5-8-2017

A Joint Model of Longitudinal Data and Informative Time with Time-Dependent Covariate

Mohammad Abdullatif Alomair

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

A JOINT MODEL OF LONGITUDINAL DATA AND INFORMATIVE TIME WITH TIME-DEPENDENT COVARIATE

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Mohammad Abdullatif Alomair

College of Education and Behavioral Sciences
Department of Applied Statistics and Research Methods
Applied Statistics and Research Methods

May 2017

This dissertation by: Mohammad Abdullatif Alomair

Entitled: *A JOINT MODEL OF LONGITUDINAL DATA AND INFORMATIVE TIME WITH TIME-DEPENDENT COVARIATE*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences in Department of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

Khalil Shafie Holighi, Ph.D., Research Advisor

Trent Lalonde Ph.D., Co-Research Advisor

Jay R. Schaffer, Ph.D., Committee Member

Heng-Yu Ku, Ph.D., Faculty Representative

Date of Dissertation Defense

Accepted by the Graduate School

Linda L. Black, Ed.D.
Associate Provost and Dean
Graduate School and International Admissions

ABSTRACT

Alomair, Mohammad Abdullatif. *A JOINT MODEL OF LONGITUDINAL DATA AND INFORMATIVE TIME WITH TIME-DEPENDENT COVARIATE.* Published Doctor of Philosophy dissertation, University of Northern Colorado, 2017.

In analysis of longitudinal data, a number of methods have been proposed. Most of the traditional longitudinal methods assume that the independent variables are not dependent on time and are the same across study. However, one of the main advantages of longitudinal studies is the ability to observe outcomes and covariates at the same time, and a researcher can define whether changes in a covariate lead to changes in the outcome of interest. In addition, the methods focused on a predetermined observation time that does not carry information about the response variable. Moreover, it is possible in real research to have time-varying covariates, unbalanced observation time, and the observation times may be informative. The usual longitudinal statistical analysis might be biased if their assumptions are not valid.

The purpose of this study was to develop a joint model of a longitudinal outcome and informative time with time-dependent covariates. In this study, a joint model and analysis of longitudinal data with possibly informative observation times and time-dependent covariates via joint probability distributions has been proposed. The maximum likelihood parameter estimates of the proposed model were obtained

from Monte Carlo simulated data by employing a nonlinear optimization in R. Furthermore, the model selection criteria and likelihood ratio test statistic were computed to select the best fitting model and for comparing nested models. Additionally, the R codes were developed for the proposed model and an application is presented on the bladder cancer data used for explanation purposes. In the application, the results show that the time-dependent covariate appear to be important predictor in the longitudinal data.

ACKNOWLEDGEMENTS

Praise belongs to Allăh, the One who has blessed me in my entire affair. Without His help, nothing can be accomplished. Next, I would like to express my deep gratitude to my supervisor Dr. Khalil Shafie for his guidance and statistical expertise has made this work possible. I especially appreciate the quick and generous responses of him. I am also grateful to the rest of my supervisory committee, Dr. Trent Lalonde, Dr. Jay Schaffe, and Dr. Heng-Yu Ku for their guidance, direction, helpful suggestions, and inspiration. I would like also to express my grateful thanks for all those who have helped or encouraged me, in any way, during my years of graduate study. My deepest gratitude is to all my family members. I thank my father, Abdullatif, my mother, Latifah and all my brothers and sister for their continuous praying and sincere wishes for my success. I would like to thank my two wonderful sons, Ammar and Abdullatif, and my two beautiful daughters, Joud and Reem, for filling my life with joy and happiness. There are no proper words to convey my heartfelt gratitude and respect for my wife, Dr. Gadir Alomair for her continuous support and great patience throughout our exhausting higher education life. Finally, I would like to thank my government for giving me an opportunity to receive a scholarshp and complete my Ph.D. study. Also, I would like to thank the Saudi Arabia Culture Mission (SACM) for supporting and helping me during my study in the United States.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Longitudinal studies are repeated measurements or collected information of individuals or a group of subjects over an interval of time where the measurements are not independent. For example, Framingham Heart Study is one of the most successful longitudinal studies; data of this study were used to produce more than 2,000 articles. In this study, 5,209 adult residents of Framingham between 30 and 62 years of age were observed every two years and information about their height, weight, blood pressure, smoking behavior, and so on was gathered (Long and Fox, 2016; Dawber, 1980). Data become valuable when exploring causal relationships that take a long time to detect. There are two ways of collecting longitudinal data: one way would be by following subjects forward in time, i.e., prospectively; another way would be by deriving several measurements on each subject from past records, i.e., retrospectively. In cross-sectional studies, a single outcome comes from each individual in contrast to longitudinal studies. The prime advantage of a longitudinal study is it investigates changes over time within individuals from differences among subjects and factors that influence the changes. Cross-sectional studies cannot detect the change over time, which makes a longitudinal study more powerful than the cross-sectional design. Moreover, fewer subjects in a longitudinal study provide a similar level of statistical power and more dependent information compared to a

cross-sectional study (Diggle, 2002; Fitzmaurice, Laird, and Ware, 2004). Despite the fact that longitudinal studies take a long time, are expensive, and are hard to analyze, they have become very popular because it is believed the problem of causality can be solved (Twisk, 2003).

The main goal of a longitudinal study is to describe a change in response over time. In addition, it can observe other factors of interest that influence the change. Assessing a within-individual change over time can only be obtained by a longitudinal study design. A cross-sectional study can be used to compare between groups that differ, i.e., in age, but cannot give information about how individuals change over a period of time. For example, believing body fat in girls increases before or around menarche and stays level for four years can be tested by two methods. Researchers would be interested in defining the increase in body fat in girls after menarche. Using a cross-sectional method could be done by taking measurements of body fat from two groups of girls at pre-menarche and at post-menarche. This method could make a comparison between the two groups but could not give an estimate of the change in body fat as the girls' age increased. On the other hand, a longitudinal study could measure a group of girls at two points in time (pre-menarche and post-menarche) and consider other factors that might affect body fat, which gives a valid estimate of change in body fat as girls' age increases (Fitzmaurice et al., 2004).

Since in longitudinal data the outcomes from one individual are correlated, they need particular statistical methods. Choosing the appropriate statistical method depends on the type of outcome variable and the covariates. In longitudinal

data, we can have continuous, binary, or count variables among others. A longitudinal study has the ability to observe an outcome and predictors at the same time, which can help define if there were changes in predictors before changes in the outcome and if those changes affected the outcome. The treatment of time-dependent covariates with longitudinal responses gives strong statistical inferences about the active relationships; however, that makes the statistical model more complicated. In most cases, longitudinal analyses are based on a regression model. Regression models could refer to a regression equation form that explains the dependence of a response mean on a set of covariates (Diggle, 2002; Fitzmaurice, Laird, and Ware, 2004; Hedeker and Gibbons, 2006).

A number of methods have been developed to accommodate different responses and independent variable types. These methods can be simple or complex depending on the study design or on the research objective. Each method works under some assumptions and for different types of situations, which means no one method can take care of all. Researchers need to choose the appropriate method based on several aspects, e.g., study design and response type. For example, the analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) methods are commonly used in the longitudinal study field under some unrealistic assumptions such as a fixed time, balanced data, and equal correlation, which is hard when following subjects, especially if the subjects are human. Other methods that can handle what ANOVA and MANOVA cannot such as unbalanced data and non-constant variances are called mixed-effects models and generalized estimating equations (GEE). However, all the approaches mentioned above would not be

applicable when the further time points for collecting observations are determined based on current outcomes for each subject, which is called informative time.

Statement of the Problem

Several types of longitudinal data appear within clinical studies, economic studies, and others. Each study has challenges and investigations. In longitudinal studies, there are some cases where the subjects either show up earlier than the supposed measurement time or drop out of the study. These cases cause irregular time, unbalanced data, and non-ignorable dropouts. All these situations indicated that there is important information that should not be ignored. Therefore, informative time needs to be of concern in these type of studies and modelled together with the longitudinal variable so it gives the right inference. In addition, time-dependent covariates appear in some studies with informative time, which need a special model to handle both of them. Since existing methods have a common assumption that the times of observations are fixed or are non-informative about the point of interest, there is a need to find a new method. In several longitudinal studies, the response variable can be correlated with observation times, e.g., some patients visit doctors more than other patients when they have a severe disease and this visit can include two different measurements–one for the output and the other for the predictor as a time-dependent covariate. Applying methods relying on a fixed time assumption on data that contain informative time or irregular time intervals can result in biased results.

Some methods have been proposed when longitudinal data and time are related. A joint model is the most useful approach; it models the longitudinal

response and the related time together. For example, Bronsert (2009) considered a joint model for normally distributed longitudinal responses and exponentially distributed times when combined. Lin (2011) and Seo (2015) extended Bronsert's study and showed the joint model could be an alternative method for analyzing longitudinal data from a normal or exponential response. However, these studies modeled the informative time with longitudinal responses for different disruptions with time-independent covariates but could not handle time-dependent covariates. In addition, Chen, May, Ibrahim, Chu, and Cole (2014) proposed a joint model for the analysis of longitudinal and survival data that accounts for both intermittently missing and left-censored time-varying covariates. They consider the left-censored time-varying covariates, which is the measurement is only accurate down to a particular limit of detection. The purpose of this study was to develop a joint model that incorporates informative time and time-dependent covariates with a longitudinal response.

Definition of Terminology

The following technical terms were used throughout this study:

Informative Schedule Data – A set of repeated observations on each subject over a given time period. The observations are measured based on the informative time for each subject determined by previous observations.

Informative Time – The time period between each measurement for each individual. The next measurement is determined by the current observation, Thus, the time interval across the study might vary for all subject.

Longitudinal Data – A set of outcomes or observations of a response that is measured repeatedly over a given period of time. The time period and points are usually predetermined by researchers before collecting the observations and limit the results on the time period used.

Time-Dependent Covariate or Time-Varying Covariate – A predictor that varies both between and within subjects, or the covariates may change their values over time.

Research Statements

This study investigated the following research questions:

Q1 How can the joint model constructed by Lin (2011) for a longitudinal response variable with a set of informative time be extended to a joint model with time-dependent covariate?

Q2 How can the maximum likelihood estimators of the proposed joint model be obtained using R?

Q3 How can the likelihood ratio test be constructed to compare the fit of two models?

Q4 How can the model selection criteria, such as AIC, AICc, or BIC, be utilized to compare different models?

Limitations

There were some limitations in this study researchers will need to consider:

1. This study was limited to outcomes from a normal distribution with a single response variable. Thus, the model should not be applied to studies where outcomes are not normally distributed and/or contain multivariate responses.

2. This study assumed the time came from an exponential distribution. This assumption should be considered before applying the results to other studies.

3. This study assumed the time-dependent covariate was an exogenous. Therefore an endogenous time-dependent covariate is not applicable on this study model.

4. Furthermore, the present study made the assumption the time-dependent covariate was normally distributed. Therefore, this model should not be applied to any study without considering this limitation.

Conclusion

In traditional analysis of longitudinal data, it is often assumed the times factor is fixed or predetermined and is the same for all subjects across the study. However, there are certain cases where the time factor can be informative, i.e., the next observation is determined by the previous outcome of the response variable. For these cases, a biased result would appear if traditional methods were used. Moreover, there are several situations where a predictor measurement can change in the given time duration as response measurement changes. Bronsert (2009) developed a joint model that could link normally distributed longitudinal responses and informative time. Lin (2011) extended the joint model to ensure the multivariate normality could be obtained from the maximum likelihood estimators and proposed model selection criteria. After that, Seo (2015) extended their joint model to the exponential family of distributions. Consequently, this study extended

their joint model to handle both the informative time and time-dependent covariates with a longitudinal response variable.

CHAPTER II

REVIEW OF LITERATURE

Traditional Methods for Longitudinal Data

The simplest form of longitudinal study is where two continuous outcome variables are measured at two different times. With this form, the primary purpose is to test the change in the outcome variable between the two times. In this situation, a paired t-test can be used. The paired t-test is used to test whether the mean difference between first and second measurements equals to some value. However, this method does not work when there are more than two measurement times, which is generally the case in longitudinal studies.

In a situation with more than two repeated measurements, the univariate repeated measures ANOVA model and MANOVA can be used. "The univariate repeated measures ANOVA model provides a natural generalization of Student (1908) paired t-test to handle more than two repeated measurements, in addition to various between-subject factors"(Fitzmaurice, 2008). To utilize any statistical techniques, several assumptions need to be checked. In ANOVA, the outcome has to be normally distributed with the assumptions of independence and homogeneity of variance; the sphericity assumption must also be met, i.e., the correlations between repeated measurement outcomes are equal and the variances of the outcome variable must be the same at each point in time. Sphericity can be known as the

variation of the difference between two responses within a subject is constant. Mauchly's test can be used to verify this assumption. Even with the restrictive assumptions, the univariate repeated measures ANOVA model can be considered a pioneer of more multilateral regression models for longitudinal data (Fitzmaurice, 2008). From a historical perspective, an undoubted appeal ANOVA was one of the few models that could realistically be fit to longitudinal data when software programming was not currently available or not as powerful. The univariate repeated-measures ANOVA model can be written as

$$\boldsymbol{Y}_{ij} = \boldsymbol{X}_{ij}^{'}\boldsymbol{\beta} + b_i + \boldsymbol{\epsilon}_{ij}, \tag{2.1}$$

where $\boldsymbol{Y}_{ij}$ is the outcome of interest,$\boldsymbol{X}_{ij}$ is a design vector, $\boldsymbol{\beta}$ is a vector of regression parameters,b_i is the random effect (Fitzmaurice, 2008).

Another method that has been used with more than two repeated measurements is multivariate repeated measures MANOVA. In MANOVA, two or more dependent variables and several differences are analyzed together. There are some assumptions for MANOVA for repeated measurements: (a) outcomes of different subjects are independent, and (b) the outcomes need to be multivariate normally distributed. In MANOVA there is a vector of responses from the ith subject at time j

$$\boldsymbol{Y}_i = (y_{i1}, y_{i2}, ..., y_{ini})', \tag{2.2}$$

where $i = 1, ..., m$ and $j = 1, ..., n_i$ and $\boldsymbol{Y}_i$ is from $N_t(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The one-sample MANOVA model is given by

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon_i}, \tag{2.3}$$

where $\boldsymbol{\mu} = n \times 1$ mean vector for timepoints, and $\boldsymbol{\epsilon_i} = n \times 1$ vector of errors, distributed as $N(\mathbf{0}, \boldsymbol{\Sigma})$ in the population.(Hedeker and Gibbons, 2006)

The advantage of MANOVA is it does not have the sphericity or compound symmetry assumptions. However, if sphericity holds, ANOVA is more powerful than MANOVA because the denominator degrees of freedom are greater for the univariate F-test of time. The sphericity assumption increases degrees of freedom, which increases the power of ANOVA (Hedeker and Gibbons, 2006).

Both repeated-measures ANOVA and MANOVA assume time intervals are equally spaced and response is normally distributed. They are affected by outliers and missing data because they are based on least-squares estimation. Due to these restrictions and unrealistic assumptions, these models have limited use for complex research situations. The previous methods investigate changes in one continuous variable over time between different groups. These approaches are not appropriate for analyzing the relationship between several predictors and the continuous outcome variable. These methods cannot test the relationship between the longitudinal response and time-dependent covariates. Therefore, these approaches were not appropriate to be used in this study with longitudinal response data, time-dependent covariates, and informative time.

Linear Mixed-Effects Models for Longitudinal Data

In practice, longitudinal studies occur with missing data, unbalanced data, time-dependent covariates, and when the aims of the study are to make inferences about the regression parameters. An alternative to traditional methods is based on linear mixed-effects models as traditional methods for longitudinal data could be difficult or impossible to apply (Fitzmaurice, 2008). A model with both fixed effects (parameters related to the entire population) and random effects (parameters related to a random individual from a population) is called a mixed-effects model (Pinheiro and Bates, 2000).

