a random intercept, b_{0i}, together with random slope(s), $(b_{1i},..., b_{q-1i})$, or only a random intercept, b_{0i}. But, the simplest case of GLMMs is naturally a model with just a random intercept.

The formulation of a random-intercept model for longitudinal count data can be as follows:

i) Conditional Mean Model: $\log(\mu_{ij}^{C}) = X_{ij}\beta + Z_{ij}b_{0ij}$
ii) Random Intercept Distribution: $b_{0i} \sim MVN(0, V)$
iii) Conditional Response Distribution: $Y_{ij}^{C} = (Y_{ij} \mid b_{0ij}) \sim Poisson(\mu_{ij}^{C})$

The Y_{ij}'s are assumed to be conditionally independent given subject-specific random intercepts, $b_{0i} = (b_{0i1}, b_{0i2}, ..., b_{0in_i})'$ and Y_{ij}'s are assumed to have Poisson distribution with conditional mean, μ_{ij}^{C}, depending on both fixed and random effects. The subject-specific random intercepts, $b_{0i} = (b_{0i1}, b_{0i2}, ..., b_{0in_i})'$ are assumed to be independent of the covariates, X_{ij}, and to have a multivariate normal distribution with zero mean and covariance matrix, V. In practice, the normality assumption for random effects may be unrealistic or invalid (Liu and Yu, 2008). Essentially, any multivariate distribution can be assumed for the random effects (Fitzmaurice and Molenberghs, 2008, Chapter 1), but multivariate normality assumption is made for mathematical convenience rather than a strong scientific

ground (Fitzmaurice and Verbeke, 2008, Chapter 2; Liu and Yu, 2008).

When the interest is on the fixed effects regression parameters for GLMMs rather than random effects; as mentioned in Chapter 2, the model fitting and inference requires the maximization of the likelihood of the data. This maximization is obtained by treating random effects, b_i's, as if they were nuisance parameters and by integrating over their distribution (Diggle et al., 2002, Chapter 9). In other words, if the i^{th} subject's contribution to the likelihood of the data is defined as

$$L_i(\beta,V) = \int_{b_i} \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid b_i, \beta)\, f(b_i \mid V)\, db_i \right], \qquad (3.4)$$

and, then, the expression in (3.5) is expected to be maximized

$$L(\beta,V) = \prod_{i=1}^{N} f_i(y_i \mid \beta, V) = \prod_{i=1}^{N} \int_{b_i} \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid b_i, \beta)\, f(b_i \mid V)\, db_i \right]. \qquad (3.5)$$

However, in GLMMs, the integral in (3.5) cannot be solved analytically since normal distribution is not conjugate to Poisson distribution. Approximation methods such as PQL and MQL and numerical integration methods such as Gaussian quadrature and adaptive Gaussian quadrature are proposed for the evaluation of the likelihood. The dimension of the integration is the dimension of the random effects, and the dimension affects the maximization of the integration. Having only one random coefficient, that's intercept, in the model ease the implementation of proposed approximation and numerical integration methods (Diggle et al., 2002, Chapter 9).

Although the normal distribution is not conjugate to Poisson distribution, Cameron and Trivedi (1998, Chapter 9) remark that multivariate normality assumption takes considerable attention in the statistical literature because if results can be obtained for random-intercept GLMMs, then it will be extended to random intercept and random slope GLMMs.

One of the most important characteristics of GLMMs is that they have the ability to accommodate complex within-subject association structures for subject-specific random intercepts. Weiss (2005, Chapter 8) lists a large number of covariance structures and detailed information on these covariance structure specifications, but among them, most commonly used ones are first order autoregressive (AR(1)), compound symmetry (CS), and unstructured (UN). The covariance structure choice can influence the results of analysis as well as the conclusion. The number of parameters, the interpretation of the covariance structure, and effects on fixed effects are some of the considerations, when selecting the covariance structure (Kincaid, 2008).

The glmmPQL function under the MASS library in R, and SAS GLIMMIX procedure, which are used to implement GLMMs for longitudinal count data offer a straightforward fitting of a wide variety of covariance structures including AR(1), CS, and UN. The SAS NLMIXED procedure does not allow a straightforward fitting of these special covariance structures, however after a considerable work, it does. Kincaid (2008) states that there is not a certain method for determining the best covariance structure. However, with the help of the computational techniques mentioned above, it is possible to specify different covariance structures for the model, and, then, to determine the appropriate covariance structure for the model by comparing the output of fit-statistics.

3.2 Missing Data Mechanisms

For a longitudinal study, missingness can occur in some measurements of the response variable Y_i, and/or of the covariates X_i. However, in this thesis, we will assume that missingness occurs only in the values of measurements for the response variable Y_i except the measurements in the first time point and that all values of measurements for the covariates X_i are fully observed.

The question of interest in incomplete longitudinal data is whether the missingness

affects the validity of statistical inference on regression parameters. The effect of missing data on inference depends on the underlying missingness mechanism it has (Carpenter, 2005).

Let the set of responses intended to be collected be $Y_i = (Y_{i1},...,Y_{in_i})'$, which denote the repeated measurements for subject i where i = 1,2,…,N. If there are some missing values for Y_i, then we can partition Y_i into two parts as $Y_i = (Y_i^o, Y_i^m)'$ where Y_i^o refers to observed responses and Y_i^m refers to missing responses. Let $X_i = (X_{i1},..., X_{in_i})'$ refer to the vector of covariates for subject i. The missing values of Y_i can be denoted by the indicators $R_i = (R_{i1},..., R_{in_i})'$ where R_i takes the value of 1 if Y_i is observed, and takes 0 otherwise (Carpenter, 2005; Schafer, 2005; Yucel, 2009).

$$R_i = \begin{cases} 1, & Y_i \text{ is observed} \\ 0, & Y_i \text{ is missing} \end{cases}$$

Then it turns out that missing data mechanisms introduced by Rubin (1976) and Little and Rubin (2002) can now be denoted mathematically as the conditional distribution of R_i given the response Y_i and covariates X_i.

The missing completely at random (MCAR) is defined such that the probability of a response being missing is independent of both observed values and unobserved values.

$$f(R_i \mid Y_i^o, Y_i^m, X_i) = f(R_i)$$

In other words, reasons yielding missingness are not related to the observed or unobserved responses. For example, in a double-blinded randomized clinical trial designed to compare the effectiveness in controlling epileptic seizures in a treatment group with that in a control group, a patient may not attend a scheduled visit due to work related reasons, not because of the reasons related to the study.

Molenberghs and Fitzmaurice (2008, Chapter 17) states that if the data follow a

MCAR type of missingness mechanism, the observed data can be considered as a random sub-sample of the complete data, which is consisted of observed and unobserved data. As a consequence, moments such as sample means, variances, and covariances and the joint distribution of the observed data do not differ from those of the complete data. One of the implications of this is that the completers, the subjects with no missing values, are considered as a random sample of the population. The other implication is that the missing component of subjects who have missing values does not differ from the corresponding components of the completers. For that reason, most statistical methods for longitudinal data analysis which are based on observed data or completers will yield valid inferences under MCAR mechanism.

Second related terminology is missing at random (MAR) conditional on observed data. It is defined such that the probability of a response being missing depends on the observed values, but not on the unobserved values.

$$f(R_i \mid Y_i^o, Y_i^m, X_i) = f(R_i \mid Y_i^o, X_i)$$

This means that reasons yielding missingness are related to the observed values but not related to the unobserved values. For instance, in the epileptic seizure example above, patients in the treatment group may drop out of the study because the treatment is causing an allergic reaction or a slight side-effect such as weight gain.

In contrast to MCAR, when the data follow a MAR conditional on observed data type of missingness mechanism, the observed data cannot be considered now as a random sub-sample of the complete data. In fact, the observed data are now a biased sample of the complete data. Therefore, sample means, variances, and covariances of the observed data are now biased estimates of those of the complete data. One of the implications of this is that the completers now cannot be considered as a random sample of the population. The other implication is that missing component of subjects that have missing values does differ from the corresponding components of the completers. As a consequence, certain

statistical methods for longitudinal data analysis which are based on observed data or completers will no further yield valid inferences under MAR conditional on observed data mechanism.

Bennett (2001) states that the missing values can be predicted from the observed data under MAR conditional on observed data mechanism and that the inference on the population parameter of interest does not depend on missing data mechanism, if the data is MAR conditional on observed data. For that reason, this type of missing data mechanism is also named as “ignorable”. Since MCAR is a special case of MAR conditional on observed data, MCAR can also be defined as an ignorable mechanism.

The last mechanism is not missing at random (NMAR). It is defined such that the probability of a response being missing depends on the unseen observations themselves even after accounting for all the available observed information.

$$f(R_i \mid Y_i^o, Y_i^m, X_i) = f(R_i \mid Y_i^o, Y_i^m, X_i)$$

In the example of epileptic seizure study, missingness is considered not at random, if a patient cannot go to hospital to report the number of seizures he had during that week, because he is too sick due to experiencing epileptic seizures in large numbers.

Missing values cannot be predicted from the observed data under NMAR mechanism. Inference on the population parameters of interest depends not only on a model for the data, but also on a model for the process that cause missing values in the data. For that reason, this type of missing data mechanism is referred to as “non-ignorable” in the likelihood setting, which means it cannot be ignored from the analysis. Therefore, most statistical methods for longitudinal data analysis which are based on the observed data will yield invalid inferences under MNAR mechanism. It is crucial that the effect of missing data mechanism should be considered in statistical analysis, otherwise results will be biased, and

amount of variability in the data cannot be estimated precisely.

3.3 Missing Data Techniques

3.3.1 Methods that ignore missing values

Before 1980's, missing data was considered as something to be gotten rid of (Yucel, 2009). For that reason, firstly, methods ignoring missing values in the data are developed. In the context of methods ignoring missing values, we will deal with complete case analysis.

Complete case analysis: In this method, all the subjects with missing values are extracted from the study, and statistical analysis is limited to only completers. If the underlying mechanism in the data is MCAR, the results of the statistical analysis will produce valid estimates since reduced data, namely the completers, would represent a randomly drawn sub-sample of the complete data. One disadvantage of complete case analysis is loss of power due to using smaller dataset.

3.3.2 Single imputation methods

Imputation, one of the most commonly used methods in handling missing data, refers to substituting or filling in missing values with imputed values. Single imputation methods estimate the values of missing data rather than ignoring it. This method prefers to "fill-in" or "impute new values" for the missing data and then treats the data as if it is complete. Thereby, it turns out to be possible to perform any statistical methods on this complete data. One disadvantage of single imputation is that it lacks of sampling variability since it imputes only one constant value for each missing value under one model.

There are many single imputation techniques which are appropriate for incomplete longitudinal data. To impute the missing value of a subject, some methods use only the information of that subject with missing data, while others use the information

of other subjects.

Let's assume an imaginary longitudinal dataset such that each row of the dataset corresponds to subjects (i = 1,2,…,N) and that each column corresponds to the repeated measures of the response, namely occasions, (j = 1,2,…,n_i). Let y_{ij} be response for subject i at occasion j and r_{ij} = 1 if y_{ij} is observed, 0 if missing.

When y_{ij} is missing, "before data" method uses the mean of all of the previously observed values, y_{i1}, $y_{i2,}$…, y_{ij-1} prior to the missing value, y_{ij} (Engels and Diehr, 2003) to impute the y_{ij}.

$$\hat{y}_{ij} = \frac{\sum_{l=1}^{j-1} r_{il} y_{il}}{\sum_{l=1}^{j-1} r_{il}}$$

A special case of "before data" method is last observation carried forward (LOCF) method, which is a commonly used imputation technique in longitudinal data. It replaces y_{ij+1}, y_{ij+2},…,y_{ini} by y_{ij} assuming that there is no change from occasion j+1 to occasion n_i. This method can be used when a subject drops out of the study after occasion j and there are no data thereafter. Another method which uses "after data" information is next observation carried backward (NOCB). It replaces the missing value, y_{ij}, with the subject's first next known value, $y_{ij+1,}$ after the missing value, y_{ij}. This method is used when a subject fails to complete baseline information (McKnight et al., 2007). LOCF and NOCB both use subject's own information to impute the missing value. For both methods, the distance between the missing value and the recent observation that will be used to impute that missing value is a common problem.

However, in this thesis, we will allow a variety of patterns of missingness to occur in data. For that reason, we will focus on two mean imputation techniques: subject mean and occasion mean imputation, and one conditional mean imputation.