Harville (1977) introduced a general class of mixed models. Then Laird and Ware (1982) proposed linear mixed-effects models for longitudinal studies. These models addressed unbalanced and missing data in normal ways (Fitzmaurice, 2008). The mixed-effects models were treated as a univariate regression with correlated errors(Davis, 2002). Laird and Ware (1982)came up with a linear mixed-effects model that included two parts–the univariate repeated-measures ANOVA and growth curve models for longitudinal data. Their model had two advantages: first, the design matrices for the fixed and random effects had few restrictions; second, the parameters could be estimated by likelihood methods. In the past, the difficulty of estimation of mixed-effects models made this method less used. Laird and Ware presented a way to fit the general class of models for longitudinal data by using an expection-maximizetion (EM) algorithm. Later, Jennrich and Schluchter (1986)

suggested other algorithms such as Fisher scoring and Newton-Raplson (Fitzmaurice, 2008).

There are several advantages that make mixed-effects models useful in longitudinal studies, e.g., the possibility to include subjects with incomplete data across time to an analysis that increases statistical power. While average change across time can be estimated in traditional methods, change over time for each subject can be estimated in mixed-effects models (Fitzmaurice et al., 2004; Hedeker and Gibbons, 2006). In addition, these models provide a flexible and powerful tool for the analysis of grouped data such as longitudinal data, repeated measures, and multilevel data (Pinheiro and Bates, 2000). Moreover, both time-independent and time-dependent covariates can be included in the model. They can also be adjusted to handle the difficulties of longitudinal data sets and allow specification of models. In mixed-effects, the between and within subjects effects can be modeled and analyzed (Davis, 2002). This method is one of the most used for analyzing longitudinal data, especially with available software.

These models have been described using a variety of names, e.g., random coefficient model (De Leeuw and Kreft, 1986), multilevel models (Goldstein, 2003; Nash, Varadhan, et al., 2011), random effects models (Diggle, 2002; Fitzmaurice, Laird, and Ware, 2004; Laird and Ware, 1982), mixed models (Longford, 1987), and hierarchical models (Lee and Nelder, 1996; Raudenbush and Bryk, 2002).

The general liner mixed-effects model is given by:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \tag{2.4}$$

where:

$\boldsymbol{y}_i$ is an $n_i \times 1$ dependent variable vector for individual i,

$\boldsymbol{x}_i$ is an $n_i \times p$ covariate matrix for individual i,

$\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed regression parameters,

$\boldsymbol{z}_i$ is a given $n_i \times r$ matrix for the random effects,

$\boldsymbol{\gamma}_i$ is an $r \times 1$ vector of random individual effects,

$\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ vector of errors and assumed to be independent of $\boldsymbol{\gamma}_i$, and with the assumptions of $\boldsymbol{\gamma_i} \sim N(\mathbf{0}, \boldsymbol{G}_i)$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{R}_i)$.

In the linear mixed model, the random components are the vectors $\boldsymbol{\gamma}_i$ and $\boldsymbol{\epsilon}_i$. These two components have the following assumptions: $E(\boldsymbol{\gamma}_i) = \mathbf{0}$, $\text{Var}(\boldsymbol{\gamma}_i) = \boldsymbol{G}, E(\boldsymbol{\epsilon}_i) = \mathbf{0_n}$, $\text{Var}(\boldsymbol{\epsilon}_i) = \boldsymbol{R}_i$. Also in this model the mean and the variance of $\boldsymbol{y_i}$ are $E(\boldsymbol{y_i}) = \boldsymbol{x\beta}$ and $var(\boldsymbol{y}_i) = \boldsymbol{z_i G z_i'} + \boldsymbol{R_i}$. When $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{ni}$ and $\boldsymbol{z} = \mathbf{0}$, the mixed model reduces to the standard linear model (Davis, 2002; Fitzmaurice, 2008; Jennrich and Schluchter, 1986; Laird and Ware, 1982).

To obtain parameter estimation for the random effects and fixed effects, maximum likelihood (ML) estimation can be used by considering the numerical solution of a constrained nonlinear optimization problem (Davis, 2002). Due to computational difficulties and bias for unbalanced designs in the maximum likelihood (ML) estimation, Patterson and Thompson (1971) suggested the

alternative restricted maximum likelihood (REML) way (Davis, 2002; Patterson and Thompson, 1971).

When a covariate is measured with an individual measurement change over time, that is called a time-dependent covariate. The time-dependent covariate can be added to the simple liner mixed-effects model as follows:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij} + u_{0i} + u_{1i} t_{ij} + \epsilon_{ij}. \tag{2.5}$$

That can also be expressed by two levels as follows:

Level 1 (within):

$$Y_{ij} = b_{0i} + b_{1i} t_{ij} + b_{2i} x_{ij} + \epsilon_{ij}. \tag{2.6}$$

Level 2(between):

$$b_{0i} = \beta_0 + u_{0i}. \tag{2.7}$$

$$b_{1i} = \beta_1 + u_{1i}. \tag{2.8}$$

$$b_{2i} = \beta_2. \tag{2.9}$$

Where:

x_{ij} is the time-varying covariate for response i at time j

b_{0i} is an outcome level for response i under the average of x_{ij} when t_{ij} and x_{ij} are equal to 0

b_{1i} is an outcome change for response i over time.

b_{2i} is a response's change in outcome due to x_{ij}

β_0 is average for responses with average of x_{ij}

β_1 is the average outcome over time change

β_2 is the average outcome difference for a unit change in x_{ij}

u_{0i} is the individual intercept deviation

u_{1i} is the individual time slope deviation

The models above assumed by adding a time-dependent covariate in a mixed-effects model, the between- and within-subjects effects were equal. To separate the within- and between-subjects effects of time-dependent covariates, one can include both the subject's average $\bar{x}_i$ and the subject's time-varying deviation $x_{ij} - \bar{x}_i$ (Neuhaus, 1998; Hedeker and Gibbons, 2006).

The models become

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_W (x_{ij} - \bar{x}_i) + \beta_B \bar{x}_i + u_{0i} + u_{1i} t_{ij} + \epsilon_{ij}. \tag{2.10}$$

Level 1 (within):

$$Y_{ij} = b_{0i} + b_{1i} t_{ij} + b_W (x_{ij} - \bar{x}_i) + \epsilon_{ij}. \tag{2.11}$$

Level 2 (between):

$$b_{0i} = \beta_0 + \beta_B \bar{x}_i + u_{0i}. \tag{2.12}$$

$$b_{1i} = \beta_1 + u_{1i}. \tag{2.13}$$

$$b_{Wi} = \beta_W. \tag{2.14}$$

In a mixed-effects model with time-dependent covariates, the model requires full data for the time-dependent covariates with the response data. However, it is most common in a longitudinal study to have missing data in time-dependent covariates. When modeling time-dependent covariates with missing data, the method starts with a single covariate by modeling the covariate process over time; for two or more models, each is processed separately but that might cause some loss of efficiency if the covariates are highly correlated (Wu, 2010). Linear mixed-effects models are not very appropriate to use when the longitudinal response is discrete because there is a dependence of the variance on the mean and the average response range is limited (Fitzmaurice, 2008). Because the linear mixed-effects model treats the time as fixed and is not an appropriate method with missing data that not missing completely at random (MCAR) of time-dependent covariates, this method was not a proper approach for this study.

Generalized Estimating Equations for Longitudinal Data

The generalized estimating equations (GEE) approach is very useful when analyzing longitudinal data, particularly when the response is discrete such as count or binary (Fitzmaurice, 2008). The marginal model method to analyze repeated measurements using the GEE method was first developed by K.-Y. Liang and Zeger (1986). This term indicates a model in which the mean response depends only on the covariates of interest and not on the previous output or random effect (Fitzmaurice, Laird, and Ware, 2004). This approach extends the generalized linear models to longitudinal data by accounting for the within-subject correlation among the measurements. Marginal models are a regression model for the response mean

with a function that links the marginal mean response to the covariates at each event and aims to make inferences about population means (Fitzmaurice, 2008). The link and variance functions for normal outcomes are shown below:

$$g(\mu_{ij}) = \mu_{ij} = x_{ij}^{'}\beta,\ v(\mu_{ij}) = 1, \text{and } var(y_{ij}) = v(\mu_{ij})\phi = \phi. \tag{2.15}$$

Marginal models do not require full distributional assumptions for the response, which gives it an advantage to use with count or binary or continuous outcomes. However, they require mean and variance (Fitzmaurice, 2008; Fitzmaurice et al., 2004; K.-Y. Liang and Zeger, 1986). The marginal models for longitudinal data contain three parts:

1. The expectation of each response is conditionally dependent on the covariates, $E(y_{ij}|x_{ij}) = \mu_{ij}$ by the link function $g(\mu_{ij}) = x_{ij}^{'}\ \beta$, Where y_{ij} is the response for subject i at time j, $x_{ij}^{'}$ is $p \times 1$ vector of covariates, β is $p \times 1$ vector of unknown parameters. $g(.)$ is the link function, and μ_{ij} is the mean response.

2. The variance of each response given the covariates is

$$var(y_{ij}) = \phi v(\mu_{ij}). \tag{2.16}$$

where $v(.)$ is a known variance function, which represents the association between the variance and the mean, and ϕ is a scale parameter that could be known or need to be estimated.

3. The association of the within-subject through the responses, given the covariates, is a function of $\boldsymbol{\alpha}$ that depends on the means. These parts extend the marginal model to the marginal models for longitudinal data (Fitzmaurice, 2008; Fitzmaurice et al., 2004; Davis, 2002).

Several advantages make the GEE method useful in the analysis of longitudinal data. For instance, GEE models are not based on distributional assumptions that make it flexible. In contrast to likelihood methods that require full distributional assumptions for the data, GEE estimates are consistent, asymptotically normal, and can deal with continuous and categorical covariates (Davis, 2002; Wu, 2010). Generalized Estimating Equations models have a few limitations, e.g., the estimates are not completely efficient compared to the maximum likelihood, do not allow for individual-specific inference, and the inference is hard when the missing data is missing completely at random (MCAR) (Hedeker and Gibbons, 2006; Wu, 2010). Generalized estimating equations does not require a fixed time interval for the measurements; however, when the time is varied for the measurements, this approach is not appropriate (Hedeker and Gibbons, 2006). The estimated parameter $\hat{\beta}$ is not a consistent estimator of β, when there are stochastic time-dependent covariates with GEE (Sullivan Pepe and Anderson, 1994; Fitzmaurice, Molenberghs, and Lipsitz, 1995; Davis, 2002). Generalized estimating equations and the mixed-effects model work well for analyzing longitudinal data since both methods make the correction about dependent observations in a within subject; the mixed-effect model allows regression coefficients to vary between subjects and by correlation structure in a GEE analysis. However, there is no clear

answer as to which one of them is better or more appropriate; in some situations, they are equally appropriate such as when an exchangeable correlation structure is appropriate and one of the estimated regression coefficients has no random variation(Twisk, 2003). Since this study assumed the time point was not fixed and included time-dependent covariates, the GEE models method was not an appropriate approach for this study.

Joint Model for Longitudinal Data

The joint model is a statistical method for analyzing longitudinal data where each subject provides two kinds of data: (a) repeated measurements from several occasions and (b) a process of occasions in time or other factors with fixed or random time (Diggle, 2002; Fitzmaurice, 2008; Y. Liang et al., 2009) and (Henderson, Diggle, and Dobson, 2000), i.e., comparing different drug treatments for chronic schizophrenia. Following up with patients once a week for eight weeks after randomization, each patient's outcome was a measure of the current severity of their symptoms at each follow-up time. Through the follow up, many patients drop out for various reasons. The interest of this study was the effect of treatments and not the time of dropout; however, because time is informative, it needs to be considered in the analysis (Fitzmaurice, 2008). Another example would be when patients in the late stage of a disease see doctors more times than those in early stages; ignoring the time informative leads to biased results (Song, Mu, and Sun, 2012). Previous examples explained cases for informative time with longitudinal outcomes where it is important to model both of them together in order to make a valid inference. The

inference is based on modeling the time distribution conditionally on the outcome measurements.

A few methods have been devolved for cases where the informative time and longitudinal outcomes were related (Bronsert, 2009; Huang, Wang, and Zhang, 2006; Y. Liang, Lu, and Ying, 2009; Lin, 2011; Lipsitz, Fitzmaurice, Ibrahim, Gelber, and Lipshultz, 2002; Ryu, Sinha, Mallick, Lipsitz, and Lipshultz, 2007; Seo, 2015; J. Sun, Park, Sun, and Zhao, 2005). For example, J. Sun et al. (2007) presented a joint model for the longitudinal and observation processes by a shared latent variable. These estimations and methods could not handle time-dependent covariates (Song, Mu, and Sun, 2012).

Bronsert (2009) proposed a model that jointly models an informative time component and a longitudinal process. The joint model in this study assumed the current outcome was dependent on the most recent outcome and current time point. In a Gaussian-exponential model, the outcome variable is normally distributed and the informative time is exponentially distributed. The pdf form Bronsert followed was:

$$f_\Theta(\boldsymbol{y_i}, \boldsymbol{t_i}) = f_\Theta(y_{i1}|t_{i1}) f_\Theta(t_{i1}) \prod_{j=2}^{ni} f_\Theta(y_{ij}|y_{i(j-1)}, t_{ij}) f_\Theta(t_{ij}|y_{i(j-1)}), \tag{2.17}$$

where $\boldsymbol{\Theta}$ is a vector of unknown parameters, $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})$ is a vector $n_i \times 1$ that includes ith subject measurements at times $\boldsymbol{t}_i = (t_{i1}, \cdots, t_i n_i)$, and y_{ij} is the outcome for the ith subject measured at the jth time point. According to this model, the joint models were developed Gaussian distributions for response and the

time distributed exponentially. Bronsert's Gaussian-Exponential model was given by:

$$f_{\Theta}(\boldsymbol{y_i}, \boldsymbol{t_i}) = \frac{1}{\sqrt{2\pi(\sigma^2)}} \exp\left(-\frac{1}{2}\frac{(y_{i1} - \boldsymbol{X}'_{i1}\boldsymbol{\beta})^2}{\sigma^2}\right) \times \prod_{j=2}^{n_i}\{\frac{1}{\sqrt{2\pi(\sigma^2)\sqrt{1-\rho_i^2}}} \exp\left(-\frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - \phi y_{i(j-1)} - \boldsymbol{X}'_{ij}\boldsymbol{\beta})^2}{\sigma^2(1-\rho_i^2)}\right) \exp(\alpha + \delta_i y_{i(j-1)}) \exp(-e^{\alpha+\delta_i y_i(j-1)} t_{ij})\}, \quad (2.18)$$

where:

$\boldsymbol{\beta}$ is the effect of the independent variables on outcomes,

$f(t_{i1})$ is the initial time point for the ith subject,

ϕ is the effect of the previous outcome on the mean response of the current outcome,

γ is the effect of current time on the mean response,

α is the constant parameter for the time process,

δ is the effect of the previous outcome on the mean time,

$\boldsymbol{X_{i1}}$ is the initial observations of k independent variables, and $\boldsymbol{X_{ij}}$ is $n \times (k+1)$ design matrix contains the observations of k independent variables, where n is the number of subjects.

Then, Lin (2011) adapted and modified Bronsert's Gaussian-exponential model by not including the term ρ_i^2 in his model because

he believed another term in the model ϕ could take care of relationships between two responses. Lin's modified model was:

$$f_{\Theta}(\boldsymbol{y_i}, \boldsymbol{t_i}) = \frac{1}{\sqrt{2\pi(\sigma^2)}} \exp(-\frac{1}{2}\frac{(y_{i1} - \boldsymbol{X}'_{i1}\boldsymbol{\beta})^2}{\sigma^2})$$
$$\times \prod_{j=2}^{ni} \{\frac{1}{\sqrt{2\pi(\sigma^2)}} \exp(-\frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - \phi y_{i(j-1)} - \boldsymbol{X}'_{ij}\boldsymbol{\beta})^2}{\sigma^2})$$
$$\exp(\alpha + \delta_i y_{i(j-1)}) \exp(-e^{\alpha+\delta_i y_i (j-1)} t_{ij})\}. \tag{2.19}$$

Later, Seo (2015) extended Bronsert's (2009) and Lin's (2011) Gaussian-exponential model to handle longitudinal outcomes from the exponential family of distributions with informative time that followed an exponential distribution. In this study, the maximum likelihood method was used for parameter estimation, a likelihood ratio test was computed for testing models, and AIC, AICc, and BIC were used for model selection.