Subject mean imputation: This method uses the information specific to the subject. The mean of known values in row i, y_{i1}, $y_{i2},\ldots,y_{ij-1}$, $y_{ij+1},y_{ij+2},\ldots,y_{ini}$ is used to replace the missing value, y_{ij} (Schafer, 2005).

$$\hat{y}_{ij} = \overline{y}_{i.} = \frac{\sum_{j=1}^{n_i} r_{ij} y_{ij}}{\sum_{j=1}^{n_i} r_{ij}} \tag{3.6}$$

Occasion mean imputation: This method does not use the information specific to the subject. Engels and Diehr (2003) also name this method as "no person data". The mean of known values in column j of that occasion, $y_{1j},y_{2j},\ldots.,y_{i-1j}$, $y_{i+1j},y_{i+2\,j},\ldots\ldots,y_{Nj}$ is used to replace the missing value, y_{ij} (Schafer, 2005).

$$\hat{y}_{ij} = \overline{y}_{.j} = \frac{\sum_{i=1}^{N} r_{ij} y_{ij}}{\sum_{i=1}^{N} r_{ij}} \tag{3.7}$$

If data is available for a particular occasion for some subjects, but not available for other subjects, then this method can be preferred. This method is reasonable when the data is MCAR but it is found that this method underestimates the variance (Bennett, 2001). Molenberghs and Verbeke (2005, Chapter 27) states that occasion mean imputation is developed mainly for continuous responses. This is also true for subject mean imputation. However, we will use rounding to convert continuous values to count values (See Appendix E).

Unlike the single imputation techniques mentioned above, there are single imputation techniques which take into account the subject's covariate information. Well-knowns are hot-deck imputation and cold-deck imputation techniques which take their origins from survey research. Hot-deck imputation groups subjects which are similar with respect to covariates. Then, the method replaces the missing values with the values of subjects whose covariates are matched, or similar. But,

complex matching algorithms are needed to be employed to match respondents. Cold-deck imputation is very similar to hot-deck imputation method. It decides subject similarity with respect to external information or knowledge of previous studies, not with respect to the information available in the dataset. It raises doubts if this external information is of good quality (Bennett, 2001).

The covariate based single imputation method, which we will also use in this thesis, is conditional mean imputation, in other words, regression imputation.

Conditional mean imputation: This method firstly requires removing all subjects with missing values in the dataset, that's complete case analysis. Afterwards, in the reduced dataset, this method regresses the responses at each occasion on the corresponding covariates. The resulting fitted regression equations for each occasion are then used to estimate the missing values in that occasion in the dataset. For this thesis, we use an AR(1) model as stated in Chapter 2. Since our response variable represents counts, we will use the log-link function to relate the responses to the covariates and the previous response. The conditional mean model for the response at the j^{th} occasion depends on the response at the $(j-1)^{th}$ occasion, as well as the covariates, such that

$$\begin{aligned} &E(Y_{ij} \mid X_{ij}, Y_{i\,j-1}) = \mu_{ij} \text{ and} \\ &\log(\mu_{ij}) = X_{ij}\beta + Y_{i\,j-1}. \end{aligned} \tag{3.8}$$

The missing values at the j^{th} occasion in the dataset can be estimated through the fitted regression equation of (3.8) for the j^{th} occasion. Contrary to other single imputation methods, this method yields less biased estimators under MAR conditional on observed data.

Implementation of subject mean, occasion mean and conditional mean imputations is easy in any statistical software. However, common disadvantage of all these single imputation methods is that it inserts a constant value for each missing value, which leads to the underestimation of the variance.

CHAPTER 4

SIMULATION STUDY

In this chapter, we consider the details of the simulation study. First, we mention about dataset characteristics such as sample size, within-subject size, and response variable, covariates, with their statistical distributions. Data generation processes under each model and true values of the parameters for each model, which are required to simulate the data, are stated in detail. Afterwards, to create missingness in the data, missing data generation scenarios are determined for each missing data mechanisms (MCAR, MAR conditional on observed data, and NMAR). Lastly, parameter estimation techniques under each model are covered.

4.1 Data Generation Scenarios

The examples in the introduction of Chapter 1 reveal that subjects of interest in longitudinal studies may be countries, cities, firms, nations, patients and so on. In this thesis, we will create a longitudinal count dataset similar to the popular epileptic seizure data (Leppik et al., 1985), which assume that subjects of interest are patients who suffer from epileptic seizures. Assuming that data collection is based on periodic scheduled visits in a hypothetical clinical trial, the total number of subjects is taken as N = 100, indexed by i = 1,2,…,100. The number of repeated measurements per subject, n_i, is assumed to be constant for all subjects and $n_i = n$ is equal to 4, indexed by j = 1, 2, 3, 4. The time interval between two consecutive repeated measurements, t_{ij}, is assumed to be same for both within subjects and between subjects, and $t_{ij} = t$ is equal to one year. Thereby, we can define the response variable of interest, Y_{ij}, as the total number of seizures that each patient experienced within one year over four successive years. Since measurements

are reported by subjects at visits, we can say that the longitudinal data are collected from all subjects every t = 1 year for n = 4 visits, see Figure 4.1. Since all subjects have equal number of repeated measurements, n = 4, and all subjects have measurements taken at the same time interval, t = 1 year for all j, we will have a balanced longitudinal study design (Weiss, 2005, Chapter 1). Thus, we prevent the complexities, which an unbalanced study design may cause.

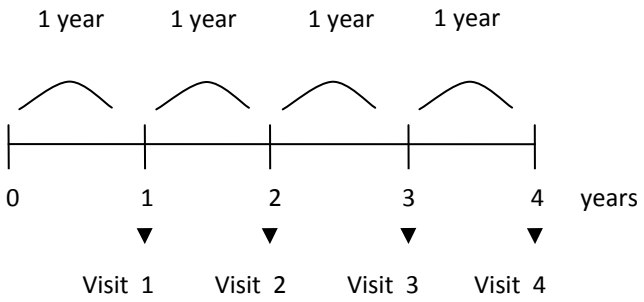

Figure 4.1 Data collection of repeated measurements for a subject

The data structure of the longitudinal response is shown in Table 4.1.

Table 4.1 Data structure of the longitudinal response

	Measurement Occasions			
Patient	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 4**
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}
2	Y_{21}	Y_{22}	Y_{23}	Y_{24}
…	…….	…….	…….	…….
99	$Y_{99,1}$	$Y_{99,2}$	$Y_{99,3}$	$Y_{99,4}$
100	$Y_{100,1}$	$Y_{100,2}$	$Y_{100,3}$	$Y_{100,4}$

Along with the longitudinal response variable, two time-independent covariates, and a time-dependent covariate, totally three covariates, are determined to relate the changes in mean response.

The first time-independent covariate, X_{1i}, is assumed to be a continuous variable

and it is generated from uniform distribution within the range of -1 and 1, for each subject.

$$X_{1i} \sim \text{Unif}(-1, 1)$$

The second time-independent covariate, X_{2i}, is assumed to be a discrete variable and is generated for each subject from binomial distribution, taking value of 0 or 1 with probability 0.5.

$$X_{2i} \sim \text{Bin}(1, 0.5)$$

Contrary to the time-independent covariates, whose values are fixed over time periods, the values of the time-dependent covariate change over time points. Since the data are collected over four time points, the measurements of X_{3i} can be expressed as

$$X_{3i} = (X_{3i1}, X_{3i2}, X_{3i3}, X_{3i4})', \text{ for each i.} \quad (4.1)$$

For that reason, for each subject, the time-dependent covariate, X_{3i}, is generated from a multivariate normal distribution with zero mean and a common AR(1) within-subject covariance structure.

In an AR(1) within-subject covariance matrix, the measurements which are closer in time are expected to be more correlated than measurements which are farther in time. An AR(1) covariance structure has two parameters, τ^2 and ρ. Here, τ^2 is the common variance of X_{3ij} for all i and j. The correlation between two measurements of a subject i, X_{3ij} and X_{3im}, is a function of the absolute value of the distance between the time points of them, so that

$$\text{Corr}(X_{3ij}, X_{3im}) = \rho^{|t_{ij} - t_{im}|},$$

where $0 < \rho < 1$. In practice, correlation between the two measurements of a longitudinal study subject is rarely found to be negative, for that reason, in

longitudinal studies, ρ is expected to range from 0 to 1 (Weiss, 2002, Chapter 8).

The within-subject covariance matrix of (4.1) is given by

$$\text{Cov}(X_{3i}) = \tau^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}, \text{ for each i,} \qquad (4.2)$$

and hence the statistical distribution of (3.1) becomes

$$X_{3i} \sim \text{MVN}_4 \left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tau^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \right], \text{ for each i.} \qquad (4.3)$$

Since we do not allow missingness in the covariates, the measurements of the time-invariant covariates, X_{1i} and X_{2i}, are assumed to be observed at study entry. Kalbfleisch and Prentice (1978) classify time-variant covariates into two categories as internal and external covariates. If the changes in the values of the time-variant covariate over time periods are affected by the subject's development over time periods, then the time-variant covariate is called as internal covariate. For example, medical sources such as the dosage of the antiepileptic drug that subject is using, and the number of antiepileptic drugs that subject is using, or metabolic sources such as lead level in blood and fever level can be considered as internal covariates. However, if the changes are not related to the subject's development over time periods, then it is called as external covariates. For instance, environmental factors such as the carbon monoxide level in the air, and toxins. Our time-variant covariate, X_{3i}, is assumed to be an external one and measurement process on the time-variant covariate, X_{3i}, can continue even if subject drops out of the study (Daniels and Hogan, 2008, Chapter 5).

The data structure of these longitudinal covariates may be illustrated as in Table 4.2.

Table 4.2 Data structure of longitudinal covariates for each subject

	Measurement Occasions			
Patient	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 4**
1	$X_{11}\ X_{21}\ X_{3,1,1}$	$X_{11}\ X_{21}\ X_{3,1,2}$	$X_{11}\ X_{2,1}\ X_{3,1,3}$	$X_{11}\ X_{21}\ X_{3,1,4}$
2	$X_{12}\ X_{22}\ X_{3,2,1}$	$X_{12}\ X_{22}\ X_{3,2,2}$	$X_{12}\ X_{22}\ X_{3,2,3}$	$X_{12}\ X_{22}\ X_{3,2,4}$
…	…….	……	……	……
99	$X_{1,99}\ X_{2,99}\ X_{3,99,1}$	$X_{1,99}\ X_{2,99}\ X_{3,99,2}$	$X_{1,99}\ X_{2,99}\ X_{3,99,3}$	$X_{11}\ X_{2,99}\ X_{3,99,4}$
100	$X_{1,100}\ X_{2,100}\ X_{3,100,1}$	$X_{1,100}\ X_{2,100}\ X_{3,100,2}$	$X_{1,100}\ X_{2,100}\ X_{3,100,3}$	$X_{1,100}\ X_{2,100}\ X_{3,100,4}$

The data generation of the response variable, Y_{ij}, is based on the corresponding regression model, where the matrix of covariates and the vector of fixed effects are assumed to be same for both models. For simulation studies, R programming language (version 2.8.1) is used. The R codes for data generation procedures are available in Appendix A-C.

4.1.1 The Log-Log-Gamma MMM

For the Log-Log-Gamma MMM, random variables, Y_{ij}, are generated for each subject from Poisson distribution with conditional mean, μ_{ij}^{C}, given by

$$E(Y_{ij} \mid b_{0ij}) = \mu_{ij}^{C} = \exp(\beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3ij} - \log(\nu_{ij}) + \log(b_{0ij}))\,,$$

$$b_{0ij} \sim \text{Gamma}(\theta_1, \theta_2) \text{ and } \nu_{ij} = E(b_{0ij})\,,$$

where Y_{ij} is the response variable of the j^{th} measurement for the i^{th} subject, $\beta_0, \beta_1, \beta_2$, and β_3 are the fixed effects regression coefficients, which are common for all subjects, X_{0i} is constant, and equal to 1, X_{1i}, and X_{2i} are the time-independent covariates for subject i, X_{3ij} is the time-dependent covariate of j^{th} measurement for the i^{th} subject, b_{0ij} is the gamma distributed random intercept coefficient for the i^{th}

subject, $\log(b_{0ij})$ is the natural logarithm of b_{0ij}, v_{ij} is the expected value of b_{0ij}, and $\log(v_{ij})$ is the natural logarithm of v_{ij}.

4.1.2 The Random-Intercept Model

For the random-intercept model, random variables, Y_{ij}, are generated for each subject from Poisson distribution with conditional mean, μ_{ij}^{c}, given by

$$E(Y_{ij} \mid b_{0ij}) = \mu_{ij}^{c} = \exp(\beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3ij} + b_{0ij}) \text{ and } b_{0i} \sim MVN_4(0, V),$$

where Y_{ij} is the response variable for the j^{th} visit of the i^{th} subject, β_0, β_1, β_2, and β_3 are the fixed-effect regression coefficients, which are common for all subjects, X_{0i} is constant, and equal to 1, X_{1i}, and X_{2i} are the time-independent covariates for subject i, X_{3ij} is the time-dependent covariate for j^{th} of n measurements in the i^{th} subject, and $b_{0i} = (b_{0i1}, b_{0i2}, b_{0i3}, b_{0i4})'$ are the random intercept coefficients for the i^{th} subject, assumed to be multivariate normally distributed with zero mean and AR(1) within-subject covariance structure, V, which is common for all subjects.

4.2 True Parameter

Section 4.1 mentions about the scenarios on the generation of the covariates. While two time-fixed covariates can be easily generated from uniform and binomial distribution, the generation of the time-varying covariate requires firstly assuming true values for the within-subject covariance parameters. After setting $\tau^2 = 1$ and $\rho = 0.9$, (4.2) turns out to be

$$\text{Cov}(X_{3i}) = \begin{pmatrix} 1.000 & 0.900 & 0.810 & 0.729 \\ 0.900 & 1.000 & 0.900 & 0.810 \\ 0.810 & 0.900 & 1.000 & 0.900 \\ 0.729 & 0.810 & 0.900 & 1.000 \end{pmatrix}, \text{ for each i,}$$

and, in turn, the statistical distribution of (4.1) becomes

$$X_{3i} \sim MVN_4\left[\begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \begin{pmatrix}1.000 & 0.900 & 0.810 & 0.729\\0.900 & 1.000 & 0.900 & 0.810\\0.810 & 0.900 & 1.000 & 0.900\\0.729 & 0.810 & 0.900 & 1.000\end{pmatrix}\right], \text{ for each i.}$$

The fixed effects regression parameters, $\beta = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})'$ in both regression models are set equal to $(2, -1.3, 1.5, 1.2)'$.

4.2.1 The Log-Log-Gamma MMM

For the Log-Log-Gamma MMM, the subject-specific random intercepts, $b_{0i} = (b_{0i1}, b_{0i2}, b_{0i3}, b_{0i4})'$ are generated from Gamma distribution assuming $\theta_1 = 1$ and $\theta_2 = 1$. In other words,

$$b_{0ij} \sim \text{Gamma}(1, 1), \text{ for each i and j.}$$

The model based generation of the response variable, Y_{ij}, at the j^{th} visit can be illustrated as follows

$$\begin{bmatrix}Y_{1j}\\Y_{2j}\\.....\\Y_{99,j}\\Y_{100,j}\end{bmatrix} = \exp\left[\begin{bmatrix}1, X_{11}, X_{21}, X_{3,1,j}\\1, X_{12}, X_{22}, X_{3,2,j}\\.....\\1, X_{1,99}, X_{2,99}, X_{3,99,j}\\1, X_{1,100}, X_{2,100}, X_{3,100,j}\end{bmatrix}\begin{bmatrix}2\\-1.3\\1.5\\1.2\end{bmatrix} - \begin{bmatrix}\log(v_{1j})\\\log(v_{2j})\\.....\\\log(v_{99,j})\\\log(v_{100,j})\end{bmatrix} + \begin{bmatrix}\log(b_{0,1,j})\\\log(b_{0,2,j})\\.....\\\log(b_{0,99,j})\\\log(b_{0,100,j})\end{bmatrix}\right]$$

4.2.2 The Random-Intercept Model

For the random-intercept model, firstly, the within-subject covariance matrix of subject-specific random intercepts, $b_{0i} = (b_{0i1}, b_{0i2}, b_{0i3}, b_{0i4})'$ are generated by setting the parameter $\tau^2 = 0.1$ and $\rho = 0.5$, such that

$$V = Var(b_{0i}) = \begin{pmatrix} 0.1000 & 0.050 & 0.025 & 0.0125 \\ 0.0500 & 0.100 & 0.050 & 0.0250 \\ 0.0250 & 0.050 & 0.100 & 0.0500 \\ 0.0125 & 0.025 & 0.050 & 0.1000 \end{pmatrix}, \text{ for each i,}$$

and, in turn, the statistical distribution of b_{0i} becomes

$$b_{0i} \sim MVN_4\left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{pmatrix} 0.1000 & 0.050 & 0.025 & 0.0125 \\ 0.0500 & 0.100 & 0.050 & 0.0250 \\ 0.0250 & 0.050 & 0.100 & 0.0500 \\ 0.0125 & 0.025 & 0.050 & 0.1000 \end{pmatrix}\right], \text{ for each i.}$$

The model based generation of the response variable, Y_{ij}, at the j^{th} visit can be illustrated as follows

$$\begin{bmatrix} Y_{1j} \\ Y_{1j} \\ \ldots\ldots \\ Y_{99,j} \\ Y_{100,j} \end{bmatrix} = \exp\left[\begin{bmatrix} 1, & X_{11}, & X_{21}, & X_{3,1,j} \\ 1, & X_{12}, & X_{22}, & X_{3,2,j} \\ & & \ldots\ldots\ldots\ldots\ldots\ldots & \\ 1, & X_{1,99}, & X_{2,99}, & X_{3,99,j} \\ 1, & X_{1,100}, & X_{2,100}, & X_{3,100,j} \end{bmatrix}\begin{bmatrix} 2 \\ -1.3 \\ 1.5 \\ 1.2 \end{bmatrix} + \begin{bmatrix} b_{0,1,j} \\ b_{0,2,j} \\ \ldots\ldots \\ b_{0,99,j} \\ b_{0,100,j} \end{bmatrix}\right]$$

So that, based on the true values mentioned above, under each regression model, 120 longitudinal count datasets are generated with the same data structure (each with 100 subjects and 4 repeated measurements for response variable, two time-independent covariates, and a time-dependent covariate) and saved.

4.3 Missing Data Generation Scenarios

As noted in Section 3.2, in this thesis we assume that missingness occurs only in 2^{nd}, 3^{rd}, or 4^{th} measurements of the response variable, $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ for all i and that no missingness occurs in the measurements of the matrix of covariates, X_i and in the 1^{st} measurement of the response variable. We allow any drop-out and intermittent missing data patterns in the measurements of the response variable,

$Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$. Hence, we may have 8 different missing data patterns in the measurements of the response variable of a subject. For convenience, the patterns are classified as completers, intermittent missingness (non-monotone missingness), and drop-outs (monotone missingness) and are depicted in Table 4.3, where O = Observed and M = Missing.

Table 4.3 Conventional classification of missing data patterns

	Measurement Occasions			
Missing Data Patterns	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 4**
	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}
Completers	O	O	O	O
Intermittent missingness	O	M	O	O
	O	O	M	O
	O	M	O	M
	O	M	M	O
Drop-out	O	O	O	M
	O	O	M	M
	O	M	M	M

As stated at the end of the Chapter 1, and reviewed in Section 3.2, we aim to create missingness in measurements of a dataset like Table 4.1, according to three missing data mechanisms, i.e. MCAR, MAR conditional on observed data, and NMAR. Scheffer (2002), Shieh (2003), Fong and Lam (2005), and Yucel (2009) provide interesting simulation scenarios about these mechanisms.

Similar to them, three missing data mechanisms are applied on the saved original datasets with generating 10% and 20% missingness in the dataset. Since, there are 400 total observations for 100 subjects with measurements taken over four periods; for 10% missingness, 40 observations are deleted, and for 20% missingness, 80 observations are deleted. No observations are deleted from the measurements of the first visit, since all subjects are assumed to have an observation in the first year.

The scenarios developed for three missing data mechanisms are:

i) In MCAR case, missing data is obtained with random deletion, so that any

observation is missing independently from other variables.

ii) In MAR conditional on observed data case, missingness is limited to observations of subjects whose X_{2i} value is 1.

iii) In NMAR case, i) an observation at the second visit is more likely to be missing if the observed response at the first visit is greater than 20, ii) an observation at the third visit was more likely made missing if the observed value at the second visit is greater than 25, and iii) an observation at the fourth visit was more likely to be missing if the observed value at the third visit is greater than 30.

For the NMAR case, we create a scenario similar to the NMAR example in Section 3.2. We expect that subjects, experiencing seizures larger than 20 after the first visit, do not return for the second visit since they are too sick. If they return for the second visit, then, again we suppose that subjects, experiencing seizures larger than 25 after the second visit, do not come back for the third visit due to sickness. However, if they come back for the third visit, lastly, once again, we assume that subjects, experiencing seizures larger than 30 after the third visit, do not return for the fourth visit due to extensive seizures.

This yields 3 missing data mechanisms $\times$ 2 types of missingness percentage = 6 different conditions for each regression model. Then, complete case analysis in Section 3.3.1, single imputation methods: occasion mean imputation, subject mean imputation, and conditional mean imputation in Section 3.3.2 are applied, in order, onto the so-called incomplete datasets which are saved safely. To impute the values of missing observations, the equations (3.6), (3.7), and (3.8) are used.

This also yields 3 missing data mechanisms $\times$ 2 types of missingness percentage $\times$ 4 missing data techniques = 24 different conditions for each regression model as seen in Figure 4.2.

The R codes for missing data generation scenarios, missing data techniques and

longitudinal data format are available in Appendix D, E and F, respectively. Now, under each regression model, the statistical evaluation of 120 datasets in each 24 conditions can be carried out.

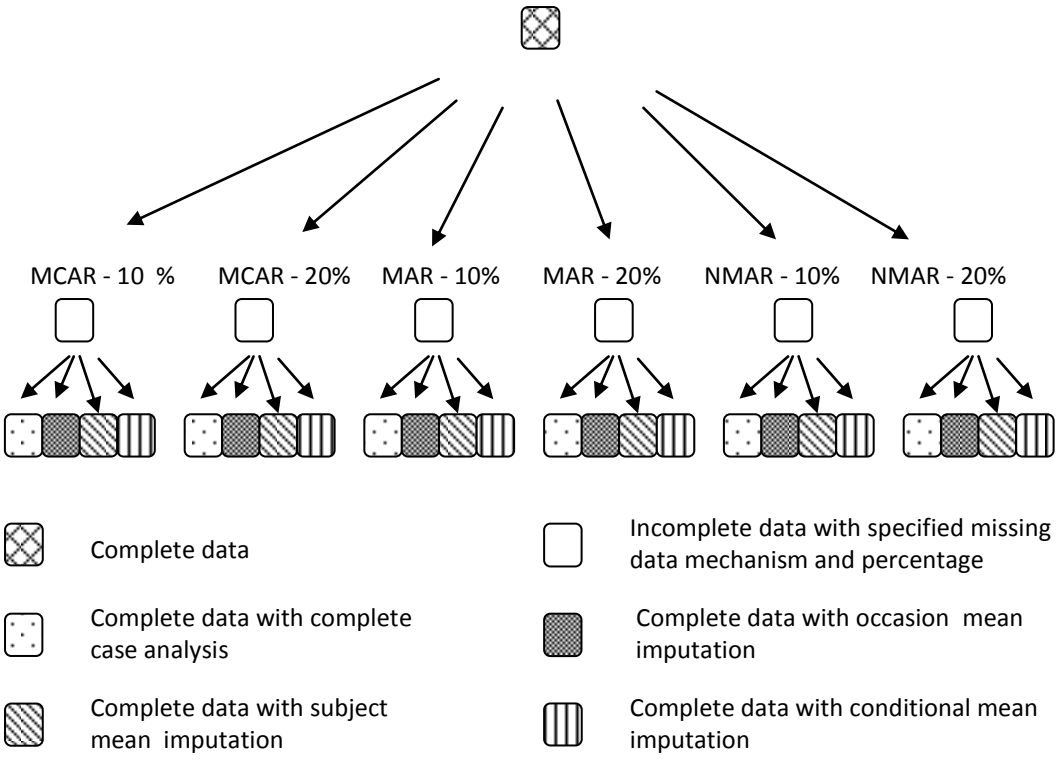

Figure 4.2 Schematic display of the simulation process

4.4 Parameter Estimation

As stated throughout the thesis, the aim is to compare the performance of the regression parameters of both regression models in 24 different conditions. Due to the special features of the Log-Log-Gamma MMM and the random-intercept model which were reviewed in detail in Sections 3.1.1 and 3.1.2, the implementation of these models, unfortunately, lacks computational tools. For that reason, for parameter estimation of the regression models, we used SAS version 9.1.3, which was the recent version available in Turkey as of the date this thesis was written. While SAS NLMIXED procedure is used for the Log-Log-Gamma MMM like

Griswold and Zeger (2004), SAS GLIMMIX procedure is preferred for the random-intercept model.