One feature of longitudinal studies is the ability to observe the change over time on the responses and determine whether changes in covariates influenced the responses of interest. For example, in the Mothers' Stress and Children's Morbidity Study (MSCM), the researchers studied the relationship between maternal employment and pediatric healthcare utilization. The researchers conducted a daily follow-up of 167 preschool children and recorded measures of child illness (response) and maternal stress (time-dependent) for 28 days (Diggle, 2002). The measurements of these covariates were often taken at the same time. Thus, any changes that happened to their values were usually ignored. i.e., they were considered as

time-independent covariates. According to Lalonde, Nguyen, Yin, Irimata, and Wilson (2013), in longitudinal studies, as the values of the responses for the same subject can change over time, the covariate values can change at different time points as well. Time-dependent covariates with longitudinal data occur and the changes in some covariates over time need to be accounted for because the change in them influences the response.

To determine the appropriate method for a longitudinal response with a time-dependent covariate, factors that impact the covariate need to be defined. Amemiya (1985) specified endogenous as "variables that are stochastically determined by measured factors within the system under observation" and exogenous as "variables are determined by factors outside the system under study" (Diggle, 2002). A covariate process is an external covariate where the covariate at time t is conditionally independent of all previous response outputs. Exogenous is the opposite of endogenous and they are given by:

$$\text{Exogenous} : f(x_{it}|H_i^y(t), H_i^x(t-1), z_i) = f(x_{it}|H_i^x(t-1), z_i), \tag{2.20}$$

$$\text{Endogenous} : f(x_{it}|H_i^y(t), H_i^x(t-1), z_i) \neq f(x_{it}|H_i^x(t-1), z_i), \tag{2.21}$$

where x_{it} is a time-dependent covariate, z_i is a collection of baseline, $f(x)$ is a density function for the continuous covariate, $H_i^x(t)$ is the history of the covariate through time $x_{i1}, x_{i2}, ..., x_{it}$, and $H_i^y(t)$ is the history of the response outputs through time $y_{i1}, y_{i2}, ..., y_{it}$.

There are two important implications of exogeneity. First is the assumption that allows the factorization of the likelihood for (x_i, y_i) :

$$
\begin{aligned}
& f(\boldsymbol{x}_i, \boldsymbol{y}_i | \boldsymbol{z}_i; \boldsymbol{\theta}) \\
& = [\prod_{t=1}^{ni} f(y_{it} | H_t^y(t-1), H_i^x(t-1), z_i; \boldsymbol{\theta}] \times [\prod_{t=1}^{ni} f(x_{it} | H_i^x(t-1), z_i; \boldsymbol{\theta})] \\
& = \mathcal{L}_y(\boldsymbol{\theta}) \times \mathcal{L}_x(\boldsymbol{\theta}). \qquad (2.22)
\end{aligned}
$$

Assuming that $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$, where $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are variation independent parameters and that $\boldsymbol{\theta_1}$ is the parameter of interest. Then x_{it} defined as strongly exogenous for the parameter $\boldsymbol{\theta_1}$. If the assumptions are satisfied, the parameter of interest can be conditioned on the time-dependent covariate without losing information (Diggle, 2002).

The second implication of exogeneity is the expectation of the response conditional on all the covariates will depend only on the covariates prior to time t, given by:

$$
E(y_{it} | x_{i1}, x_{i2}, ..., x_{it}, ..., x_{ini}, z_i) = E(y_{it} | x_{i1}, x_{i2}, ..., x_{it-1}, z_i). \qquad (2.23)
$$

Y. Liang, Lu, and Ying, 2009 proposed a joint model for the analysis of longitudinal outcomes with irregular, informative time, and external time-dependent covariates based on latent variables. They used two models: (a) a semiparametric mixed effects model for the longitudinal outcomes, and (b) a frailty model used for observation times. The association among longitudinal outcomes, times, and external covariates

was modeled via latent variables. Their method required a specific distributional assumption of the frailty variable and the relationship between the two latent variables. Maximum likelihood estimation has been broadly used for joint models; however, in their model, this could be complicated to use because their model contained two nonparametric components. Moreover, their method could not be used for cases where time-dependent covariates depended on the longitudinal variable or informative time depended on time-dependent covariates.

Later,Song, Mu, and Sun (2012) presented a joint model of longitudinal data with time-dependent covariates and informative observation times via latent variables. They assumed time was informative and the longitudinal variable depended on time-dependent covariates. The joint model did not require specific assumptions on the distributions of the latent variables, which made it flexible. The joint model used in their study via z_1 and z_2, two latent variables, was:

$$E[y(t)|x(t), z_1, z_2] = \mu_0(t) + \boldsymbol{\beta}_0'\boldsymbol{x}(t) + z_1, \tag{2.24}$$

$$E[dN(t)|x(t), z_1, z_2] = z_2 exp(\gamma_0'\boldsymbol{x}(t))d\Lambda_0(t), \tag{2.25}$$

where:

$y(t)$ is the longitudinal response variable related to (z_1, z_2) through z_1,

$x(t)$ is the time-dependent covariates,

$N(t)$ is the number of observation times before or at time t (time informative), which is correlated to $y(t)$ and related to (z_1, z_2) through z_2,

$\mu_0(t)$ is an unspecified smooth function of t,

β_0 is a vector of unknown regression parameters,

γ_0 is vector of parameters,

$\Lambda_0(t)$ is an unknown non-decreasing function.

They assumed that the $E[z_1|x(t)] = 0, E[z_2|x(t)] = 1$, and $E[z_1 z_2|x(t)] = E[z_1 z_2]$. To check model (1), they used cumulative sums of residuals, and model (2) based on the recurrent event data. Their method could not deal with the situation where the informative observation times or time-dependent covariates depended on the longitudinal outcome. In addition, their model and estimation method was not that simple in the calculation.

Conclusion

Longitudinal studies have the power to measure change in outcomes and/or predictors at the individual level over time. A number of methods have been proposed for different outcomes types and researchers' design issues. However, most of these methods require a common assumption that time intervals are fixed and predetermined. In addition, a few methods consider the impact of time-dependent covariates on the outcomes. Also, rare methods consider situations where both time-dependent covariates and informative observation times occur together with longitudinal outcomes. According to the presented methods used for longitudinal data, the joint model seemed the appropriate method to consider for this study since it does not require time to be fixed. Finally, there is an obvious need for a method that can model longitudinal data with informative time and time-dependent covariates jointly for ease in interpretation.

CHAPTER III

METHODOLOGY

According to the review of literature, few approaches can be applied to observations collected over time that consider time dependent covariates along with the informative time. Although these approaches have been used in the past, the results were not that useful because the approaches had several restrictions, e.g., the time was fixed and the models and estimation methods were complicated. Thus, to simplify the analysis, this study aimed to extend the Gaussian-exponential model developed by Bronsert (2009) and extended by Lin (2011) and Seo (2015). This study assumed the responses and time-dependent covariates followed a normal distribution and time followed an exponential distribution.

Notation and Joint Model

The outcome for the ith individual measured at time j is donated by y_{ij}; so the ith individual has a vector of outcomes $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})'$ measured with a vector of a time-dependent covariate $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{in_i})'$ collected at a vector of time $\boldsymbol{t}_i = (t_{i1}, t_{i2}, ..., t_{in_i})'$. Note that the individuals range from $i = 1, ..., m$ and the outcomes, time-dependent covariate, and the time range from $j = 1, ..., n_i$, where n_i allows the measured time to vary from one individual to another individual. The joint distribution of $\boldsymbol{y}_i$, $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$ is in general $f_{\boldsymbol{\Theta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i)$,

$$f_{\boldsymbol{\Theta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i) = f_{\boldsymbol{\Theta}}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i) f_{\boldsymbol{\Theta}}(\boldsymbol{t}_i) f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i), \tag{3.1}$$

where $\boldsymbol{\Theta}$ is a vector of unknown parameters.

In order to explain the necessary methods that need to help answer each research question. The research questions marked in Chapter 1 are recalled. This dissertation addressed the following questions:

Q1 How can the joint model constructed by Lin (2011) for a longitudinal response variable with a set of informative time be extended to a joint model with time-dependent covariate?

Q2 How can the maximum likelihood estimators of the proposed joint model be obtained using R?

Q3 How can the likelihood ratio test be constructed to compare the fit of two models?

Q4 How can the model selection criteria, such as AIC, AICc, or BIC, be utilized to compare different models?

General Model

This section provides the basic background to answer the first research question:

Q1 How can the joint model constructed by Lin (2011) for a longitudinal response variable with a set of informative time be extended to a joint model with time-dependent covariate?

Offering an example here would help understand the idea of the general model. Let us say one individual has three outcomes $\boldsymbol{y} = (y_1, y_2, y_3)'$, measured with three observations of time-dependent covariate $\boldsymbol{x} = (x_1, x_2, x_3)'$, at three time points $\boldsymbol{t} = (t_1, t_2, t_3)'$, with a fixed independent variable $\boldsymbol{Z}$. The joint distribution of the outcomes, time-dependent covariate, and time points is

$f_{\Theta}(y_1, y_2, y_3, x_1, x_2, x_3, t_1, t_2, t_3|\boldsymbol{Z})$.

This joint distribution has the products of conditional and marginal distributions as follows:

$$
\begin{aligned}
&f_\Theta(y_1, y_2, y_3, x_1, x_2, x_3, t_1, t_2, t_3 | \boldsymbol{Z}) \\
&\quad = f_\Theta(y_3, x_3, t_3 | y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}) f_\Theta(y_2, x_2, t_2, y_1, x_1, t_1. \boldsymbol{Z}) \\
&\quad = f_\Theta(y_3 | y_2, x_2, t_2, y_1, x_1, t_1, x_3, t_3, \boldsymbol{Z}) f_\Theta(x_3 | y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}) \\
&\qquad \times f_\Theta(t_3 | y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}) f_\Theta(y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}). \quad (3.2)
\end{aligned}
$$

Based on the following assumptions:

- y_3 is independent of t_2, y_1, x_1, t_1 and x_3 given $y_2, x_2, t_3, \boldsymbol{Z}$, i.e.,

  $f_\Theta(y_3 | y_2, x_2, t_2, y_1, x_1, t_1, x_3, t_3, \boldsymbol{Z}) = f_\Theta(y_3 | y_2, x_2, t_3, \boldsymbol{Z})$,

- t_3 is independent of x_2, t_2, y_1, x_1, t_1 and $\boldsymbol{Z}$ given y_2, i.e.,

  $f_\Theta(t_3 | y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}) = f_\Theta(t_3 | y_2)$,

- x_3 is independent of y_2, t_2, y_1, x_1,and t_1 given x_2 and $\boldsymbol{Z}$, i.e.,

  $f_\Theta(x_3 | y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}) = f_\Theta(x_3 | x_2, \boldsymbol{Z})$.

Then the model can be expressed as:

$$
\begin{aligned}
&f_\Theta(y_1, y_2, y_3, x_1, x_2, x_3, t_1, t_2, t_3, \boldsymbol{Z}) \\
&= f_\Theta(y_3 | y_2, x_2, t_3, \boldsymbol{Z}) f_\Theta(t_3 | y_2) f_\Theta(x_3 | x_2, \boldsymbol{Z}) f_\Theta(y_2, x_2, t_2, y_1, x_1, t_1, \boldsymbol{Z}). \quad (3.3)
\end{aligned}
$$

The right hand can be expressed as:

$$
= f_\Theta(y_3 | y_2, x_2, t_3, \boldsymbol{Z}) f_\Theta(t_3 | y_2) f_\Theta(x_3 | x_2, \boldsymbol{Z}) f_\Theta(y_2, x_2, t_2 | y_1, x_1, t_1, \boldsymbol{Z}) f_\Theta(y_1, x_1, t_1, \boldsymbol{Z})
$$

$$= f_\Theta(y_3|y_2, x_2, t_3, \boldsymbol{Z}) f_\Theta(t_3|y_2) f_\Theta(x_3|x_2, \boldsymbol{Z}) f_\Theta(y_2|y_1, x_1, t_1, x_2, t_2, \boldsymbol{Z}) f_\Theta(t_2|y_1, x_1, t_1, \boldsymbol{Z})$$
$$\times f_\Theta(x_2|y_1, x_1, t_1, \boldsymbol{Z}) f_\Theta(y_1, x_1, t_1, \boldsymbol{Z}). \tag{3.4}$$

Based on the following assumptions:

- y_2 is independent of x_2 and t_1 given $y_1, x_1, t_2, \boldsymbol{Z}$, i.e.,

  $f_\Theta(y_2|y_1, x_1, t_1, x_2, t_2, \boldsymbol{Z}) = f_\Theta(y_2|y_1, x_1, t_2, \boldsymbol{Z})$

- t_2 is independent of x_1, $\boldsymbol{Z}$ and t_1 given y_1, i.e., $f_\Theta(t_2|y_1, x_1, t_1, \boldsymbol{Z}) = f_\Theta(t_2|y_1)$

- x_2 is independent of y_1 and t_1 given $x_1, \boldsymbol{Z}$, i.e.,

  $f_\Theta(x_2|y_1, x_1, t_1, \boldsymbol{Z}) = f_\Theta(x_2|x_1, \boldsymbol{Z})$.

Then the model can be expressed as:

$$= f_\Theta(y_3|y_2, x_2, t_3, \boldsymbol{Z}) f_\Theta(t_3|y_2) f_\Theta(x_3|x_2, \boldsymbol{Z}) f_\Theta(y_2|y_1, x_1, t_2, \boldsymbol{Z})$$
$$\times f_\Theta(t_2|y_1) f_\Theta(x_2|x_1, \boldsymbol{Z}) f_\Theta(y_1, x_1, t_1, \boldsymbol{Z}). \tag{3.5}$$

The right hand can be expressed as:

$$= f_\Theta(y_3|y_2, x_2, t_3, \boldsymbol{Z}) f_\Theta(t_3|y_2) f_\Theta(x_3|x_2, \boldsymbol{Z}) f_\Theta(y_2|y_1, x_1, t_2, \boldsymbol{Z}) f_\Theta(t_2|y_1)$$
$$\times f_\Theta(x_2|x_1, \boldsymbol{Z}) f_\Theta(y_1|x_1, t_1, \boldsymbol{Z}) f_\Theta(x_1) f_\Theta(t_1). \tag{3.6}$$

Finally, rewrite the equation by observation order:

$$= f_\Theta(y_1|x_1, t_1, \boldsymbol{Z}) f_\Theta(t_1) f_\Theta(x_1) f_\Theta(y_2|y_1, x_1, t_2, \boldsymbol{Z}) f_\Theta(y_3|y_2, x_2, t_3, \boldsymbol{Z}) f_\Theta(t_2|y_1)$$
$$\times f_\Theta(t_3|y_2) f_\Theta(x_2|x_1, \boldsymbol{Z}) f_\Theta(x_3|x_2, \boldsymbol{Z}). \tag{3.7}$$

So,the general model for ith individual measured at n_i times points with a one step dependency has the following general form:

$$f_\Theta(\boldsymbol{y_i}, \boldsymbol{x_i}, \boldsymbol{t_i}|\boldsymbol{z_i}) =$$
$$= f_\Theta(y_{i1}|x_{i1}, t_{i1}, \boldsymbol{Z}_i) f(x_{i1}) f(t_{i1}) [\prod_{j=2}^{n_i} f_\Theta(y_{ij}|y_{i(j-1)}, x_{i(j-1)}, t_{ij}, \boldsymbol{Z}_i) f_\Theta(t_{ij}|y_{i(j-1)})]$$
$$\times [\prod_{j=2}^{n_i} f_\Theta(x_{ij}|x_{i(j-1)}, \boldsymbol{Z}_i]. \tag{3.8}$$

Thus, the initial observation, y_{i1} is conditioned on the initial observation of the time-dependent covariate, x_{i1}, and on time of observation, t_{i1}. In addition, subsequent observations of the response variable, y_{ij} are conditioned on the most recent previous observation, y_{ij-1}, on the most recent observation of time-dependent covariates, x_{ij-1}, and on time of observation, t_{ij}. For the purpose of the likelihood function, $f(x_{i1})$ and $f(t_{i1})$ are assumed not to depend on Θ_i , so they can be ignored.