The NLMIXED procedure is a built-in SAS procedure, whereas the GLIMMIX procedure is an add-on procedure in SAS 9.1.3 which requires to be downloaded from the web site of SAS.

The NLMIXED procedure is an appropriate choice for nonlinear mixed models, in which random effects are allowed to enter the linear predictor of the model nonlinearly. It specifies the conditional distribution for the response variable given the random effects, either by standard distributions such as normal, binomial, and Poisson or by general distributions that can be coded using SAS programming statements. The only distribution available for random effects is normal distribution. The way of model specification in the NLMIXED procedure has a high degree of flexibility, compared to other SAS procedures (Molenberghs and Verbeke, 2005, Chapter 15). This advantage enables any non-normal distribution of interest for random effects to be implemented within the numerical integration techniques available in the NLMIXED procedure via the probability integral transformation (Nelson et al., 2006). As stated at the end of the Section 3.1.2, when the random effects are normally distributed, the NLMIXED procedure does not offer a straightforward option for the specification of any within-subject covariance structure. But, by the help of the its flexibility, it is possible to allow the within-subject covariance matrix of the random effects to be an AR(1) and the like, when specifying the mean and covariance components of the normal distribution (Molenberghs and Verbeke, 2005, Chapter 1; Xu et al., 2007). Xu et al. (2007) states that the NLMIXED procedure in SAS 9.1.3 allows only up to three random effects per subject, which is a significant disadvantage of the NLMIXED procedure in SAS 9.1.3.

On the other hand, the GLIMMIX procedure is an appropriate choice for generalized linear mixed models, in which random effects are restricted to enter linear predictor linearly. It specifies the conditional distribution for the response

variable given the random effects to have any distribution in the exponential family, and only normal distribution for random effects. As stated at the end of Section 3.1.2, any within-subject covariance structure for normally distributed random effects can be modeled directly in the GLIMMIX procedure. This procedure is especially recommended for models when the number of random effects per subject is large (Flom et al., 2006). The reason why we have used two different SAS procedures for these models is that while the NLMIXED procedure accommodates the Log-Log-Gamma MMM, it could not handle the random-intercept model, since the number of random intercept coefficients per subject in our model is four, which is beyond the NLMIXED procedure capacity. In a similar fashion, while the GLIMMIX procedure perfectly accommodates the random-intercept model, it cannot handle the Log-Log-Gamma MMM, since it does not allow random effects to have a distribution other than normal distribution.

However, the most essential difference between the two SAS procedures is the estimation techniques that they use (Flom et al., 2006). The likelihood of the data can be written as

$$L(\beta \mid y_{ij}, b_{ij}) = \prod_{i=1}^{N} f_i(y_i \mid \beta, b_i) = \prod_{i=1}^{N} \int_{b_i} \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid b_{ij}, \beta)\, f(b_i \mid \theta)\, db_i \right], \qquad (4.4)$$

where (4.4) is the general version of (3.5) for any random-effects model and θ is the vector of parameters for the distribution of b_i. The estimation in the NLMIXED procedure is based on maximizing the likelihood in (4.4). The maximization requires the computation of the integrals in (4.4) over the distribution of random effects. However, generally it does not provide an analytical solution for the maximization in (4.4). The NLMIXED procedure computes the integrals in (4.4) by numerical integration methods such as Gaussian quadrature or adaptive Gaussian quadrature.

Within the framework of NLMIXED procedure, to fit the Log-Log-Gamma MMM we make the model specification parallel to Griswold and Zeger (2004) and to accommodate gamma distributed random effects, we use probability

integral transformation (PIT) like Nelson et al. (2006). Similar to them, a_i is assumed to be a random effect from standard normal distribution, such that $a_i \sim N(0,1)$, and then by the use of PIT, it can be shown that $\Phi(a_i) = u_i \sim Unif(0,1)$ where $\Phi(.)$ is the cumulative distribution function (CDF) of the standard normal distribution. Again by the help of PIT, it can also be shown that $F_\theta(b_i) = u_i \sim Unif(0,1)$ where $F_\theta(.)$ is the cumulative distribution function (CDF) of the gamma distribution of b_i, with $\theta = (\theta_1, \theta_2)$. Then it turns out that $b_i = F_\theta^{-1}(u_i) = F_\theta^{-1}(\Phi(a_i))$ has the gamma distribution of interest, where $F_\theta^{-1}(.)$ is the inverse CDF of gamma distribution. The equation (4.4) can now be rewritten in terms of random effects, a_i, such that

$$L(\beta \mid y_{ij}, a_i) = \prod_{i=1}^{N} f_i(y_i \mid \beta, a_i) = \prod_{i=1}^{N} \int_{b_i} \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid F_\theta^{-1}(\Phi(a_i)), \beta)\, \phi(a_i)\, da_i \right], \qquad (4.5)$$

where $\phi(.)$ is the standard normal distribution density function. Nelson et al. (2006) suggest that the likelihood in (4.5) can be approximated by the Gaussian quadrature numerical integration technique well. The approximation with Gaussian quadrature to integrals in (4.5) is achieved such that i^{th} subject's likelihood is approximated by a weighted sum

$$L_i(\beta \mid y_{ij}, a_i) = \int_{a_i} \left[\prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid F_\theta^{-1}(\Phi(a_i)), \beta)\, \phi(a_i)\, da_i \right]$$

$$\approx \sum_{q=1}^{Q} \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid F_\theta^{-1}(\Phi(z_q)), \beta)\, \phi(z_q)\, w_q ,$$

and, thus, the likelihood which is expected to be maximized turns out that

$$L(\beta \mid y_{ij}) = \prod_{i=1}^{N} \sum_{q=1}^{Q} \prod_{j=1}^{n_i} f_{ij}(y_{ij} \mid F_\theta^{-1}(\Phi(z_q)), \beta)\, \phi(z_q)\, w_q ,$$

where z_q is quadrature point and indexed by q = 1,...,Q, Q is the order of

approximation, w_q is the standard Gauss-Hermite weight. Since the approximations will be more accurate as the Q increases, we use Gaussian quadrature with 30 points like Griswold and Zeger (2004) and Nelson et al. (2006). The values of z_q and w_q can be obtained from tables (Abramowitz and Stegun, 1972, Table 25.10).

Gaussian quadrature with 40 and 50 points are also tried, but no significant difference are observed, compared to Gaussian quadrature with 30 points.

The NLMIXED procedure requires specification of initial values for all parameters in the model. However, one limitation of the NLMIXED procedure occurs when specifying gamma distributed random effects. This procedure only allows θ_2 being equal to 1 (Nelson et al., 2006). This limitation causes us to assume θ_1 is equal to θ_2 and to assign only the value of 1 to θ_1 so that we provide the value of 1 to θ_2. For detail, we refer the reader to the SAS codes which are given in Appendix G.

On the other side, the GLIMMIX procedure is based on the linearization of the general linear mixed models, that's it transforms the GLMM into a linear mixed model such that,

$$Y_{ij} = \exp(X_{ij}\beta + Z_{ij}b_{ij}) \Rightarrow Y_{ij} \approx (X_{ij}\beta + Z_{ij}b_{ij})$$

Linearization of Y_{ij} is achieved by expanding $\exp(X_{ij}\beta + Z_{ij}b_{ij})$ with some order of the Taylor series around some point. The order of the Taylor approximation, with the point around the approximation is carried out, yield different linearization methods such as PQL and MQL (Molenberghs and Verbeke, 2005, Chapter 14). While both methods are based on a linear Taylor series expansion, MQL differs from PQL in that it completely disregards the random effects in the linearization. The resulting linear mixed model, then, can be fitted by either maximum likelihood estimation or restricted maximum likelihood (REML) (Harville, 1977) estimation.

Within the framework of the GLIMMIX procedure, the random-intercept model is fitted by using PQL, based on REML for the linear mixed models. The option in the GLIMMIX procedure is the “method=RSPL”, which is the default method. For detail, we refer the reader to the SAS codes which are given in Appendix H.

Further information on the description of the NLMIXED and GLIMMIX procedures and their options can be obtained from SAS Institute Inc. (2000) and SAS Institute Inc. (2004).

CHAPTER 5

FINDINGS and DISCUSSION

In this chapter, we draw conclusions from the simulation study, and make several comparisons such as different amount of missing data, missing data mechanisms, and missing data techniques under both Log-Log-Gamma MMM and random-intercept model and mention the general performance of both regression models.

Figures 5.1 and 5.2 show the profile plots, which draw response patterns for each subject, of one of the 120 epileptic seizure counts data generated from the Log-Log-Gamma model, and the random-intercept model, respectively.

For the data of the Log-Log-Gamma model, most of the subject profiles change within the range of 0 and 50, while for the data of the random-intercept model, most change within the range of 0 and 100. Although there are a few exceptions at the top subject profiles in both figures, the bottom subject profiles cross another in both figures.

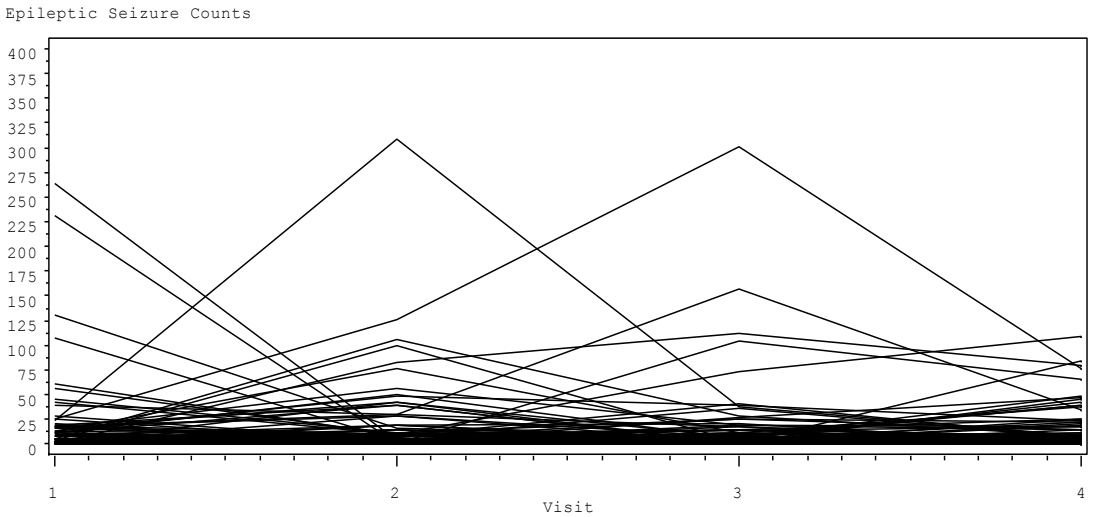

Figure 5.1 An Epileptic seizure count data which is generated under the Log-Log-Gamma model

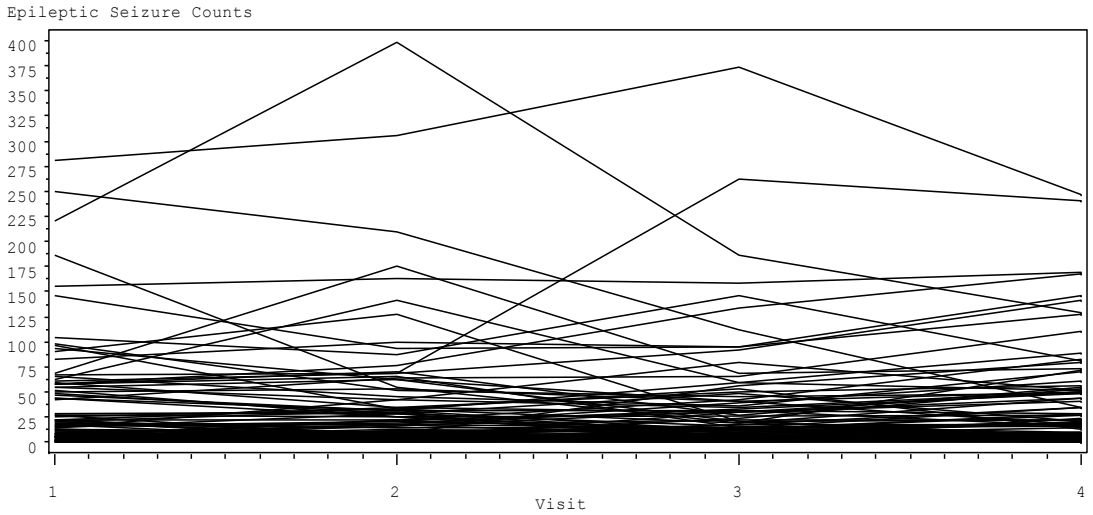

Figure 5.2 An Epileptic seizure count data which is generated under the random-intercept model

Table 5.1 Summary statistics for each visit at each model

	Log-Log-Gamma MMM			
Statistic	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 4**
Mean	15.39	16.19	14.04	12.38
Variance/ Mean	97.304	88.072	99.105	33.301
	Random-Intercept Model			
Statistic	**Visit 1**	**Visit 2**	**Visit 3**	**Visit 4**
Mean	34.76	35.09	32.64	34.90
Variance/ Mean	78.507	104.572	92.442	67.558

Table 5.1 shows that the variance to the mean ratio is pretty larger than 1 at each visit for each regression model. This means that the epileptic seizure counts data for each regression model exhibit a high degree of overdispersion. Furthermore, the magnitude of the overdispersion changes across the visits as well.