Gaussian Exponential Model

This joint model is for continuous outcomes normally distributed with a time-dependent covariate that is also normally distributed and informative time that is exponentially distributed. The Gaussian-exponential model (GE) can be expressed to handle time-dependent covariates as follows:

$$f_\Theta(\boldsymbol{y_i}, \boldsymbol{x_i}, \boldsymbol{t_i})$$
$$= \frac{1}{\sqrt{2\pi(\sigma_y^2)}} \exp(-\frac{1}{2} \frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}) f(x_{i1}) f(t_{i1})$$
$$\times \prod_{j=2}^{n_i} \frac{1}{\sqrt{2\pi(\sigma_y^2)}} \exp(-\frac{1}{2} \frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})$$

$$\times \exp(\alpha + \delta_i y_{i(j-1)}) \exp(-e^{\alpha+\delta_i y_i(j-1)} t_{ij})$$
$$\times \prod_{j=2}^{n_i} \frac{1}{\sqrt{2\pi(\sigma_x^2)}} \exp(-\frac{1}{2} \frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}), \tag{3.9}$$

where :

$\boldsymbol{z}_y$ is the $m \times (k_y + 1)$ design matrix for the independent variables that are related to the observations, y_{ij}, where m is the number of individuals and k_y is the number of independent variables with y,

$\boldsymbol{z}_x$ is the $m \times (k_x + 1)$ design matrix for the independent variables that are related to the time dependent covariate observations, x_{ij}, where m is the number of individuals and k_x is the number of independent variables with x,

y_{i1} is the initial observation,

x_{i1} is the initial time-dependent covariate,

t_{i1} is the initial time of observation,

y_{ij} is the jth observation for the ith individual,

x_{ij} is the jth time dependent covariate for the ith individual,

t_{ij} is the jth time point for the ith individual,

$\boldsymbol{\beta}_y$ is the effect of the independent variables on outcomes,

$\boldsymbol{\beta}_x$ is the effect of the independent variables on time-dependent covariate observation,

γ is the effect of current time on the mean response,

ϕ is the effect of the previous outcome on the mean response of the current outcome,

ψ is the effect of the previous time-dependent covariate on the mean response of the current outcome,

α is the constant parameter for the time process,

δ is the effect of the previous outcome on the mean time,

η is the effect of the previous time-dependent covariate on the current time dependent covariate.

The likelihood function, which is the product of the density functions for m individuals, is

$$
\begin{aligned}
&L(\boldsymbol{\Theta}; \boldsymbol{y}_1, ...\boldsymbol{y}, \boldsymbol{x}_1, ..., \boldsymbol{x}_m, \boldsymbol{t}_1, ..., \boldsymbol{t}_m) = \prod_{i=1}^{m} f_{\boldsymbol{\Theta}}(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{t}_i) \\
&= \prod_{i=1}^{m} \{ f_{\Theta}(y_{i1}|x_{i1}, t_{i1}, \boldsymbol{z}_{iy}) [\prod_{j=2}^{n_i} f_{\Theta}(y_{ij}|y_{i(j-1)}, x_{i(j-1)}, t_{ij}, \boldsymbol{z}_{iy}) f_{\Theta}(t_{ij}|y_{i(j-1)})] \\
&\times [\prod_{j=2}^{n_i} f_{\Theta}(x_{ij}|x_{i(j-1)}, \boldsymbol{z}_{ix})] \} \\
&= \prod_{i=1}^{m} \{ \frac{1}{\sqrt{2\pi(\sigma_y^2)}} \exp(-\frac{1}{2} \frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}) \\
&\times \prod_{j=2}^{n_i} \frac{1}{\sqrt{2\pi(\sigma_y^2)}} \exp(-\frac{1}{2} \frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}) \\
&\times \exp(\alpha + \delta_i y_{i(j-1)}) \exp(-e^{\alpha + \delta_i y_i (j-1)} t_{ij}) \\
&\times \prod_{j=2}^{n_i} \frac{1}{\sqrt{2\pi(\sigma_x^2)}} \exp(-\frac{1}{2} \frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}) \}. \qquad (3.10)
\end{aligned}
$$

The log-likelihood function for the *ith* individual in the above model becomes

$$
ln(L) = \ell_i = log[\frac{1}{\sqrt{2\pi(\sigma_y^2)}} \exp(-\frac{1}{2} \frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})
$$

$$\times \prod_{j=2}^{n_i} \frac{1}{\sqrt{2\pi(\sigma_y^2)}} \exp(-\frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})$$

$$\times \exp(\alpha + \delta_i y_{i(j-1)}) \exp(-e^{\alpha+\delta_i y_i(j-1)} t_{ij})$$

$$\times \prod_{j=2}^{n_i} \frac{1}{\sqrt{2\pi(\sigma_x^2)}} \exp(-\frac{1}{2}\frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta})^2}{\sigma_x^2})]$$

$$= log[\frac{1}{\sqrt{2\pi(\sigma_y^2)}}] + [(-\frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})]$$

$$+ \Sigma_{j=2}^{n_i}\{log[\frac{1}{\sqrt{2\pi(\sigma_y^2)}}] + [-\frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}]\}$$

$$+ \Sigma_{j=2}^{n_i}[(\alpha + \delta_i y_{i(j-1)}) + (-e^{\alpha+\delta_i y_i(j-1)} t_{ij})]$$

$$+ \Sigma_{j=2}^{n_i}\{log[\frac{1}{\sqrt{2\pi(\sigma_x^2)}}] + [-\frac{1}{2}\frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}]\}. \quad (3.11)$$

The log-likelihood function for the GE model for all individuals, which is the sum of the log-likelihood for each of m individual, is given by:

$$\ell = \Sigma_{i=1}^m \ell_i = \Sigma_{i=1}^m [log[\frac{1}{\sqrt{2\pi(\sigma_y^2)}}] + [(-\frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})]]$$

$$+ \Sigma_{i=1}^m \Sigma_{j=2}^{n_i}\{log[\frac{1}{\sqrt{2\pi(\sigma_y^2)}}]$$

$$+ [-\frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}]\}$$

$$+ \Sigma_{i=1}^m \Sigma_{j=2}^{n_i}[(\alpha + \delta_i y_{i(j-1)}) + (-e^{\alpha+\delta_i y_i(j-1)} t_{ij})]$$

$$+ \Sigma_{i=1}^m \Sigma_{j=2}^{n_i}\{log[\frac{1}{\sqrt{2\pi(\sigma_x^2)}}] + [(-\frac{1}{2}\frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2})]\}. \quad (3.12)$$

Or,

$$\ell = c + \Sigma_{i=1}^m [-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})]$$

$$+\Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_y^2)-\frac{1}{2}\frac{(y_{ij}-\gamma t_{ij}-y_{i(j-1)}\phi-x'_{(ij-1)}\psi-\boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}]$$

$$+\Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[(\alpha+\delta_i y_{i(j-1)})-e^{\alpha+\delta_i y_i(j-1)}t_{ij})]$$

$$+\Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_x^2)-\frac{1}{2}\frac{(x_{ij}-x'_{i(j-1)}\eta-\boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}], \tag{3.13}$$

where $c=\Sigma_{i=1}^{m}log\frac{1}{\sqrt{2\pi}}+\Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}log\frac{1}{\sqrt{2\pi}}+\Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}log\frac{1}{\sqrt{2\pi}}$.

Parameter Estimation

This section provides the basic background to answer the second research question:

Q2 How can the maximum likelihood estimators of the proposed joint model be obtained using R?

The estimation of the parameters can be obtained by the maximum likelihood estimation (MLE). This approach has multiple properties especially when there are a sufficiently large sample such as consistency, efficiency, and parameterization invariance. In this method, when the sample size is large, the parameter estimates are close to the true population parameters. In this study, the likelihood function was separated into three processes: the outcome process, the time process, and the time-dependent covariate process. Since each process has parameters that are independent of the other parameters in the other process and that can be proved by taking the derivative. For simplicity, the log-likelihood function was used to find the parameter estimates. Bronsert (2009) and Lin (2011) used a function in SAS/IML (called NLPDD) to find the maximum likelihood estimation from the log-likelihood function. The NLPDD is a nonlinear optimization function that combines

quasi-Newton and trust-region methods. Seo (2015) used a function in R called maxLik. This function uses maximum likelihood estimation and non-linear optimization and includes the Newton-Raphson maximization. In this current study, maxLik was used because it included a unified way of calling different optimizers.

Hypothesis Testing

This section provides the basic background to answer the third research question:

Q3 How can the likelihood ratio test be constructed to compare the fit of two models?

In the field of statistics, making decisions from a sample of a population is called inferential statistics. There are major methods for statistical inference: frequentist, Bayesian, and likelihood. Frequentist does require repeated sampling, Bayesian does require prior distribution, while Likelihood does not require the notion of repeated sampling or prior distribution where the parameter can be obtained by examination of the likelihood function (Rohde, 2014). Likelihood approach has several advantages such as easy to apply and easy to compute. There are several approaches for hypothesis testing using maximum likelihood estimators: Wald test, the score test, and the likelihood ratio test. Utilizing these approaches, significance tests of parameters and confidence intervals can be performed. The likelihood ratio test compares between the maximized log likelihoods of the two nested models (full vs. reduced models). To meet the requirement of nested models, all of the parameters in the reduced model must be contained in the general model. Where the full model contains the all parameters while the reduced model contains some of

the parameters. The hypotheses are maximized log-likelihood value for the null hypothesis H_0 and maximized log-likelihood value for the alternative hypothesis H_a. When the difference between the log-likelihood for the two models is statistically significant, then the full model is more appropriate than the reduced model. Using Rohde (1991)definition of generalized likelihood ratio: let $X = (X_1, ..., X_m)$ where $X_1, ..., X_m$ have joint pdf $f(\boldsymbol{x}; \theta_1, ..., \theta_k)$ for $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Omega_0}$ versus $H_a : \boldsymbol{\theta} \in \boldsymbol{\Omega} - \boldsymbol{\Omega_0}$ the test statistic is given by

$$\lambda(\boldsymbol{x}) = \frac{max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}_0} f(\boldsymbol{x}; \boldsymbol{\theta})}{max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} f(\boldsymbol{x}; \boldsymbol{\theta})} = \frac{f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_0)}{f(\boldsymbol{x}; \hat{\boldsymbol{\theta}})}, \tag{3.14}$$

we reject H_0 if $\lambda(\boldsymbol{x})$ is small compared to significance level of the test. Where $\lambda(\boldsymbol{x})$ is a valid test statistic that is not a function of unknown parameters, $\boldsymbol{\Omega}$ is the the total possible parameter space of $\boldsymbol{\theta}$, and $\boldsymbol{\Omega_0}$ is the the total possible estimate parameter space of $\hat{\boldsymbol{\theta}}$. This study used the likelihood ratio test because this method gives useful interpretations, ease of calculation, and is consistent with the previous study.

Model Selection

This section provides the basic background to answer the fourth research question:

Q4 How can the model selection criteria, such as AIC, AICc, or BIC, be utilized to compare different models?

Model selection is an important part of any statistical analysis. This study considered several model selection criteria such as the Akaike information criterion (AIC), the Akaike information criterion with correction (AICc), and the Bayesian

information criterion (BIC). The AIC measures the deviation between the fitted model and the true model and is defined as

$$\text{AIC}= 2k - 2ln(L),$$

where $ln(L)$ is the maximized value of the log-likelihood function of the model and k is the number of estimated parameters in the model. For example, the AIC for the Gaussian Exponential model becomes

$$\begin{aligned} AIC = 2k - 2\{&\Sigma_{i=1}^{m}[-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_x)^2}{\sigma_y^2})] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[(\alpha + \delta_i y_{i(j-1)}) - e^{\alpha+\delta_i y_i(j-1)}t_{ij})] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_x^2) - \frac{1}{2}\frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}]\}. \end{aligned} \tag{3.15}$$

The AIC picks the model that fits the data well with fewest parameters; a model with a low AIC value is better. Since the AIC is not consistent when the sample size, n, becomes large, the Akaike information criterion with correction (AICc) is used and is given by

$$AIC_C = AIC + \frac{2k(k+1)}{n-k-1}. \tag{3.16}$$

Bayesian information criterion (BIC) is given by

$$BIC = -2\ ln(L) + k\ log(n). \tag{3.17}$$

The model with a lower AIC and BIC values is better. The BIC is not desirable to use due to the potential to choose a model that is too simple; however, in simulation studies, the BIC was found to perform very well and outperformed the AIC as sample size increased (Fitzmaurice, 2008; Posada and Buckley, 2004).

Data Simulation

Most of the parameter values and simulation conditions were adopted from Bronsert (2009), Lin (2011), and Seo (2015) because this study was an extension to their studies. Monte Carlo simulations were used with SAS/IML and/or R in the previous studies to verify the properties of the MLEs for their models. In addition, parameters were assumed to be the same across subjects for simplicity of model form. The joint model developed in this current study assumed the observations and time-dependent covariate followed the normal distribution while the observations for the time intervals followed the exponential distribution.

The simulated data design contains two categorical variables with three levels each and two continuous variables associated with the response variable and another categorical variable with three levels and one continuous variable associated with the time-dependent covariate. The expression in regression term can be as shown below

$$y_{ij} = \beta_0 + \beta_{1y} z_{i1y} + \beta_{2y} z_{i2y} + \beta_{3y} z_{i3y} + \beta_{4y} z_{i4y} + \beta_{5y} z_{i5y} + \beta_{6y} z_{i6y} + \epsilon_{ijy}. \tag{3.18}$$

$$x_{ij} = \beta_0 + \beta_{1x} z_{i1x} + \beta_{2x} z_{i2x} + \beta_{3x} z_{i3x} + \epsilon_{ijx}. \tag{3.19}$$

The initial outcome was generated from the normal distribution with the mean adjusted by the initial time-dependent covariate and independent variables. Then

sequence outcomes were generated from the normal distribution with the average outcome calculated by the current time of observation, previous observed outcome, previous observed time-dependent covariate, and current independent variables. The times of observation then followed the exponential distribution with the mean adjusted by the previous outcome. Finally, the time-dependent covariate followed the normal distribution with the mean adjusted by the previous time-dependent covariate and the current independent variables. All these terms with the fixed parameter values were adopted from previous studies except for the parameters that are associated with the time-dependent covariate. They are however chosen to approximate the initial values from the previous studies (see Table 1).