The evaluation of regression parameters estimates obtained from of 120 simulated datasets under each 24 conditions (3 missing data mechanisms × 2 types of missingness percentage × 4 missing data techniques) for each regression model is performed by the quantities: Mean absolute error (MAE) and mean square error (MSE). MAE is the average absolute difference between the true value of a parameter and its estimates. MSE is the average squared difference between the

true value of a parameter and its estimates.

We firstly sort the 120 estimates of each regression parameter under each 24 conditions for each model and, then, trim 5% of the smallest estimates from the lower end and 5% of the largest estimates from the upper end. Discarding 6 largest estimates from the upper end and 6 smallest estimates from the lower end results in 108 regression parameters estimates, but this disregards potential outliers, and consequently provides more robust results. Hence, for each regression parameter across the 24 conditions, MAE and MSE values are computed by

$$\text{MAE} = \frac{\sum_{r=1}^{108} |\beta_p - \hat{\beta}_{pr}|}{108} \quad \text{and} \quad \text{MSE} = \frac{\sum_{r=1}^{108} (\beta_p - \hat{\beta}_{pr})^2}{108},$$

where β_p is the true value of the regression parameter of interest indexed by p = 0,1,2,3; r is the simulation index taking values between 1 and 108, and $\hat{\beta}_{pr}$ is the estimate of the p^{th} regression parameter of interest in the r^{th} simulation. The MAE results are tabulated in Tables 5.2 and 5.4, and the MSE results are summarized in Tables 5.3 and 5.5. In addition to rows reserved for missing data techniques, the values in row labeled “complete data” in Tables 5.2 - 5.5 refer to the MAE and MSE values of parameter estimates after the corresponding model is fitted on the complete data which have no missing values. The results are discussed in the following sections in detail.

Table 5.2 Mean Absolute Error of parameters under the Log-Log-Gamma MMM

	Log-Log-Gamma MMM			
	$\beta_0 = 2$		$\beta_1 = -1.3$	
	10%	**20%**	**10%**	**20%**
MCAR				
Complete Data	0.6029		0.2114	
Complete Case	0.4880	0.5046	0.2284	0.2062
Subject Mean Imputation	0.5204	0.4442	0.2208	0.2336
Occasion Mean Imputation	0.1815	0.2173	0.2566	0.3540
Conditional Mean Imputation	0.2103	0.5786	0.2103	0.2192
MAR conditional on observed data				
Complete Data	0.5883		0.2080	
Complete Case	0.6059	0.6098	0.1914	0.1802
Subject Mean Imputation	0.4771	0.3625	0.2401	0.2624
Occasion Mean Imputation	0.4693	0.3991	0.2645	0.2645
Conditional Mean Imputation	0.6142	0.5663	0.2188	0.2024
NMAR				
Complete Data	0.5586		0.2089	
Complete Case	0.6631	0.7770	0.2031	0.1902
Subject Mean Imputation	0.3619	0.2530	0.2392	0.2724
Occasion Mean Imputation	0.5710	0.6532	0.1982	0.2754
Conditional Mean Imputation	0.5973	0.6429	0.2237	0.2033
	$\beta_2 = 1.5$		$\beta_3 = 1.2$	
	10%	**20%**	**10%**	**20%**
MCAR				
Complete Data	0.1813		0.1688	
Complete Case	0.2023	0.2246	0.1830	0.2487
Subject Mean Imputation	0.2296	0.2360	0.1916	0.2962
Occasion Mean Imputation	0.3075	0.3931	0.2129	0.3152
Conditional Mean Imputation	0.2104	0.1926	0.1944	0.1634
MAR conditional on observed data				
Complete Data	0.2362		0.1579	
Complete Case	0.2303	0.2848	0.1745	0.1882
Subject Mean Imputation	0.2538	0.2673	0.2382	0.4175
Occasion Mean Imputation	0.2417	0.2207	0.1780	0.2934
Conditional Mean Imputation	0.2404	0.2341	0.1593	0.1796
NMAR				
Complete Data	0.1954		0.1687	
Complete Case	0.1988	0.2065	0.1942	0.1880
Subject Mean Imputation	0.2161	0.2653	0.3816	0.7348
Occasion Mean Imputation	0.2104	0.2576	0.2256	0.2640
Conditional Mean Imputation	0.2003	0.2673	0.2133	0.2621

Table 5.3 Mean Square Error of parameters under the Log-Log-Gamma MMM

	Log-Log-Gamma MMM				
	$\beta_0 = 2$			$\beta_1 = -1.3$	
	10%	**20%**		**10%**	**20%**
MCAR					
Complete Data	0.3968		0.0656		
Complete Case	0.2967	0.3292		0.0836	0.0722
Subject Mean Imputation	0.3159	0.2363		0.0671	0.0850
Occasion Mean Imputation	0.0469	0.0724		0.0968	0.1857
Conditional Mean Imputation	0.4074	0.3682		0.0677	0.0722
MAR conditional on observed data					
Complete Data	0.3940		0.0667		
Complete Case	0.4107	0.4119		0.0582	0.0530
Subject Mean Imputation	0.2587	0.1799		0.0818	0.0973
Occasion Mean Imputation	0.2736	0.1992		0.1032	0.1160
Conditional Mean Imputation	0.4230	0.3584		0.0741	0.0606
NMAR					
Complete Data	0.3463		0.0581		
Complete Case	0.4663	0.6117		0.0600	0.0518
Subject Mean Imputation	0.1700	0.0937		0.0894	0.1170
Occasion Mean Imputation	0.3782	0.4551		0.0567	0.1062
Conditional Mean Imputation	0.4027	0.4598		0.0813	0.0636
	$\beta_2 = 1.5$			$\beta_3 = 1.2$	
	10%	**20%**		**10%**	**20%**
MCAR					
Complete Data	0.0515		0.0440		
Complete Case	0.0601	0.0722		0.0533	0.1023
Subject Mean Imputation	0.0804	0.0775		0.0569	0.1248
Occasion Mean Imputation	0.1377	0.2226		0.0695	0.1329
Conditional Mean Imputation	0.0678	0.0574		0.0524	0.0412
MAR conditional on observed data					
Complete Data	0.0789		0.0355		
Complete Case	0.0789	0.1151		0.0441	0.0555
Subject Mean Imputation	0.0985	0.1126		0.0779	0.2140
Occasion Mean Imputation	0.0813	0.0705		0.0502	0.1171
Conditional Mean Imputation	0.0801	0.0835		0.0386	0.0465
NMAR					
Complete Data	0.0571		0.0434		
Complete Case	0.0561	0.0616		0.0565	0.0512
Subject Mean Imputation	0.0701	0.1193		0.1756	0.5666
Occasion Mean Imputation	0.0649	0.0926		0.0736	0.1040
Conditional Mean Imputation	0.0613	0.0957		0.0627	0.1014

Table 5.4 Mean Absolute Error of parameters under the Random-Intercept Model

	Random-Intercept Model			
	$\beta_0 = 2$		$\beta_1 = -1.3$	
	10%	**20%**	**10%**	**20%**
MCAR				
Complete Data	0.0442		0.0343	
Complete Case	0.0541	0.0579	0.0436	0.0570
Subject Mean Imputation	0.0667	0.0945	0.0390	0.0536
Occasion Mean Imputation	0.4425	0.7708	0.2708	0.4678
Conditional Mean Imputation	0.0477	0.0577	0.0354	0.0429
MAR conditional on observed data				
Complete Data	0.0481		0.0346	
Complete Case	0.0458	0.0478	0.0387	0.0470
Subject Mean Imputation	0.0888	0.1189	0.0439	0.0656
Occasion Mean Imputation	0.0759	0.1003	0.1940	0.3211
Conditional Mean Imputation	0.0550	0.0576	0.0400	0.0446
NMAR				
Complete Data	0.0411		0.0350	
Complete Case	0.0349	0.0352	0.0430	0.0614
Subject Mean Imputation	0.1008	0.1325	0.0489	0.0818
Occasion Mean Imputation	0.0911	0.1017	0.0666	0.1448
Conditional Mean Imputation	0.0446	0.0510	0.0349	0.0499
	$\beta_2 = 1.5$		$\beta_3 = 1.2$	
	10%	**20%**	**10%**	**20%**
MCAR				
Complete Data	0.0457		0.0415	
Complete Case	0.0544	0.0650	0.0508	0.0513
Subject Mean Imputation	0.0552	0.0718	0.1305	0.2705
Occasion Mean Imputation	0.3014	0.5297	0.1511	0.2714
Conditional Mean Imputation	0.0479	0.0508	0.0436	0.0483
MAR conditional on observed data				
Complete Data	0.0475		0.0412	
Complete Case	0.0574	0.1049	0.0458	0.0510
Subject Mean Imputation	0.0562	0.0966	0.1893	0.4318
Occasion Mean Imputation	0.1324	0.1532	0.1252	0.2229
Conditional Mean Imputation	0.0522	0.0739	0.0476	0.0548
NMAR				
Complete Data	0.0379		0.0390	
Complete Case	0.0522	0.0707	0.0494	0.0609
Subject Mean Imputation	0.0635	0.0926	0.2271	0.5346
Occasion Mean Imputation	0.0887	0.1741	0.1112	0.1888
Conditional Mean Imputation	0.0450	0.0588	0.0531	0.0882

Table 5.5 Mean Square Error of parameters under the Random-Intercept Model

	Random-Intercept Model				
	$\beta_0 = 2$			$\beta_1 = -1.3$	
		10%	**20%**	**10%**	**20%**
MCAR					
Complete Data	0.0027			0.0016	
Complete Case		0.0043	0.0047	0.0028	0.0050
Subject Mean Imputation		0.0062	0.0121	0.0023	0.0043
Occasion Mean Imputation		0.2023	0.6063	0.0803	0.2289
Conditional Mean Imputation		0.0034	0.0048	0.0017	0.0026
MAR conditional on observed data					
Complete Data	0.0033			0.0018	
Complete Case		0.0032	0.0033	0.0023	0.0033
Subject Mean Imputation		0.0100	0.0168	0.0029	0.0065
Occasion Mean Imputation		0.0074	0.0128	0.0411	0.1076
Conditional Mean Imputation		0.0042	0.0044	0.0024	0.0029
NMAR					
Complete Data	0.0025			0.0017	
Complete Case		0.0018	0.0017	0.0026	0.0054
Subject Mean Imputation		0.0123	0.0208	0.0038	0.0098
Occasion Mean Imputation		0.0111	0.0146	0.0061	0.0246
Conditional Mean Imputation		0.0028	0.0041	0.0018	0.0035
	$\beta_2 = 1.5$			$\beta_3 = 1.2$	
		10%	**20%**	**10%**	**20%**
MCAR					
Complete Data	0.0030			0.0027	
Complete Case		0.0045	0.0060	0.0040	0.0041
Subject Mean Imputation		0.0043	0.0077	0.0215	0.0824
Occasion Mean Imputation		0.0999	0.2923	0.0288	0.0876
Conditional Mean Imputation		0.0035	0.0038	0.0029	0.0035
MAR conditional on observed data					
Complete Data	0.0031			0.0025	
Complete Case		0.0048	0.0160	0.0031	0.0037
Subject Mean Imputation		0.0051	0.0147	0.0421	0.1996
Occasion Mean Imputation		0.0247	0.0336	0.0229	0.0674
Conditional Mean Imputation		0.0038	0.0074	0.0035	0.0045
NMAR					
Complete Data	0.0021			0.0023	
Complete Case		0.0039	0.0071	0.0036	0.0053
Subject Mean Imputation		0.0057	0.0138	0.0588	0.3019
Occasion Mean Imputation		0.0112	0.0371	0.0187	0.0478
Conditional Mean Imputation		0.0030	0.0053	0.0048	0.0208

5.1 Comparison of different amount of Missing Data

In this section, the effect of the amount of missing data on missing data techniques is analyzed under each model.

5.1.1 The Log-Log-Gamma MMM

The results obtained under the Log-Log-Gamma MMM are shown in Tables 5.2 and 5.3. MAE and MSE values at 20% missingness are larger than those at 10% missingness across regression parameters, β_1, β_2 and β_3, in most of the 12 conditions (3 missing data mechanisms $\times$ 4 missing data techniques). In others, the difference in the values is too small. This means that, as expected, as the missingness in the data increases, the regression parameters are less accurately and precisely estimated.

Tables 5.2 and 5.3 illustrate interesting findings. While estimating the regression parameters of the time-independent covariates, β_1 and β_2, the missingness in the data influences occasion mean imputation mostly; consequently it displays the worst performance in estimates. It has the largest MAE and MSE values both at 10% and 20% missingness, compared to the MAE and MSE values of other techniques. It is also worthy to mention that when the missingness percentage in the data increases from 10% to 20%, the MAE and MSE values of occasion mean imputation grows at a higher rate compared to other techniques. For example, in Table 5.2, for β_2, under MCAR, there exists approximately 28% increase in the MAE values of occasion mean imputation, when the missingness percentage in the data is changed from 10% to 20%. However, slight increases appear for other techniques, under the same condition.

On the other hand, while estimating the regression parameter of the time-dependent covariate, β_3, firstly subject mean imputation, afterwards, occasion mean imputation are mostly affected by the missingness in the data. As Tables

5.2 and 5.3, for β_3, reveal, subject mean imputation has the largest MAE and MSE values both at 10% and 20% missingness. Changing missingness percentage from 10% to 20% in the data causes subject mean imputation to exhibit a subversive effect both on MAE and MSE values. To illustrate, in Table 5.2, under MAR conditional on observed data, for β_3, subject mean imputation at 20% missingness yield nearly twice as large MAE values as that at 10% missingness. However, complete case analysis, or conditional mean imputation appears to be more robust to missingness percentage changes under the same condition.

5.1.2 The Random-Intercept Model

The results obtained under the random-intercept model are reported in Tables 5.4 and 4.5. Across regression parameters, β_1, β_2 and β_3, MAE and MSE values at 20% missingness are larger than those at 10% missingness, in all of the 12 conditions (3 missing data mechanisms $\times$ 4 missing data techniques).

The random-intercept model yields identical results with the Log-Log-Gamma MMM in the context of missingness percentage in the data. As illustrated in the columns labeled β_1 and β_2 in Tables 5.4 and 5.5, while estimating β_1 and β_2, the missingness percentage in the data affects occasion mean imputation mostly. Accordingly, it produces the largest MAE and MSE values both at 10% and 20% missingness, and displays the highest rate of change in MAE and MSE values, when the missingness percentage in the data changes from 10% to 20%.

Similar to the case in the Log-Log-Gamma MMM, while estimating β_3, subject mean imputation, afterwards, occasion mean imputation are mostly affected by the missingness percentage in the data. Naturally, as in Tables 5.4 and 5.5, for β_3 reveal, subject mean imputation has the largest MAE and MSE values both at 10% and 20% missingness, compared to the MAE and MSE values of other techniques.

5.2 Comparison of Missing Data Mechanisms

This section discusses the comparison of the missing data mechanisms under each model. To aid in the interpretation of results in terms of the missing data mechanisms, we follow the procedure of Newman (2003). That's we combine the MAE values and MSE values across all regression parameters, β_o, β_1, β_2 and β_3 and tabulate the results in Tables 5.6 and 5.7.

5.2.1 The Log-Log-Gamma MMM

For the Log-Log-Gamma MMM, Table 5.6 shows that all missing data techniques exhibit larger average MAE and MSE values under NMAR compared to the values under MAR conditional on observed data and MCAR, both at 10% and 20% missingness. Not surprisingly, this is the expected case in the framework of missing data theory, since parameter estimates which are estimated under MCAR or MAR conditional on observed data are expected to be less biased than those of under NMAR. Moreover, larger average MAE and MSE values are observed under MAR conditional on observed data than under MCAR. One exception is occasion mean imputation which gives slightly lower average MAE and MSE values under MAR conditional on observed data than MCAR at 20% missingness in the data, as seen in the bold digits in Table 5.6.

Table 5.6 Average MAE and MSE values of missing data techniques across all regression parameters including the intercept

	Log-Log-Gamma MMM			
	Average MAE		**Average MSE**	
	10%	**20%**	**10%**	**20%**
MCAR				
Complete Case	0.275	0.296	0.123	0.144
Subject Mean Imputation	0.291	0.303	0.130	0.131
Occasion Mean Imputation	0.240	**0.320**	0.088	**0.153**
Conditional Mean Imputation	0.206	0.288	0.149	0.135

Table 5.6 (Cont'd)

MAR conditional on observed data				
Complete Case	0.301	0.316	0.148	0.159
Subject Mean Imputation	0.302	0.327	0.129	0.151
Occasion Mean Imputation	0.288	**0.294**	0.127	**0.126**
Conditional Mean Imputation	0.308	0.296	0.154	0.137
NMAR				
Complete Case	0.315	0.340	0.160	0.194
Subject Mean Imputation	0.300	0.381	0.126	0.224
Occasion Mean Imputation	0.301	0.363	0.143	0.189
Conditional Mean Imputation	0.309	0.344	0.152	0.180

5.2.2 The Random-Intercept Model

For the random-intercept model, Table 5.7 illustrates that only the use of subject mean yields larger average MAE and MSE values under NMAR compared to the values under MAR conditional on observed data and MCAR. While no significant differences are observed in average MAE and MSE values of complete case analysis and conditional mean imputation under three missing data mechanisms, the average MAE and MSE values of occasion mean imputation are contrary to the expected pattern in the context of missing data theory. When the missingness mechanism changes from MCAR to MAR conditional on observed data, then, to NMAR, there appear a sharp decrease in average MAE and MSE values of occasion mean imputation both at 10% and at 20% missingness in the data, as seen in the bold digits in Table 5.7.

Table 5.7 Average MAE and MSE values of missing data techniques across all regression parameters including the intercept

	Random-Intercept Model			
	Average MAE		**Average MSE**	
	10%	**20%**	**10%**	**20%**
MCAR				
Complete Case	0.051	0.058	0.004	0.005
Subject Mean Imputation	0.073	0.123	0.009	0.027
Occasion Mean Imputation	**0.291**	**0.510**	**0.103**	**0.304**
Conditional Mean Imputation	0.044	0.050	0.003	0.004

Table 5.7 (Cont'd)

MAR conditional on observed data				
Complete Case	0.047	0.063	0.003	0.007
Subject Mean Imputation	0.095	0.178	0.015	0.059
Occasion Mean Imputation	**0.132**	**0.199**	**0.024**	**0.055**
Conditional Mean Imputation	0.049	0.058	0.003	0.005
NMAR				
Complete Case	0.045	0.057	0.003	0.005
Subject Mean Imputation	0.110	0.210	0.020	0.087
Occasion Mean Imputation	**0.089**	**0.152**	**0.012**	**0.031**
Conditional Mean Imputation	0.044	0.062	0.003	0.008

5.3 Comparison of Missing Data Techniques

In this section, the performance of the missing data techniques is discussed generally.

5.3.1 The Log-Log-Gamma MMM

Under the Log-Log-Gamma MMM, occasion mean imputation performs poorly among other missing data techniques, and incompetence of occasion mean imputation is stressed especially when estimating the regression parameters, β_1 and β_2, since it results in the largest average MAE and MSE values (see Tables 5.2 and 5.3). Furthermore, this missing data technique is the most sensitive to the missingness percentage in the data. Under MCAR, for β_3, occasion mean continues to perform poorly. However, when it is MAR conditional on observed data or NMAR, occasion mean imputation is no longer the worst technique for estimating the regression parameter, β_3, and performs better than subject mean imputation.

In general, the failure of occasion mean imputation may be due to the fact that it does not use information related to subject while imputing missing values. Consequently, it may not capture the trend within the values of a subject.

Subject mean imputation performs better than occasion mean imputation, in

estimating the regression parameters, β_1 and β_2. Interestingly, however, for β_3 regression parameter, subject mean imputation displays worse results than occasion mean imputation under MAR conditional on observed data and NMAR. As Tables 5.2 and 5.3 show that, when the missing data mechanism is NMAR, subject mean imputation gives the worst estimate for β_3 compared to other missing data techniques.

General performance of subject mean imputation over occasion mean imputation can be explained as its use of information related to subject while imputing missing values.

Complete case analysis does not deal with imputation task, and uses less information, unlike occasion and subject mean imputation which use all available information. Our simulation study shows that across regression parameters, β_1, β_2 and β_3, and under all conditions, complete case analysis performs at least as good as occasion and subject mean imputation in terms of producing smaller or similar average MAE and MSE values.

Conditional mean imputation outperforms subject and occasion mean imputation in producing accurate and precise parameter estimates for regression parameters, β_1, β_2 and β_3 and gives similar results with complete case analysis. Conditional mean imputation gives the smallest MAE and MSE values, for β_3, under MAR conditional on observed data (0.1593 and 0.0386, respectively: see Tables 5.2 and 5.3).

5.3.2 The Random-Intercept Model

The random-intercept model provides substantially similar results with the Log-Log-Gamma MMM. Like mentioned above for the Log-Log-Gamma MMM, occasion mean imputation also turns out to be the ineffective method in terms of accuracy and precision under the random intercept model. It displays the same

behaviors across the regression parameters, β_1, β_2, and β_3, and under all conditions under the Log-Log-Gamma MMM.

The behaviors of subject mean imputation and complete case analysis under the random-intercept model give parallel results to the cases under the Log-Log-Gamma MMM. Conditional mean imputation turns out be superior to the unconditional mean imputations, for the regression parameters, β_1, β_2 and β_3, regardless of any missing data mechanism, and give similar results with complete case analysis. The only exception is for β_3 under NMAR, for which complete case analysis performs better than conditional mean imputation.

5.4 Comparison of Regression Models

Inspection of the values in row labeled "complete data" in Tables 5.2 - 5.5 points out that the Log-Log-Gamma MMM yields larger MAE and MSE values than the random-intercept model. This shows that the regression parameters under the Log-Log-Gamma MMM are less accurately and precisely estimated compared to those under the random-intercept model. However, this does not directly mean the Log-Log-Gamma MMM should be abandoned in favour of the random-intercept model. The difference between the MAE and MSE values of the Log-Log-Gamma MMM and the random-intercept model can be due to the dissimilarity of estimation methods in SAS NLMIXED and SAS GLIMMIX procedures. An updated release of SAS NLMIXED procedure which allows more than three random effects per subject will accommodate the random-intercept model described in this thesis. Then, it will be possible to make more fair comparisons between these two regression models and to tell which one is the best. Although regression model selection is subject to the question of research interest, it is worthy to remind that one shortcoming of the random-intercept model is that the results of the model are subject to the within-subject covariance structure choice in the multivariate distribution. However, the Log-Log-Gamma MMM is free of this assumption and, besides; it handles the

overdispersion problem wisely. The features of the Log-Log-Gamma MMM and the random-intercept model are discussed in detail in Sections 3.1.1 and 3.1.2.