Table 1

Parameter Values for Simulations

β_{0y}	β_{1y}	β_{2y}	β_{3y}	β_{4y}	β_{5y}	β_{6y}	β_{0x}	β_{1x}	β_{2x}	β_{3x}	ϕ	η	ψ	γ	α	δ	σ_y	σ_x
0.4	0.2	0.3	0.1	0.3	0.4	0.9	0.5	0.6	0.2	0.7	0.8	0.7	0.6	0.1	2	0.01	1	1
0.4	0.2	0.3	0.1	0.3	0.4	0.9	0.5	0.6	0.2	0.7	0.8	0.7	0.6	0.1	1	0.02	1	1
0.4	0.2	0.3	0.1	0.3	0.4	0.9	0.5	0.6	0.2	0.7	0.8	0.7	0.6	0.1	2	0.01	2	2
0.4	0.2	0.3	0.1	0.3	0.4	0.9	0.5	0.6	0.2	0.7	0	0	0	0.1	1	0.02	2	2
0.4	0.2	0.3	0.1	0.3	0.4	0.9	0.5	0.6	0.2	0.7	0	0	0	0.1	2	0.01	0.5	0.5
0.4	0.2	0.3	0.1	0.3	0.4	0.9	0.5	0.6	0.2	0.7	0.8	0.7	0.6	0.1	1	0.02	0.5	0.5

To check the assumption that the MLE is distributed as a multivariate normal, the simulation design included different sample sizes, number of observations, and design structures. Based on the literature, a number of researchers used a sample size of less than 200 and they believed it to be enough to see if the multivariate normality test shows a trend as sample size increases; for the replications, some of

the researchers used 500, 1,000, and 5,000 replications (Y. Liang et al., 2009; Lipsitz et al., 2002; Qiu et al., 2016; Song et al., 2012). This study considered the sample size to be less than 200 and the replications to be 2,000. Table 2 shows five sample sizes–(18, 36, 54, 90, and 180)–with four observation designs that were balanced or unbalanced with a different number of observations. One reason to have different sample sizes is to see if there is a trend in multivariate normality as the sample size increases. For example, in a sample size of 90 when the number of observations is 10 and 5, that means 45 subjects have 10 outcomes and the other 45 subjects have 5 outcomes, which leads to a total number of 675 observations.

Table 2
Simulation Designs

Scheme Number	Sample Size	Number of Observations	Design Structure	Total number of Observations
1	18	10	Balanced	180
2		5 and 3	Unbalanced	72
3		10 and 5	Unbalanced	135
4		20 and 6	Unbalanced	234
5	36	10	Balanced	360
6		5 and 3	Unbalanced	144
7		10 and 5	Unbalanced	270
8		20 and 6	Unbalanced	468
9	54	10	Balanced	540
10		5 and 3	Unbalanced	216
11		10 and 5	Unbalanced	405
12		20 and 6	Unbalanced	702
13	90	10	Balanced	900
14		5 and 3	Unbalanced	360
15		10 and 5	Unbalanced	675
16		20 and 6	Unbalanced	1170
17	180	10	Balanced	1800
18		5 and 3	Unbalanced	720
19		10 and 5	Unbalanced	1350
20		20 and 6	Unbalanced	2340

Two procedures were replicated 2,000 times: (a) the simulation for each sample size with the number of observations and (b) the calculation of the estimators. In addition, multivariate normality was tested with 2,000 sets of estimators by using the Henze-Zirkler test in R via the HZ.test function in the MVN package (Mecklin and Mundfrom, 2004). Model selection was the next step after parameter estimation and testing. Finally, for application on real data, the model is used to analyze the bladder cancer data provided in R in the package called survival. The data set contained 85 bladder cancer patients who were randomly assigned to two groups the placebo group (47) and the thiotepa treatment group (38). At each clinical visit, observation times in a month and the number of bladder tumors that occurred between clinical visits were gathered. Moreover, two covariates were measured the number of initial tumors and the size of the largest initial tumor. To demonstrate the performance of the R code, the Gaussian Exponential model was applied to the bladder dataset. In this dataset the time is informative and the time intervals is irregular, because the future visiting time is scheduled based on the recurrence of bladder tumor at the time of measurement. The outcome, y_{ij} is the natural logarithm of the number of observed tumours at time j plus 1 to avoid 0, $i = 1, ..., 85$. For time- independet covariate, let $z_i = 0$ if the patient was in the placebo group and 1 if the patient was in the thiotepa group, the number and the size of the initial tumors as continues variables, and time-dependet covariate, x_ij to be the natural logarithm of the total number of observed tumors within the last 6 months plus 1.

CHAPTER IV

RESULTS

The main goals of this study were to develop, evaluate, and construct R codes for the joint model with longitudinal outcomes and informative time with time-dependent covariate presented in the previous chapter. This chapter presents the simulation study to evaluate the asymptotic normality of the maximum likelihood estimators of the joint model. To achieve these goals, the researcher attempted to study the following research questions:

Q1 How can the joint model constructed by Lin (2011) for a longitudinal response variable with a set of informative time be extended to a joint model with time-dependent covariate?

Q2 How can the maximum likelihood estimators of the proposed joint model be obtained using R?

Q3 How can the likelihood ratio test be constructed to compare the fit of two models?

Q4 How can the model selection criteria, such as AIC, AICc, or BIC, be utilized to compare different models?

This chapter contains several sections as follows: In the first section, the components of the proposed Gaussian-Exponential Model are presented. The second section explains the simulation steps to evaluate the developed model by estimating model parameters. The third section shows the simulation results of the multivariate normality tests for the Gaussian-Exponential Model. In the fourth

section, the operations of how to conduct the likelihood ratio test are presented. Finally, the information criteria are computed for model selection criteria.

Gaussian-Exponential Model

The first research question was answered in chapter III by constructing the model. This model has three components that are independent of each other; thus, the log likelihood function can be written separately for each component.

The log likelihood function for the proposed Gaussian-Exponential model in the previous chapter was

$$\begin{aligned}
\ell =& c + \Sigma_{i=1}^{m}[-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2})] \\
&+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}] \\
&+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[(\alpha + \delta_i y_{i(j-1)}) - e^{\alpha+\delta_i y_i(j-1)}t_{ij})] \\
&+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_x^2) - \frac{1}{2}\frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}], \qquad (4.1) \\
\text{where} \quad & c = \Sigma_{i=1}^{m} log\frac{1}{\sqrt{2\pi}} + \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i} log\frac{1}{\sqrt{2\pi}} + \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i} log\frac{1}{\sqrt{2\pi}}.
\end{aligned}$$

As can be seen, there are three processes: the first two terms represent the outcome process, the third term is the time process, and the last term is the time-dependent covariate process. For better illustration, the parameter vector $\boldsymbol{\beta}_y$ is only present in the first two terms, the parameters that are associated with the time process α and δ can only be seen in the third term, and the parameter vector $\boldsymbol{\beta}_{\boldsymbol{x}}$ is shown in the fourth term only. Thus, the log likelihood function can be maximized independently for each component. To demonstrate, let

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \boldsymbol{\theta}_3 \end{bmatrix},$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ are three different sets of parameters, then the log likelihood function of $\boldsymbol{\theta}$ can be written as $\ell(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2) + \ell_3(\boldsymbol{\theta}_3)$. After that,

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ell_1(\boldsymbol{\theta_1})}{\partial \boldsymbol{\theta_1}} \\ \frac{\partial \ell_2(\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2}} \\ \frac{\partial \ell_3(\boldsymbol{\theta_3})}{\partial \boldsymbol{\theta_3}} \end{bmatrix} \text{ and } \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 \ell_1(\boldsymbol{\theta_1})}{\partial \boldsymbol{\theta_1} \partial \boldsymbol{\theta_1'}} & 0 & 0 \\ 0 & \frac{\partial^2 \ell_2(\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2} \partial \boldsymbol{\theta_2'}} & 0 \\ 0 & 0 & \frac{\partial^2 \ell_3(\boldsymbol{\theta_3})}{\partial \boldsymbol{\theta_3} \partial \boldsymbol{\theta_3'}} \end{bmatrix}. \tag{4.2}$$

Thus, the Fisher information matrix will be a diagonal block matrix of the form

$$E(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}) = \begin{bmatrix} E(\frac{\partial^2 \ell_1(\boldsymbol{\theta_1})}{\partial \boldsymbol{\theta_1} \partial \boldsymbol{\theta_1'}}) & 0 & 0 \\ 0 & E(\frac{\partial^2 \ell_2(\boldsymbol{\theta_2})}{\partial \boldsymbol{\theta_2} \partial \boldsymbol{\theta_2'}}) & 0 \\ 0 & 0 & E(\frac{\partial^2 \ell_3(\boldsymbol{\theta_3})}{\partial \boldsymbol{\theta_3} \partial \boldsymbol{\theta_3'}}) \end{bmatrix}. \tag{4.3}$$

Steps of Simulation

Step 1: A design matrix related to the outcome was generated with two continuous and two categorical variables with three levels each.

Step 2: A design matrix that related to the time-dependent covariate was generated with one continuous and one categorical variable with three levels each.

Step 3: A dataset that contains four variables (outcomes, time-dependent covariate, time, and subject) was created based on the relations among previous and current outcomes, the time-dependent covariate, and the current time with the fixed parameter values shown in Table 1.

Step 4: A nonlinear parameter optimization function in R called maxLik was utilized to compute the maximum likelihood estimators.

Step 5: A standard error from the Hessian matrix was used to standardize the parameter estimates.

Step 6: The previous steps were repeated 2,000 times, and then the Henze-Zirkler test was used to test the multivariate normality of the 20,000 sets of parameter estimates in outcome process, 10,000 sets of parameter estimates in outcome time-dependent covariate, and 4,000 sets of parameter estimates in time process. This test was conducted for the parameter estimates of the outcome, time-dependent covariate, and time processes separately. The total results come out of the above steps was 120 simulation conditions, since there were five sample sizes, four different observations, and six parameter schemes on the simulation designs.

This section aims to answer the second research question, where the R codes for the above steps can handle an outcome from Gaussian distribution, with the time-dependent covariate, which follows a normal distribution, and informative time, which follows an exponential distribution presented in Appendix B. After

having all the input to use the R program, the codes compute the estimators, likelihood ratio test statistic, and AIC, AICc, and BIC.

Simulation Results

Even though there were six different parameter schemes, the maximum likelihood estimators seem to be increasingly multivariate normal as the sample size number increased in the three processes (outcome, time-dependent covariate, and informative time), as shown in the figures (Figure 1, Figure 5, and Figure 9). In all the figures shown below, blue circles represent the balanced simulation design, and red triangles represent the unbalanced simulation design. In addition, the significance level of $\alpha = 0.05$ was utilized for the multivariate normality test for each parameter scheme. In addition, it looks like when the number of observations increases, the maximum likelihood estimators become multivariate normal for the three processes, as shown in the figures (Figure 3, Figure 7, and Figure 11). These general findings for the outcome process and time process are similar to what the previous studies found. Since there are three processes in this study, the researcher presents the results for each one separately.

Outcome Process

In this process, there were 10 parameters to be tested. As Table 3 and Figure 1 which plots the p-value of the multivariate normality test against the number of subject shows that most of the cases were failed to show significant evidence of non-normality when the number of subject became 90 or larger. Yet when the number of subjects is 18 or 36, the multivariate normality was showed evidence of non-normality. This process requires 1,000 observations in order to show

the multivariate normal as Figure 3 display that. Figure 2 was plotted the number of subjects versus normality case and Figure 4 was plotted the number of observations against the normality case as well, which was categorized by "Normal" when a p-value of the test was greater than 0.05 otherwise, it was "on-Normal". By looking at the Figure 1 through 4, it seemed unclear which of the design structures (balanced and unbalanced) became multivariate normality first.

Time-dependent Covariate Process

For the time-dependent covariate process, there were five parameters to be investigated. As can be seen in Figures 5 and 7 and Table 4, the multivariate normality test results for the six parameter schemes were against the number of subjects and number of observations. Most of the test results that show multivariate normality can be satisfied when the number of subjects became larger than 90 and the number of observations exceeded 1,000 observations. In this process, the multivariate normality is not stable when the sample sizes are 18, 36, and 54, which is a similar result as in the outcome process. In term of the different schemes, some of scheme give better multivariate normality results than the other which can indicate that fitting the right parameters can help to have best results.

Time Process

The multivariate normality results from two parameters in the time process are presented in Table 5. From Figures 9 and 10, it is clear that most of the cases have evidence of multivariate normality even with a small number of subjects, as shown in all schemes except scheme 3 which as mentioned in the previous process the values of the parameters can be affect the multivariate normality results. In

terms of the number of observations, Figures 11 and 12 represent the number of observation against the p-value of the multivariate normality test and the normality categories. The time process needs 500 observations in order to have multivariate normality, while the outcome process and time-dependent covariate required 1,000 observations. In general, when there are more parameters to test, it becomes more difficult to gain multivariate normality results.

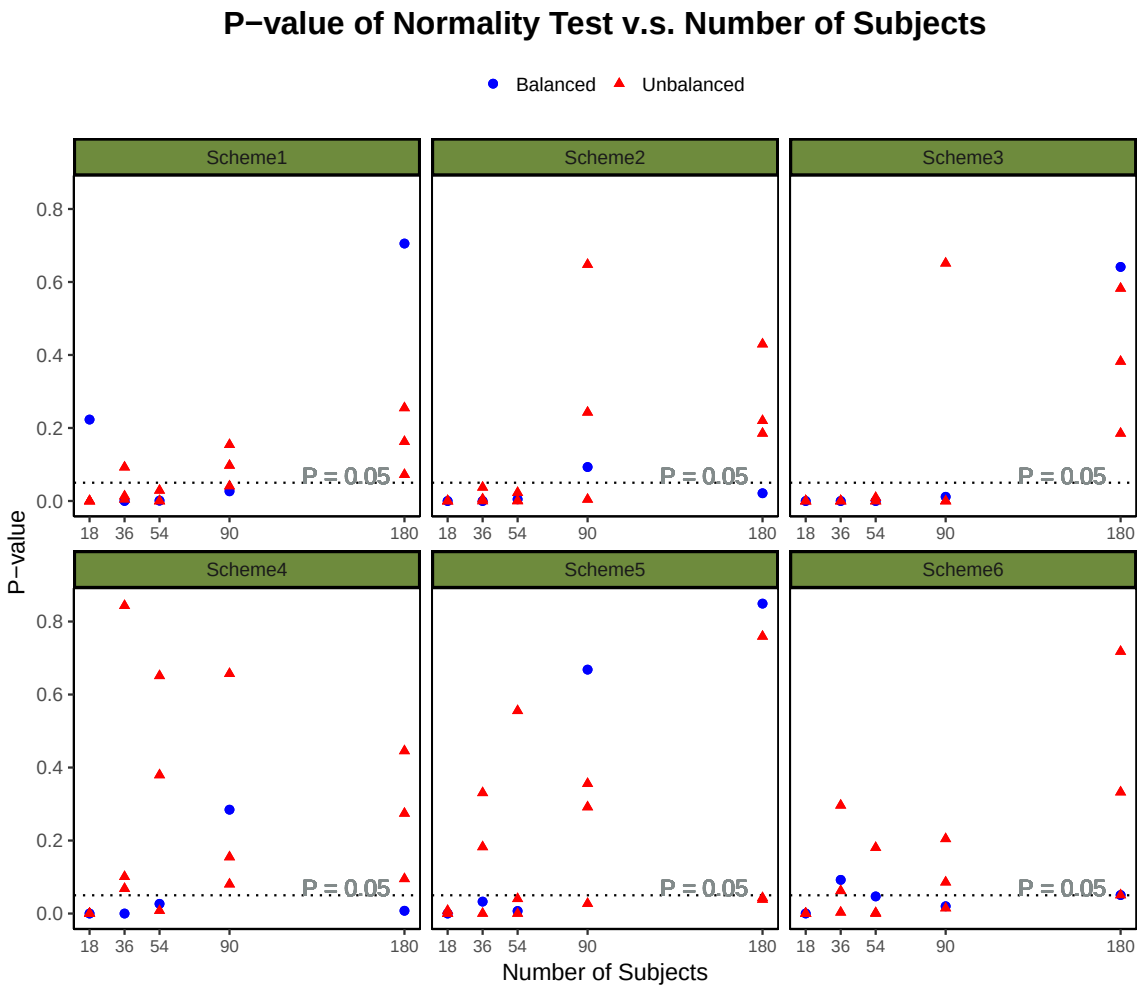

Figure 1. Outcome Process of the Gaussian-Exponential Model (Subjects).

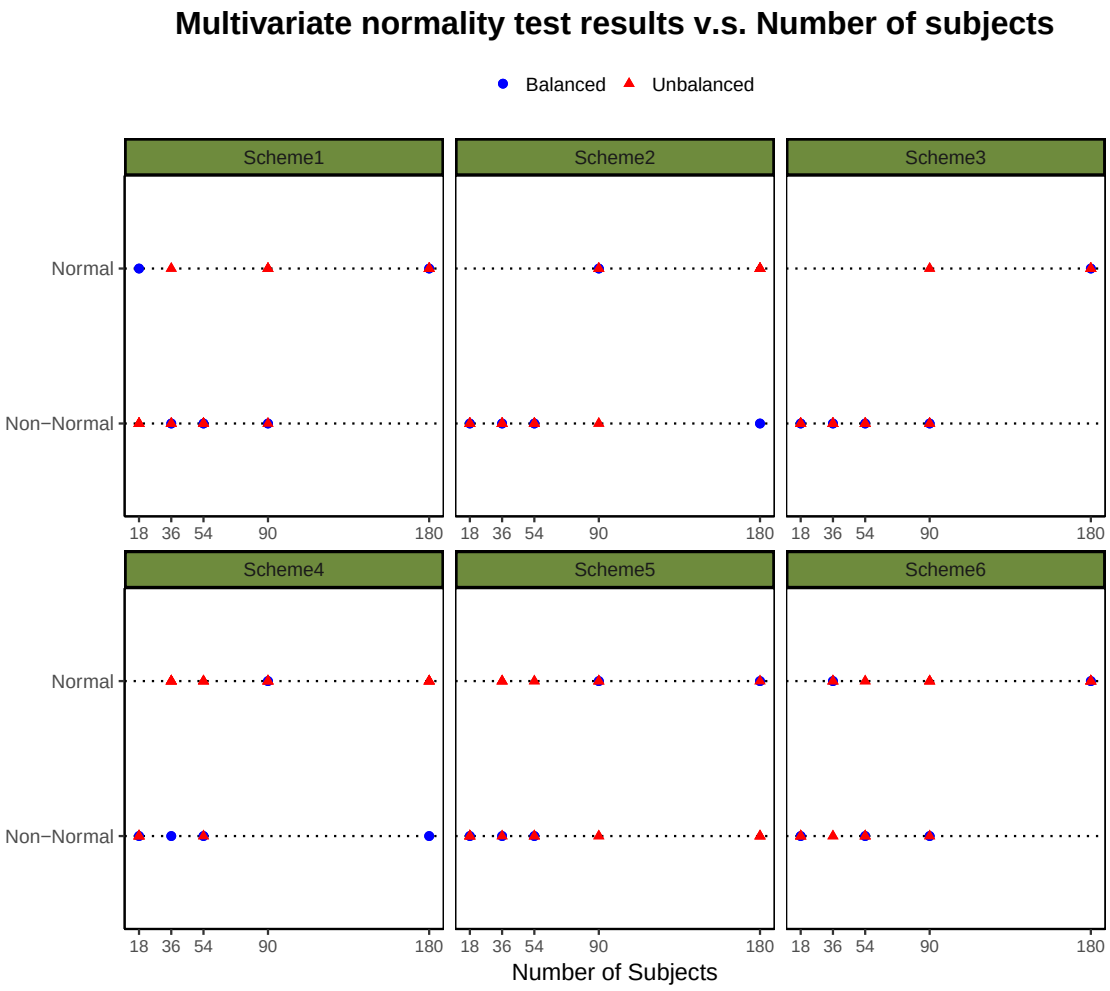

Figure 2. Outcome Process of the Gaussian-Exponential Model (Subjects).

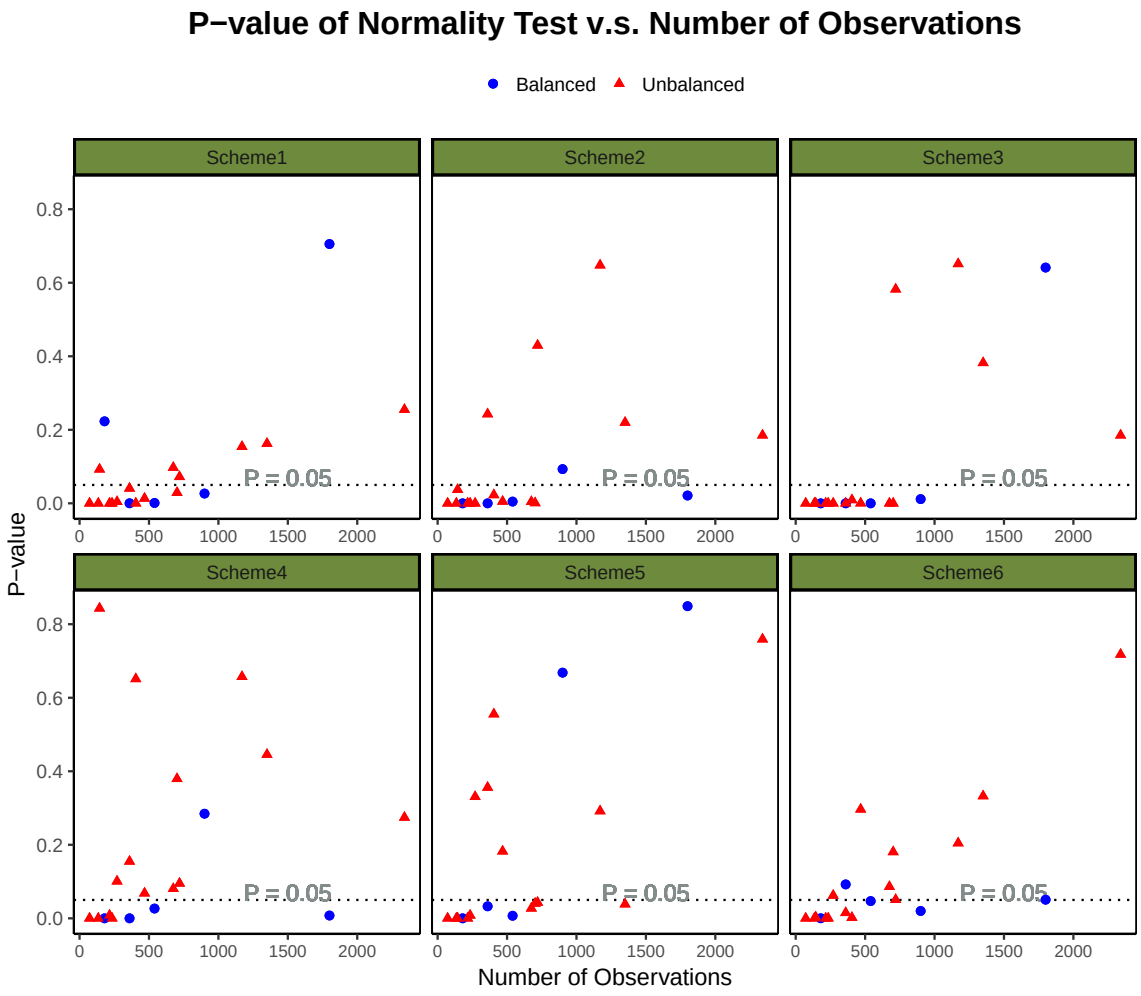

Figure 3. Outcome Process of the Gaussian-Exponential Model (Observations).

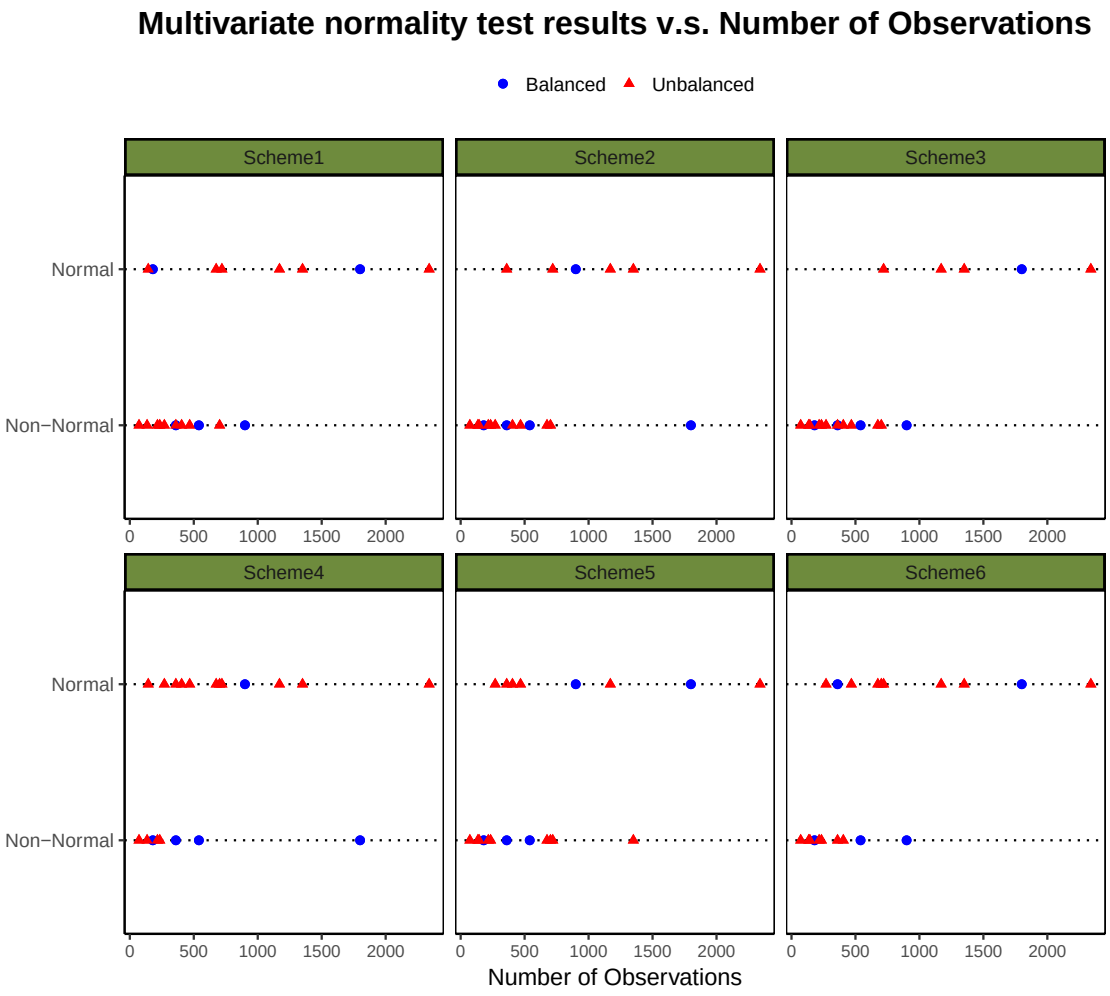

Figure 4. Outcome Process of the Gaussian-Exponential Model (Observations).

Table 3
Outcome Process of the Gaussian-Exponential Model: p-values of the Multivariate Normality Test for different parameter schemes and sample sizes

			Parameter Scheme					
Sample Scheme	Sample Size	Number of Obs.	1 p-value	2 p-value	3 p-value	4 p-value	5 p-value	6 p-value
1	18	180	0.00001	0.00001	0.00001	0.01893	0.25306	0.74697
2		72	0.00043	0.00001	0.00001	0.00001	0.00001	0.00001
3		135	0.00001	0.39425	0.00001	0.00016	0.00001	0.53843
4		234	0.00112	0.00855	0.00001	0.00023	0.29524	0.29286
5	36	360	0.64927	0.03972	0.091200	0.18606	0.83539	0.06618
6		144	0.07788	0.09712	0.01148	0.00740	0.69916	0.00098
7		270	0.06816	0.00001	0.00001	0.03982	0.67496	0.00631
8		468	0.37814	0.03943	0.08623	0.00016	0.80871	0.73476
9	54	540	0.95085	0.52148	0.01621	0.44887	0.19415	0.61150
10		216	0.00001	0.26435	0.00001	0.00011	0.00011	0.00016
11		405	0.46104	0.62916	0.19883	0.09479	0.06940	0.05871
12		702	0.59427	0.18153	0.00001	0.01605	0.31042	0.70631
13	90	900	0.80195	0.09630	0.67365	0.17567	0.22969	0.71188
14		360	0.14246	0.74305	0.14170	0.38963	0.04453	0.07450
15		675	0.00001	0.01679	0.00001	0.30071	0.00001	0.00001
16		1170	0.21147	0.10249	0.00001	0.91115	0.60275	0.78654
17	180	1800	0.18450	0.78151	0.00001	0.52464	0.78731	0.00001
18		720	0.00001	0.89535	0.00348	0.42743	0.71613	0.00001
19		1350	0.50515	0.64972	0.00001	0.18729	0.04919	0.83036
20		2340	0.00001	0.01801	0.01801	0.09816	0.94620	0.18858

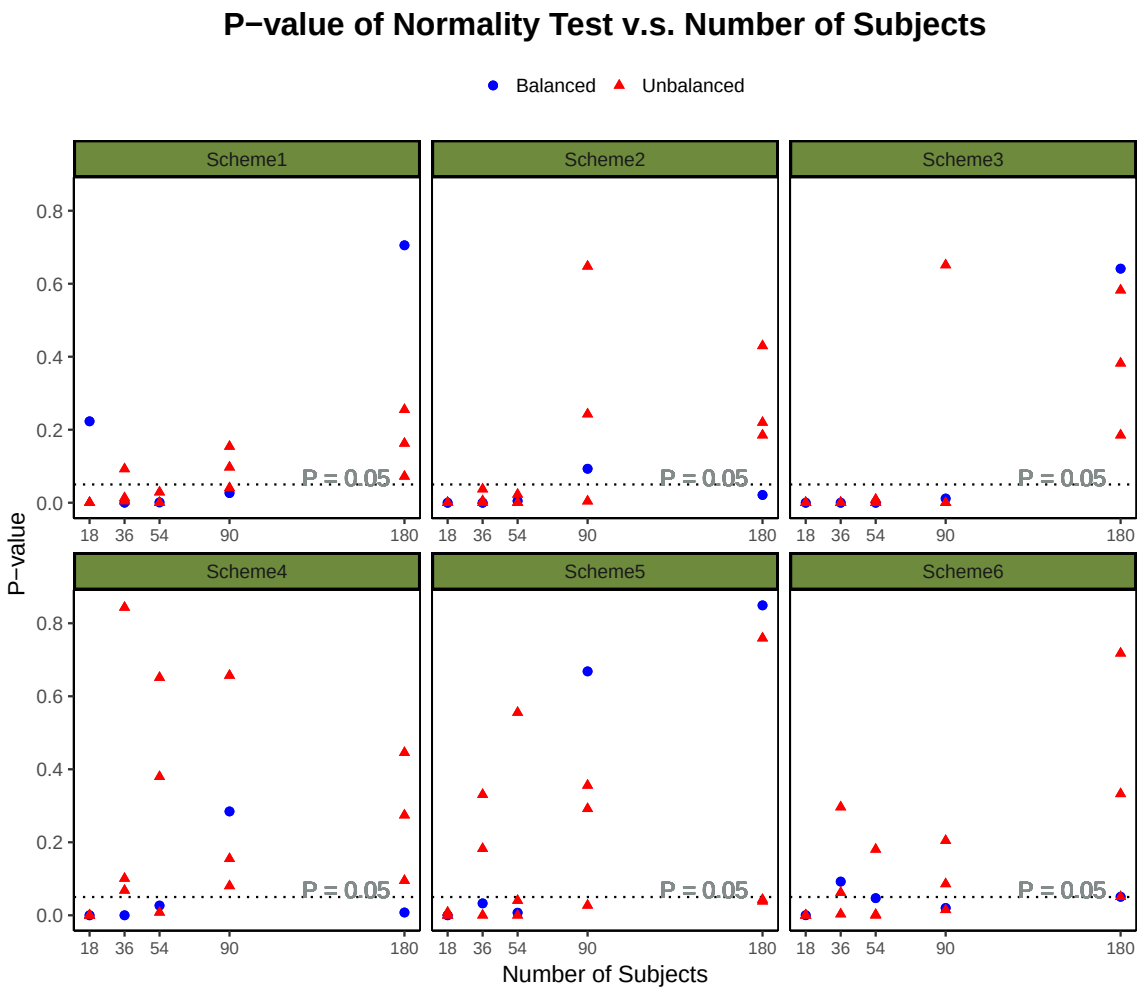

Figure 5. Time-Dependent Covariate Process of the Gaussian-Exponential Model (Subjects).