CHAPTER 6

CONCLUSION

Among regression models dealing with longitudinal count data in the literature, this thesis focuses on the Log-Log-Gamma marginalized multilevel model, which was developed by Griswold and Zeger (2004), and the random-intercept model. Log-Log-Gamma MMM is a likelihood-based model and offers a GLM for the mean response model, and a nonlinear mixed model for the within-subject association model. Separation of the model for mean response from that for within-subject association eases the interpretation of regression parameters of interest. Moreover, the Log-Log-Gamma MMM specifies a gamma distribution for the random effects which is conjugate to the Poisson distribution of conditional mean model. This is a great advantage over normally distributed random effects model since the Poisson-gamma mixture is able to remedy the overdispersion problem.

One of the most frequently encountered problems in longitudinal studies is the missing values in the data due to data collection process over a sequence of time periods. This leads the longitudinal data to be incomplete. To facilitate the work of the regression models against missingness, missing data techniques can be utilized. For instance, the missing values in the data can be either ignored by the use of complete case analysis, or filled in by the imputation methods such as subject, occasion, and conditional mean imputation. However, the missingness percentage in the data and the missingness mechanism that the data have, whether it is MCAR, MAR conditional on observed data, or NMAR, affect the performance of missing data techniques. For that reason, these concepts should be taken into account as well.

To make an effective assessment, for the Log-Log-Gamma MMM, only a random intercept is assumed as a random coefficient in the linear predictor. As a competitor regression model, the random-intercept model from generalized linear mixed models is preferred.

For the simulation study, 100 subjects with four repeated measurements, three covariates, where two are time-independent, and one is time-dependent, are determined. After generating and saving original longitudinal count datasets, under each model, 24 different conditions are created by multiplying three missing data mechanisms, two types of missingness percentage and four 4 missing data techniques. Missingness is limited to 2^{nd}, 3^{rd}, or 4^{th} measurements of the response variable and not allowed in the measurements of the matrix of covariates, and in the 1^{st} measurement of the response variable. Any drop-out or intermittent missing data patterns in the measurements of the response variable are welcomed.

While data generation process is achieved by R version 2.8.1, the statistical evaluation of the models is achieved by SAS version 9.1.3. Due to the lack of computational advances, while SAS NLMIXED procedure is preferred for the Log-Log-Gamma MMM, SAS GLIMMIX procedure is used for the random-intercept model.

Based on the statistical evaluation quantities, mean absolute error and mean square error, the simulation study supports the missing data theory and proves that missingness percentage in the data, and the missingness mechanism that the data have, influence the performance of missing data techniques under both regression model.

Under both regression models, while generally occasion mean imputation displays the worst performance in the estimates, conditional mean imputation shows a superior performance, over occasion and subject mean imputation, regardless of missing data amount and missing data mechanisms, and gives parallel results with

complete case analysis. Although, behaviors are similar under the models, the Log-Log-Gamma MMM yields larger MAE and MSE values than the random-intercept model.

Longitudinal count data lack availability in different statistical software. Model fitting of longitudinal count data in different statistical software with different estimation techniques will enable improvement in usage frequency, inference capabilities and comparison. Furthermore, as Nelson et al. (2006) stresses, non-normal random effects are taking progressive attention not only from longitudinal data analysis field, but also from different areas in statistics, and are more realistic than normally distributed random effects. However, non-normal random effects suffer from the lack of computational implementation as well.

As a future work, missingness in the data can be admitted to the matrix of covariates, in addition to the measurements of response variable. However, although the implementation of incomplete longitudinal continuous data by multiple imputation technique is offered by the pan function under the PAN library in R, or SAS MI procedure, the implementation of incomplete longitudinal count data by multiple imputation technique is still suffering from computational availability in statistical software. Filling in missing values in an incomplete longitudinal count data by multiple imputation technique can be a good extension of this thesis. Sensitivity analysis on model fit under the normality assumption is currently under investigation when random effects come from a nonnormal distribution.

REFERENCES

Abramowitz, M., and Stegun, I. (eds.) (1972). Handbook of Mathematical Functions. New York: Dover.

Barron, D.N. (1992). The Analysis of count data: Overdispersion and Autocorrelation. Sociological Methodology, 22, 179-220.

Bennett, D.A. (2001). How can I deal with missing data in my study? Australian & New Zealand Journal of Public Health, 25, 464–469**.**

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. Journal of American Statistical Association, 88, 9-25.

Cameron, A.C., and Trivedi, P.K. (1998). Regression Analysis of Count Data. Econometric Society Monographs 30. New York: Cambridge University Press.

Carpenter, J. (2005). Missing value mechanism. Retrieved from http://www.lshtm.ac.uk/msu/missingdata/jargon_web/node3.html (18 May 2009).

Daniels, M.J., and Hogan, J.W. (2008). Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. New York: Chapman & Hall.

Diggle, P.J., Heagerty, P.J., Liang, K.Y., and Zeger, S.L. (2002). Analysis of Longitudinal Data. Oxford Science Publications. Oxford: Clarendon Press.

Engels, J.M., and Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology, 56 (10), 968-76.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications. New York: John Wiley.

Fitzmaurice, G., and Molenberghs, G. (2008). Advances in longitudinal data analysis: A historical perspective. In *Longitudinal Data Analysis: A Handbook of Modern*

Statistical Methods. (pp 3-27). Garrett Fitzmaurice, Marie Davidian, Geert Molenberghs and Geert Verbeke. (Co-Eds.). Boca Raton, FL: Chapman & Hall/CRC Press.

Fitzmaurice, G., and Verbeke, G. (2008). Parametric modeling of longitudinal data: Introduction and overview. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. (pp 32-42). Garrett Fitzmaurice, Marie Davidian, Geert Molenberghs and Geert Verbeke. (Co-Eds.). Boca Raton, FL: Chapman & Hall/CRC Press.

Flom, P.L., McMahon, J.M., and Pouget, E.R. (2006). Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with binary outcomes. Northeast SAS Users Group (NESUG) Proceedings. SAS Inc.: Cary, NC.

Fong, D.Y.T., and Lam, K.S.L. (2005). "Use of Multiple Imputation on Linear Mixed Model and Generalized Estimating Equations for Longitudinal Data Analysis." Presented at The XVII IEA World Congress of Epidemiology, held in Bangkok, Thailand, 21-25 August 2005.

Greenwood, M., and Yule, G. U. (1920). An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. Journal of the Royal Statistical Society, 83, 255-279.

Griswold, M.E., and Zeger, S.L. (2004). On Marginalized Multilevel Models and their Computation. The Johns Hopkins University, Department of Biostatistics Working Papers.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of American Statistical Association, 72, 320-340.

Heagerty, P.J., and Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference. Statistical Science, 75 (1), 1-26.

Ilk, O. (2008). Multivariate Longitudinal Data Analysis: Models for Binary Response and Exploratory Tools for Binary and Continuous Response. Saarbrucken: Verlag Dr. Muller (VDM).

Joseph, G. I., and Geert, M. (2009). Missing data methods in longitudinal studies: a review. Test, 18 (1), 1–43.

Jowaheer, V., and Sutradhar, B.C. (2002). Analyzing longitudinal count data with overdispersion. Biometrika, 89 (2), 389-399.

Kaciroti, N.A., Raghunathan, T.E., Schork, M.A., and Clark, N.M. (2008). A Bayesian model for longitudinal count data with non-ignorable dropout**.** Journal of the Royal Statistical Society Series C-Applied Statistics, 57, 521-534.

Kalbfleisch, J.D., and Prentice R.L. (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley and Sons.

Kincaid, C. (2008). Guidelines for Selecting the Covariance Structure in Mixed Model Analysis. Retrieved from www2.sas.com/proceedings/sugi30/198-30.pdf. (18 May 2009).

Leppik, I.E., Dreifuss, F.E., Bowman-Cloyd**,** T., Santilli, N., Jacobs. M., Crosby, C., Cloyd, J., Stockman, J., Graves, N., Sutula, T., Welty, T., Vickery, J., Brundage, R., Gumnit, R., and Gutierres, A. (1985). A double-blind crossover evaluation of progabide in partial seizures. Neurology, 35, 285.

Li, J., Yang, X., Wu, Y., and Shoptaw, S. (2007). A random-effects markov transition model for Poisson-distributed repeated measures with non-ignorable missing values. Statistics in Medicine, 26, 2519-2532.

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

Lipsitz, S., and Fitzmaurice, G. (2008). Generalized estimating equations for longitudinal data analysis. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. (pp 43-78). Garrett Fitzmaurice, Marie Davidian, Geert Molenberghs and Geert Verbeke. (Co-Eds.). Boca Raton, FL: Chapman & Hall/CRC Press.

Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.

Liu, L., and Yu, Z. (2008). A likelihood reformulation method in non-normal random effects models. Statistics in medicine, 27, 3105-3124.

McKnight, P.E., McKnight, K.M., Sidani, .S., and Figueredo, A.J. (2007). Missing Data: A Gentle Introduction. New York: The Guilford Press.

Molenberghs, G., and Verbeke, G. (2005). Models for Discrete Longitudinal Data. New York: Springer.

Molenberghs, G., and Fitzmaurice, G. (2008). Incomplete data: Introduction and overview. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. (pp 395-408). Garrett Fitzmaurice, Marie Davidian, Geert Molenberghs and Geert Verbeke. (Co-Eds.). Boca Raton, FL: Chapman & Hall/CRC Press.

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A, 135, 370-384.

Nelson, K.P., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J., Parzen, M., Strawderman, R. (2006). Use of the Probability Integral Transformation to Fit Nonlinear Mixed-Effects Models With Nonnormal Random Effects. Journal of Computational & Graphical Statistics ,15 (1), 39-57.

Newman, D.A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. Organizational Research Methods, 6, 328-362.

Rabe-Hesketh, S.R., and Skrondal, A. (2008). Generalized linear mixed-effects models. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. (pp 79-106). Garrett Fitzmaurice, Marie Davidian, Geert Molenberghs and Geert Verbeke. (Co-Eds.). Boca Raton, FL: Chapman & Hall/CRC Press.

Roth, P.L. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47(3), 537-560.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592. SAS Institute Inc. (2000).

SAS/STAT User's Guide, Version 8, Chapter 46. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2004). SAS Online Doc 9.1.3. Cary, NC: SAS Institute Inc.

Schafer, J. (2005). Missing Data in Longitudinal Studies: A Review. Retrieved from www.stat.psu.edu/~jls/aaps_schafer.pdf . (18 May 2009).

Scheffer, J. (2002). Dealing with Missing Data. Res. Lett. Inf. Math. Sci., 3, 153-160.

Shieh, Y.Y. (2003). Imputation Methods On General Linear Mixed Models of Longitudinal Studies. Retrieved from www.fcsm.gov/03papers/Shieh.pdf (25 May 2009).

Thall, P. F., and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. Biometrics, 26 (3), 657-671.

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

Weiss, R.E. (2005). Modeling longitudinal data. New York: Springer.

Xu, S., Jones, R.H., and Grunwald, G. (2007). Analysis of Longitudinal Count Data with Serial Correlation. Biometrical Journal, 49(3), 416-428.

Yang, Z., Hardin, J.W., Addy, C.L., and Vuong, Q.H. (2007). Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model. Biometrical Journal, 49 (4), 565-584.

Yucel, R. (2009). "Analysis of (incomplete) longitudinal data." Presented at the Pre-conference course of the 5th Conference Eastern Mediterranean Region of the International Biometric Society, held in Istanbul, Turkey, 10-14 May 2009.