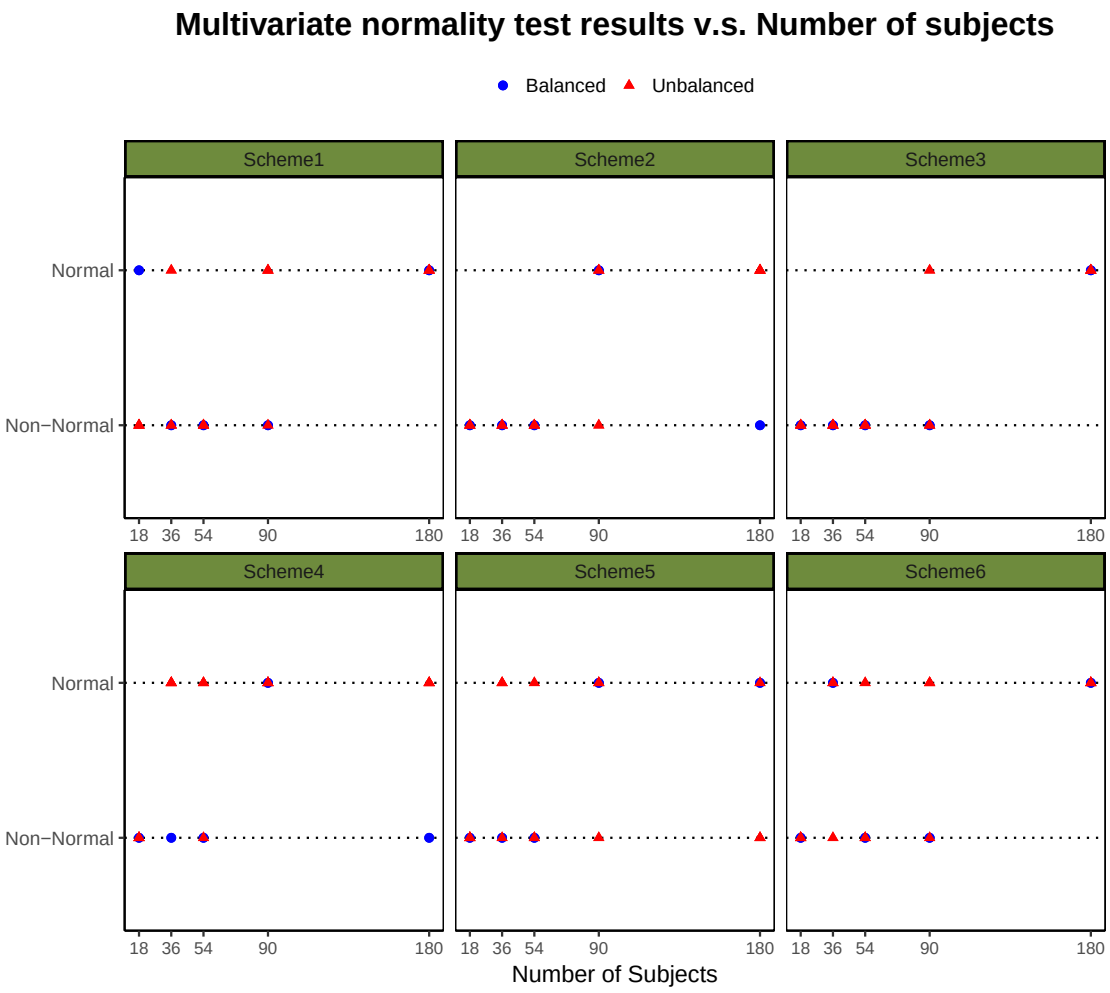

Figure 6. Time-Dependent Covariate Process of the Gaussian-Exponential Model (Subjects).

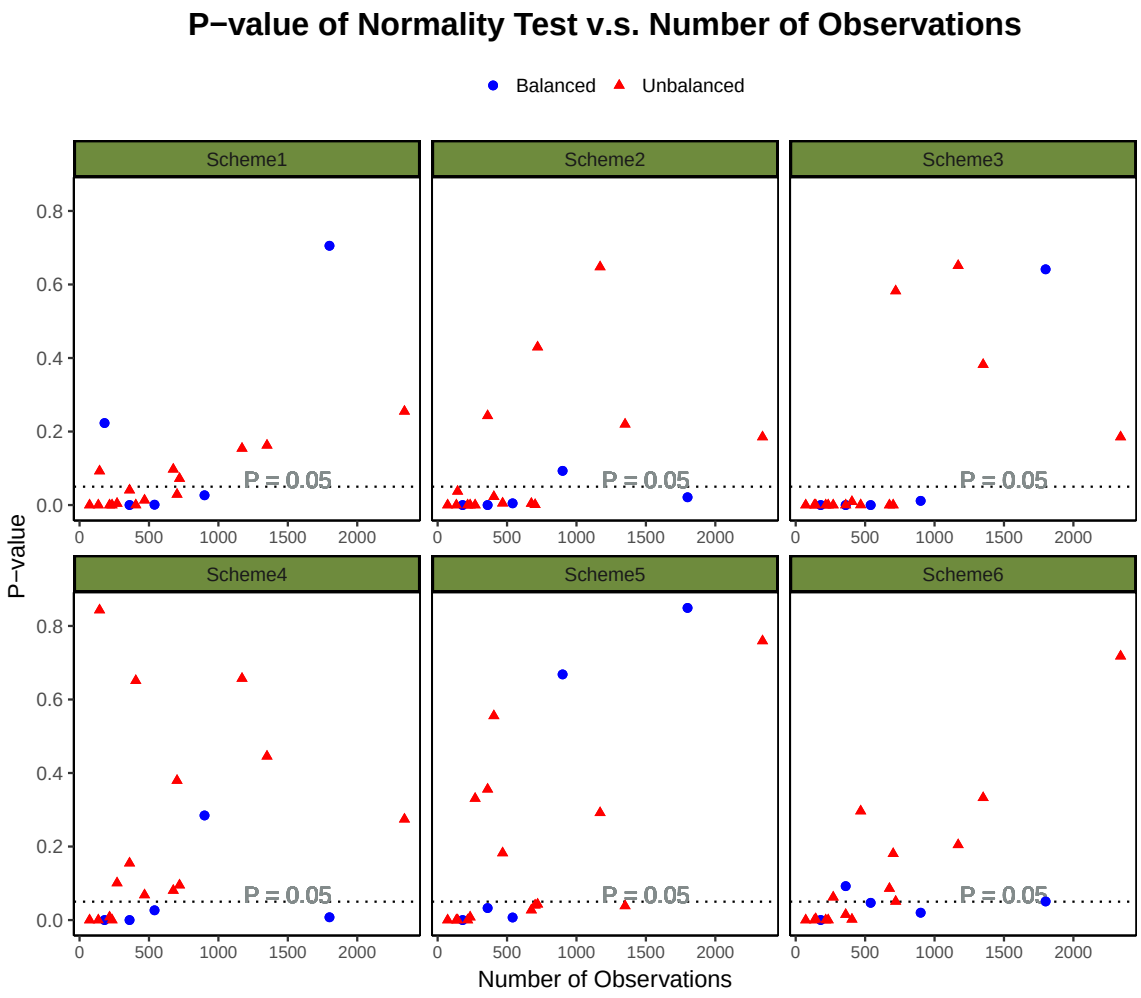

Figure 7. Time-Dependent Covariate Process of the Gaussian-Exponential Model (Observations) .

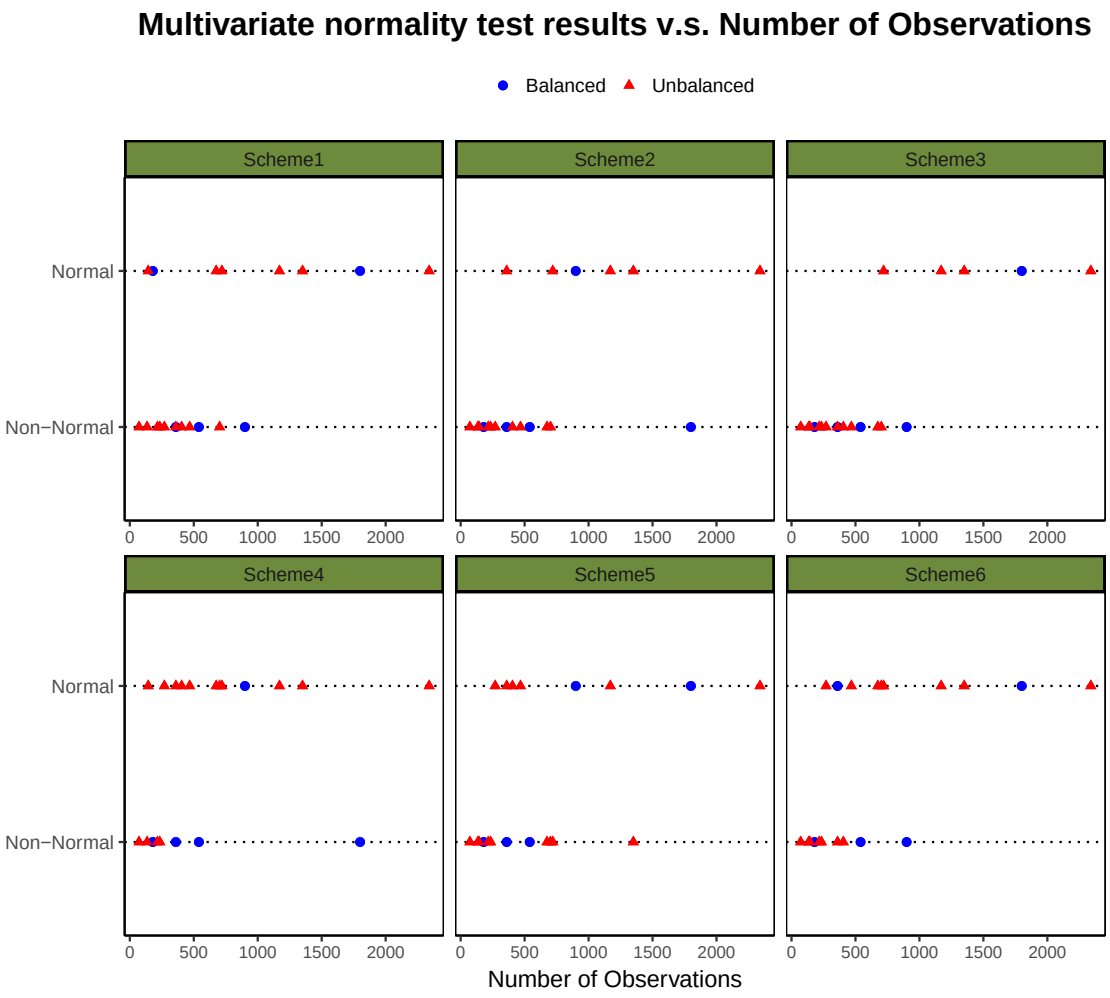

Figure 8. Time-Dependent Covariate Process of the Gaussian-Exponential Model (Observations).

Table 4
Time-Dependent covariate Process of the Gaussian-Exponential Model: p-values of the Multivariate Normality Test for different parameter schemes and sample sizes

			Parameter Scheme					
Sample Scheme	Sample Size	Number of Obs.	1 p-value	2 p-value	3 p-value	4 p-value	5 p-value	6 p-value
1	18	180	0.22308	0.00001	0.00001	0.00001	0.00001	0.00018
2		72	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
3		135	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
4		234	0.00001	0.00001	0.00001	0.00011	0.00796	0.000135
5	36	360	0.00017	0.00001	0.00001	0.00001	0.03271	0.09231
6		144	0.09253	0.03730	0.00001	0.84331	0.00010	0.00322
7		270	0.00427	0.00001	0.00001	0.10078	0.33065	0.06219
8		468	0.01324	0.00507	0.00030	0.06820	0.18245	0.29644
9	54	540	0.00099	0.00479	0.00001	0.02663	0.00698	0.04694
10		216	0.00001	0.00001	0.00001	0.00800	0.00001	0.00001
11		405	0.00014	0.02279	0.00913	0.65124	0.55543	0.00192
12		702	0.02893	0.00096	0.00001	0.37952	0.04034	0.18051
13	90	900	0.02664	0.09308	0.01149	0.28455	0.66828	0.01999
14		360	0.04063	0.24280	0.00010	0.15479	0.35583	0.014969
15		675	0.09727	0.00426	0.00001	0.08042	0.02715	0.08584
16		1170	0.15404	0.64767	0.65124	0.65706	0.29174	0.20460
17	180	1800	0.70533	0.02124	0.64132	0.00762	0.84912	0.05071
18		720	0.07212	0.42969	0.58233	0.09503	0.04301	0.05071
19		1350	0.16253	0.21980	0.38211	0.4455	0.03846	0.33262
20		2340	0.25488	0.18513	0.18513	0.274277	0.75924	0.71776

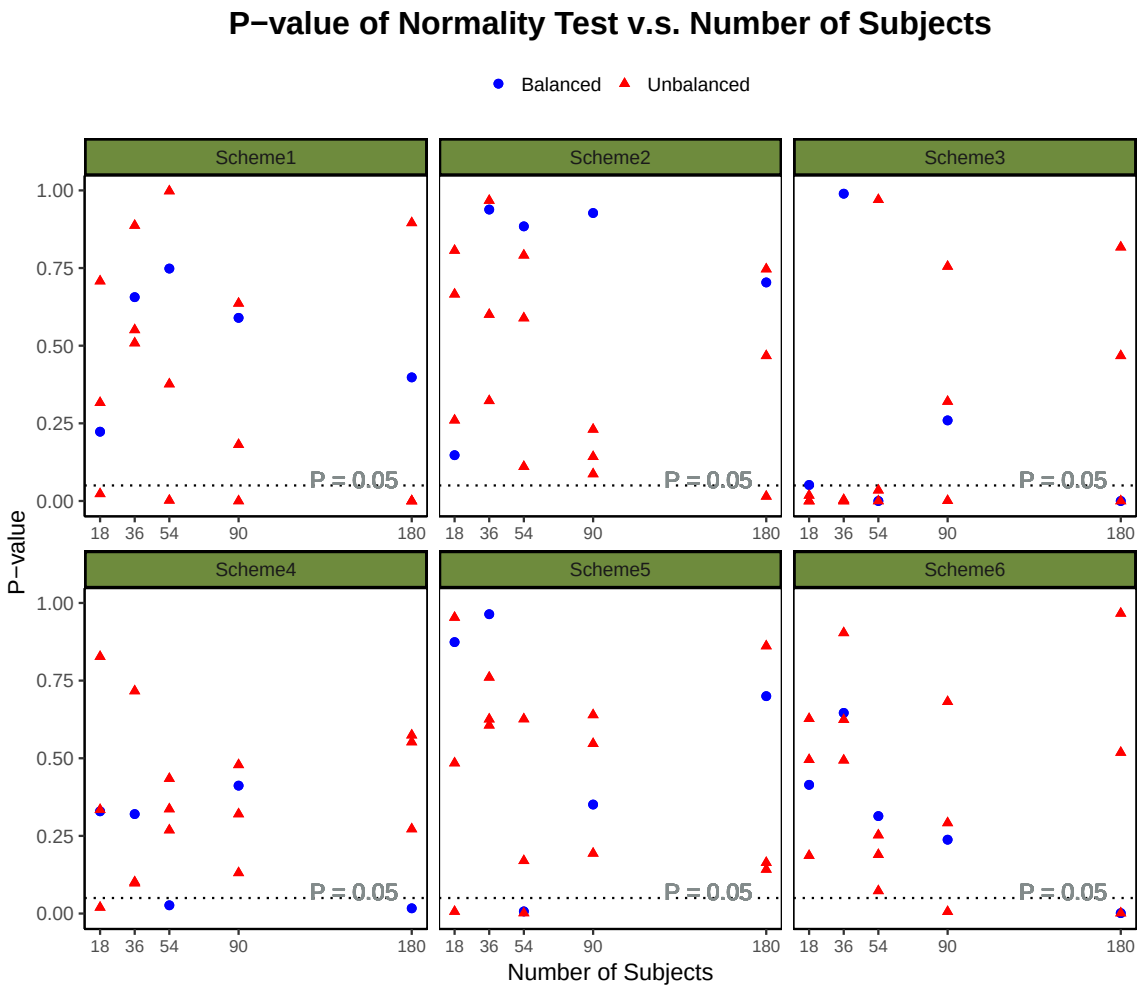

Figure 9. Time Process of the Gaussian-Exponential Model (Subjects).