APPENDIX A

R CODES FOR THE GENERATION OF COVARIATES MATRIX

```
#  Generating the elements of matrix of time-independent covariates, Xf.
#  Each row of the matrix, Xf, corresponds to value of j^th visit (j = 1,2,3,4) for the i^th
#  subject (i = 1,2,…,100).


Xf<-matrix(0,100,3)
for(i in 1:dim(Xf)[1])Xf[i,]=cbind(1,runif(1, min=-1, max=1), rbinom(1,1,0.5) )


#  Defining a common AR(1) within-subject covariance matrix, Vx3,  for the time-
#  dependent covariate of all subjects (i = 1,2,…,100).

rho <- 0.9
sigma2 <-1
times<-1:4
H <- abs(outer(times, times, "-"))
Vx3<- sigma2* rho^H

#  Generating the elements of matrix of time-dependent covariates, Xt.
#  Each row of the matrix corresponds to value of j^th visit (j = 1,2,3,4) for the i^th
#  subject (i = 1,2,…,100).
#  Install package mvtnorm package

library(mvtnorm)


Xt<-matrix(0,100,4)
for(i in 1:dim(Xt)[1]) Xt [i,]=rmvnorm(1,mean=c(0,0,0,0),sigma=Vx3)
```

APPENDIX B

R CODES FOR THE GENERATION OF DATA UNDER LOG-LOG-GAMMA MMM

```
#  Generating the elements of matrix of random intercepts, C. Each row of the
#  matrix, C, corresponds to value of j^th visit (j = 1,2,3,4) for the i^th subject
#  (i = 1,2,…,100). Each random intercept follow a gamma distribution with a shape
#  and a scale parameter equal to 1.

C<-matrix(0,100,4)
for(i in 1:dim(C)[1]) {
for(j in 1:dim(C)[2]) {
C[i,j]=rgamma(1, shape=1, rate = 1, scale = 1)
}
}

#  Defining the true parameter for the fixed effects regression coefficients

 beta_m<-c(2,-1.3,1.5, 1.2)

#  Generating the elements of matrix of delta, as stated in the Section 3.1.1, for
#  each  subject (i = 1,2,…,100).

delta<-((cbind((cbind(Xf,Xt[,1])%*%beta_m),(cbind(Xf,Xt[,2])%*%beta_m),
(cbind(Xf,Xt[,3])%*% beta_m), (cbind(Xf,Xt[,4])%*% beta_m)))- log(1))

#  Defining the elements of matrix of mu_c for each subject (i = 1,2,…,100)

mu_c <-cbind((exp(delta[,1]+log(C[,1]))),(exp(delta[,2]+ log(C[,2]))) , (exp(delta[,3] +
log(C[,3]))) , (exp(delta[,4] + log(C[,4]))))

#  Generating the elements of matrix of response, Y. Each row of the response
#  matrix, # Y, corresponds to value of j^th visit (j = 1,2,3,4) in the i^th subject
#  (i = 1,2,…,100).

Y<-matrix(0,100,4)
for(j in 1:dim(Y)[2]){
for(i in 1:dim(Y)[1]){
Y[i,j]= rpois(1, mu_c[i,j])
}
}
```

APPENDIX C

R CODES FOR THE GENERATION OF DATA UNDER RANDOM-INTERCEPT MODEL

```
#  Defining a common AR(1) within-subject covariance matrix, V of random-
#  intercepts for all subjects (i = 1,2,…,100).

rho <- 0.5
sigma2 <-0.1
times<-1:4
H <- abs(outer(times, times, "-"))
V<-(sigma2*rho^H)

#  Generating the elements of matrix of random intercepts, A. Each row of the
#  matrix, # A corresponds to value of j^th visit (j = 1,2,3,4) in the i^th subject
#  (i = 1,2,…,100). The random-intercepts within a subject has zero mean vector, and
#  AR(1)   covariance matrix, V.

A<-matrix(0,100,4)
i=dim(A)[1]
for(i in 1:dim(A)[1])A[i,]=rmvnorm(1,mean=c(0,0,0,0),sigma=V)

#  Defining the true parameter for the fixed effects regression coefficients

 beta<-c(2,-1.3,1.5, 1.2)

#  Defining the elements of matrix of mu_c for each subject (i = 1,2,…,100)

mu_c <-cbind(  (exp((cbind(Xf,Xt[,1]) %*% beta)+ A[,1])), (exp((cbind(Xf,Xt[,2]) %*%
beta)+ A[,2])), (exp((cbind(Xf,Xt[,3]) %*% beta)+ A[,3])), exp((cbind(Xf,Xt[,4]) %*%
beta)+A[,4])) )

#  Generating the elements of matrix of response, Y. Each row of the response
#  matrix, # Y corresponds to value of j^th visit (j = 1,2,3,4) for the i^th subject
#  (i = 1,2,…,100).

Y<-matrix(0,100,4)
for(j in 1:dim(Y)[2]) {
for(i in 1:dim(Y)[1]) {
Y[i,j]= rpois(1, mu_c[i,j])
}
}
```

APPENDIX D

R CODES FOR MISSING DATA GENERATION SCENARIOS

```
#  For MCAR type missing data mechanism
#  generating 10% missingness in the response matrix, Y.

while(sum(is.na(Y))<40)  {
id<-sample(1:100, 1,replace=T)
ocassion<-sample(2:4,1, replace=T)
Y[id, ocassion]<-NA
}
Ymis<-Y
sum(is.na(Ymis))

#  For MAR conditional on observed data type missing data mechanism
#  generating 10% missingness in the response matrix, Y.

a<-data.frame(Y,Xf)
while(sum(is.na(Y))<40)  {
row.sample<-((1:nrow(a))[a$X3.1=="1"])
id<-sample(row.sample, 1,replace=T)
ocassion<-sample(2:4,1, replace=T)
Y[id, ocassion]<-NA
}
Ymis<-Y
sum(is.na(Ymis))

#  For NMAR type missing data mechanism
#  Generating 10% missingness in the response matrix, Y.

while(sum(is.na(Y))<40)  {
row.sample<-((1:nrow(Y))[Y[,1]>20 |Y[,2]>25 |Y[,3]>30 ])
id<-sample(row.sample, 1,replace=T)
ocassion<-sample(2:4,1, replace=T)
Y[id, ocassion]<-NA
}
Ymis<-Y
sum(is.na(Ymis))

#  For 20% missingness in the response matrix, Y, replace 40 with 80 in the R codes
```

APPENDIX E

R CODES FOR MISSING DATA TECHNIQUES

```
#  Complete case

b<-data.frame(Ymis,Xf,Xt)
br<-na.exclude(b)


#  Subject mean imputation

for( i in 1:dim(Ymis)[1]) {
Ymis[i,][is.na(Ymis[i,])]<-round(mean(Ymis[i,],na.rm=T))
}
Ymeanr<-Ymis

#  Occasion mean imputation

for(j in 2:dim(Ymis)[2]) {
Ymis[,j][is.na(Ymis[,j])]<-round(mean(Ymis[,j],na.rm=T))
}
Ymeanc<-Ymis


#  Conditional mean imputation

#  Columns 1,2,3,and 4 refer to response matrix, Y, column 5 refers to vector of
#  ones, column 5, 6 and 7 refer to time-independent covariates, and lastly columns
#  8,9, 10, and 11 refer to time-independent covariates.
#  Coef matrix is composed of coefficients of the fitted equations.

Coef <-rbind(coef (glm( br[,2]~ br[,1]+ br[,6]+ br[,7]+ br[,9], family=poisson,
data=br)),
coef(glm( br[,3]~ br[,2]+ br[,6]+ br[,7]+ br[,10], family=poisson, data=br) ),
coef(glm( br[,4]~ br[,3]+ br[,6]+ br[,7]+ br[,11], family=poisson, data=br) ))

for(j in 2:dim(Ymis)[2] ) {  for(i in 1:dim(Ymis)[1] )
Ymis[i,j][is.na(Ymis[i,j])]<-round(exp(Coef[j-1,1]+(Coef[j-1,2]*Y[i,j-1])+(Coef[j-1,3]
*Xf[i,][2])+(Coef[j-1,4] *Xf[i,][3]) + (Coef[j-1,5] *Xt[i,j] )))
}
Ycon<-Ymis
```

APPENDIX F

R CODES FOR LONGITUDINAL DATA FORMAT

```
#  In any data frame, the id column refers to the subject id number, occasion
#  column refers to occasion number for that subject. The column Ynew refers to
#  the response for the j^th of 4 occasions for the i^th subject (i = 1,2,…,100).
#  The first column of the matrix Xnew corresponds to the time-independent
#  covariate, X_1, the second column refers to the time-independent covariate
#  X_2, and lastly, the third column refers to the time-dependent covariate X_3.

#  Longitudinal data format for the complete dataset

Ynew<-as.vector(t(Y))
id <- c(rep(1:100,each=4))
occasion <-c(rep(1:4,100))
Xtt<- as.vector(t(Xt))
Xnew<-cbind( rep(Xf[,2], each=4), rep(Xf[,3], each=4),Xtt )
data1<-data.frame(id,occasion,Ynew, Xnew)


#  Longitudinal data format for complete dataset after complete case analysis

Ynew<- br[,c(1,2,3,4)]
d<- nrow(Ynew)
Ynew<-as.vector(t(Ynew))
id <- c(rep(1:d,each=4))
occasion <-c(rep(1:4, d))
Xfr<- br[,c(5,6,7)]
Xtr<- br[,c(8,9,10,11)]
Xtr<- as.vector(t(Xtr))
Xnew<-cbind( rep(Xfr[,2], each=4), rep(Xfr[,3], each=4),Xtr )
data2<-data.frame(id, occasion,Ynew, Xnew)


#   Longitudinal data format for the complete dataset after subject mean imputation

Ynew<-as.vector(t(Ymeanr))
id <- c(rep(1:100,each=4))
occasion <-c(rep(1:4,100))
Xtt<- as.vector(t(Xt))
Xnew<-cbind( rep(Xf[,2], each=4),              rep(Xf[,3], each=4),Xtt )
```

```
data3<-data.frame(id, occasion,Ynew, Xnew)

#  Longitudinal data format for the complete dataset after occasion mean
#  imputation

Ynew<-as.vector(t(Ymeanc))
id <- c(rep(1:100,each=4))
ocassion <-c(rep(1:4,100))
Xtt<- as.vector(t(Xt))
Xnew<-cbind( rep(Xf[,2], each=4), rep(Xf[,3], each=4),Xtt )
data4<-data.frame(id, ocassion,Ynew, Xnew)


#  Longitudinal data format for the complete dataset after conditional mean
#  imputation

Ynew<-as.vector(t(Ycon))
id <- c(rep(1:100,each=4))
occasion <-c(rep(1:4,100))
Xtt<- as.vector(t(Xt))
Xnew<-cbind( rep(Xf[,2], each=4), rep(Xf[,3], each=4),Xtt)
data5<-data.frame(id, occasion,Ynew, Xnew)
```

APPENDIX G

SAS CODES FOR LOG-LOG-GAMMA MMM

(Adapted from Griswold and Zeger (2004) and Nelson et al. (2006))

```
data data1;
infile "C:\Documents and Settings\GulINAN\Desktop\data1.txt" DELIMITER='09'x;
input no id occasion y x1 x2 xt;
run;

proc sort data=data1;
by id;
run;

proc nlmixed data=data1 noad fd qpoints=30;
PARMS theta1=1 beta0_m=2  beta1_m=-1.3 beta2_m=1.5 beta3_m=1.2;
eta_m= beta0_m + beta1_m*x1+ beta2_m*x2+ beta3_m*xt ;
mu_m=exp(eta_m);
ui= CDF('Normal',ai);
if (ui > 0.9999 ) then ui=0.9999;
bi2=quantile('GAMMA',ui, theta1);
bi1=theta1*bi2;
v=theta1*theta1;
delta=eta_m-log(v);
eta_c = delta + log(bi1);
mu_c=exp(eta_c);
Model y ~ Poisson(mu_c);
Random ai ~ Normal(0,1) subject=id;
run;

/*   noad = refers to nonadaptive Gaussian quadrature */
/*   fd   = specifies that all derivatives be computed using finite difference
approximations. FD is equivalent to FD=100 */
/*    qpoints = refers to the number of quadrature points to be used during
evaluation of integrals */
/*   subject = refers to subjects in the model */
/*   eta_m = specifies the linear predictor of the mean model */
/*   eta_c = specifies the linear predictor of the association model */
/*   mu_c = relates the linear predictor to the association model through the link
function */
```

APPENDIX H

SAS CODES FOR RANDOM-INTERCEPT MODEL

```
data data1;
infile "C:\Documents and Settings\GulINAN\Desktop\data1.txt" DELIMITER='09'x;
input no id occasion y x1 x2 xt;
run;

proc glimmix data=data1 MAXOPT=500;
class id ;
model y = x1 x2 xt / dist=p link=log s;
random intercept / subject=id type=ar(1);
run;


/*  dist = refers to conditional distribution of the data  */
/*  link = refers to the link function in the model  */
/*  random intercept = specifies a random intercept in the model  */
/*  subject = refers to subjects in the model  */
/*  type = refers to within-subject covariance structure in the model  */
```