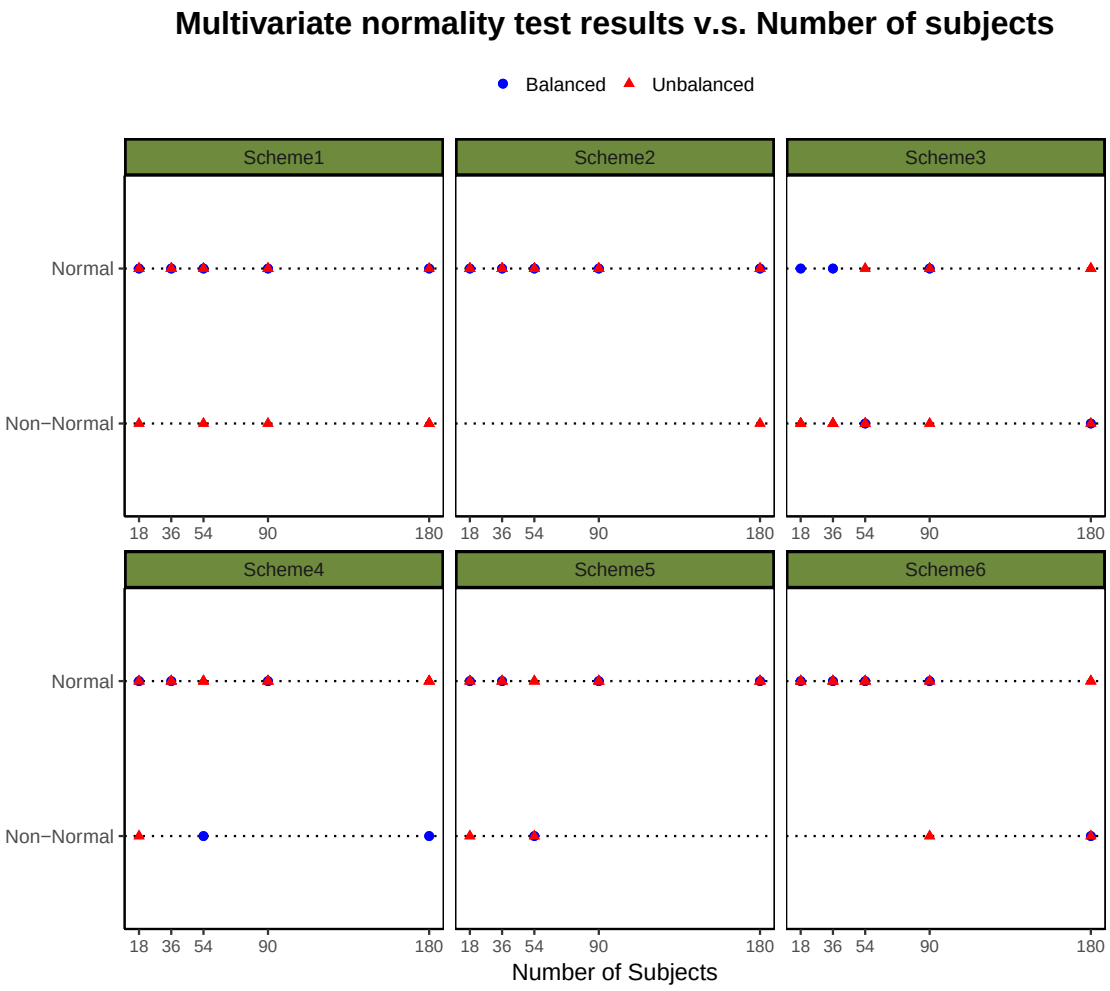

Figure 10. Time Process of the Gaussian-Exponential Model (Subjects).

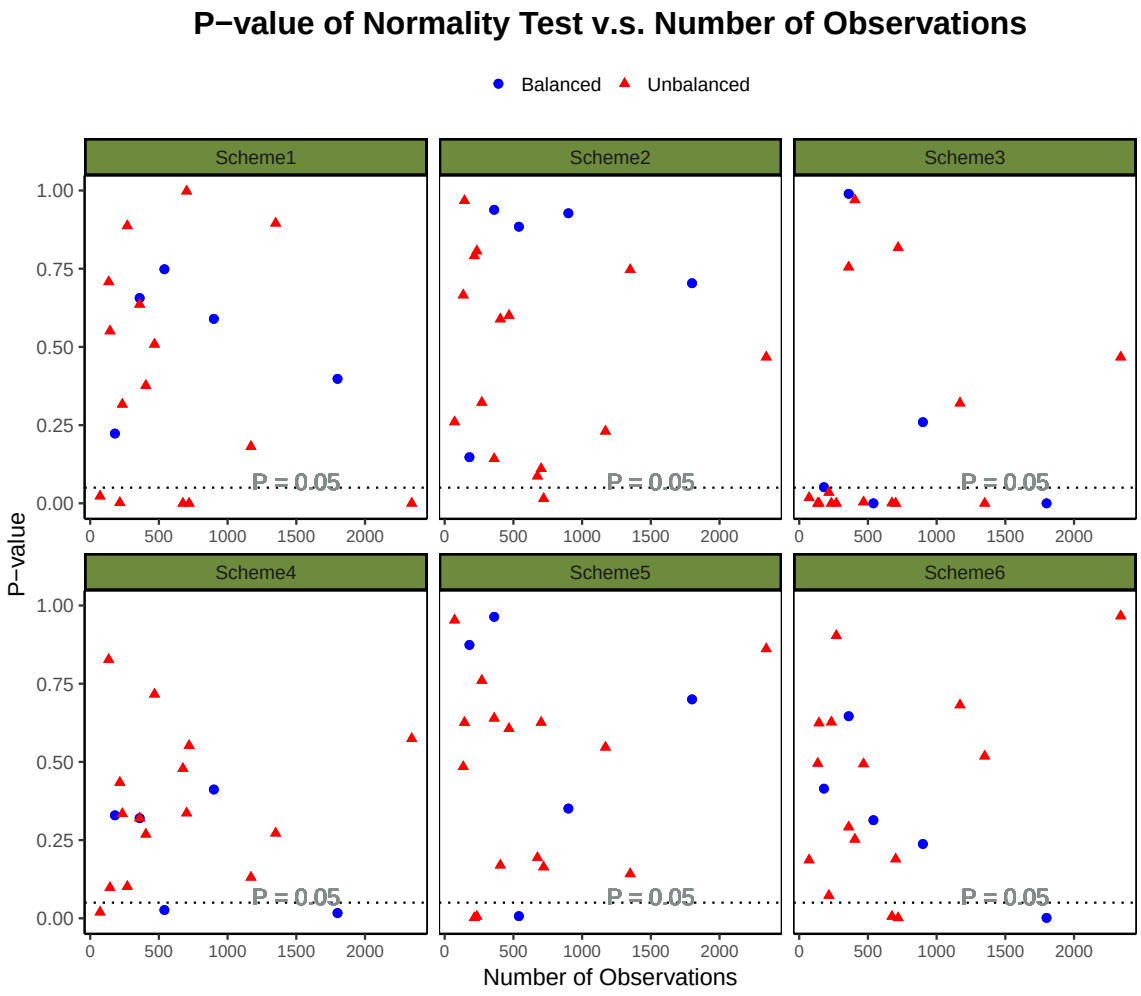

Figure 11. Time Process of the Gaussian-Exponential Model (Observations).

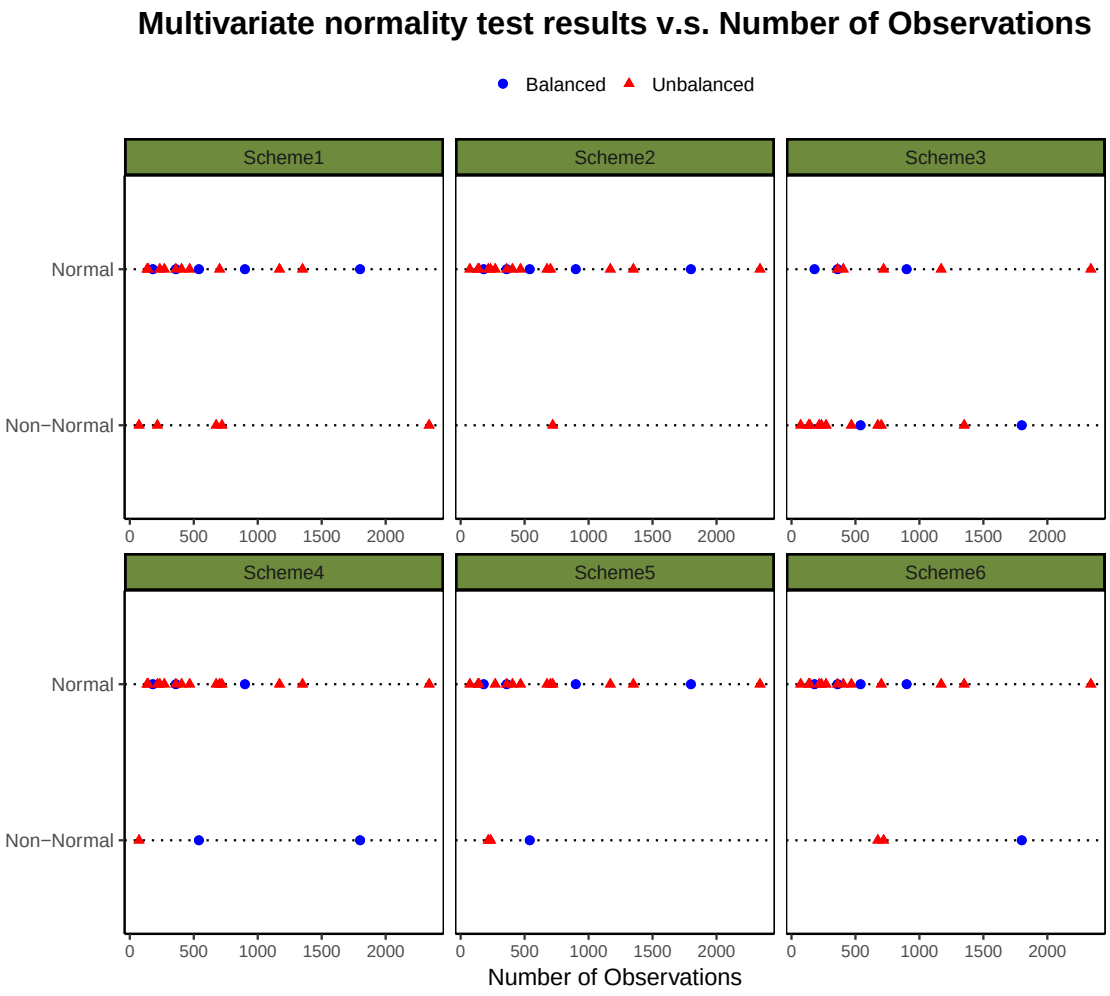

Figure 12. Time Process of the Gaussian-Exponential Model (Observations).

Table 5
Time Process of the Gaussian-Exponential Model: p-values of the Multivariate Normality Test for different parameter schemes and sample sizes

			Parameter Scheme					
Sample Scheme	Sample Size	Number of Obs.	1 p-value	2 p-value	3 p-value	4 p-value	5 p-value	6 p-value
1	18	180	0.22308	0.14755	0.05167	0.32928	0.87397	0.41446
2		72	0.02296	0.26009	0.01794	0.01962	0.95321	0.18646
3		135	0.70792	0.66569	0.00001	0.82716	0.48441	0.49516
4		234	0.31646	0.80649	0.00014	0.33437	0.00633	0.62719
5	36	360	0.65634	0.93838	0.98944	0.32025	0.96384	0.64607
6		144	0.55080	0.96734	0.00031	0.09799	0.62582	0.62432
7		270	0.88690	0.32249	0.00001	0.10156	0.76036	0.90359
8		468	0.50842	0.60029	0.00402	0.71645	0.60615	0.49357
9	54	540	0.74836	0.88432	0.00001	0.02663	0.00698	0.31379
10		216	0.00200	0.79112	0.03471	0.43434	0.00172	0.07293
11		405	0.37644	0.58897	0.96992	0.26868	0.16986	0.25264
12		702	0.99790	0.11057	0.00001	0.33649	0.62623	0.18941
13	90	900	0.58969	0.92725	0.25960	0.41185	0.35095	0.23774
14		360	0.63623	0.14239	0.75502	0.32038	0.63948	0.29173
15		675	0.00001	0.08691	0.00095	0.47864	0.19360	0.00587
16		1170	0.18127	0.23043	0.32025	0.13090	0.54681	0.68196
17	180	1800	0.39805	0.70370	0.00001	0.01693	0.69998	0.00143
18		720	0.00001	0.01476	0.81714	0.55214	0.16386	0.00143
19		1350	0.89536	0.74668	0.00001	0.27215	0.14197	0.51839
20		2340	0.00001	0.46735	0.46735	0.57429	0.86126	0.96630

Likelihood Ratio Test

In order to answer research question three, the simulation results show that the maximum likelihood estimators for the parameters of the joint model have the asymptotic normal distribution. Thus, the asymptotic normality assumption of the likelihood ratio test has been verified under sufficient sample size. Therefore, this test can be used for the joint model for comparing nested models.

The likelihood ratio test statistic can be presented as a twice difference in the maximized log-likelihoods of two nested models as they can be expressed,

$$2(\hat{l}_{full} - \hat{l}_{red}) \sim \chi^2_{df_{full} - df_{red}} \ , \tag{4.4}$$

where the $\hat{l}_{full}$ is the the maximized log-likelihood of the full model and $\hat{l}_{red}$ is the the maximized log-likelihood of the reduced model. This test statistic can be compared to a chi-squared distribution. By giving an example of how this test can be utilized, conisider the following nested models,

The maximized log-likelihood of the full model of the Gaussian-Exponential model is,

$$\begin{aligned}\hat{\ell}_{full} =& \Sigma_{i=1}^{m}[-\frac{1}{2}log(\sigma^2) - \frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_i\boldsymbol{\beta})^2}{\sigma^2})] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma^2) - \frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_i\boldsymbol{\beta})^2}{\sigma^2}].\end{aligned} \tag{4.5}$$

The maximized log-likelihood of the reduced model of the Gaussian-Exponential model is,

$$\begin{aligned}\hat{\ell}_{red} =&\Sigma_{i=1}^{m}[-\frac{1}{2}log(\sigma^2) - \frac{1}{2}\frac{(y_{i1} - x'_{i1} \times 0 - \boldsymbol{z}'_i \times \mathbf{0})^2}{\sigma^2})] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma^2) - \frac{1}{2}\frac{(y_{ij} - 0 \times t_{ij} - y_{i(j-1)} \times 0 - x'_{(ij-1)} \times 0 - \boldsymbol{z}'_i \times \mathbf{0})^2}{\sigma^2}].\end{aligned} \tag{4.6}$$

Then the likelihood ratio test statistic can be computed by multiplying two to the difference between the two maximized log-likelihoods.

Information Criteria

With respect to research question four, the information criteria AIC, AICc, and BIC are computed. The AIC is given by $2k - 2\ell(\hat{\Theta})$, where $\ell(\hat{\Theta})$ is the maximized log-likelihood evaluated at $\hat{\Theta}$ and k is the number of parameters. Also, the AICc is expressed by $AIC + \frac{2k(k+1)}{n-k-1}$ and the BIC can be expressed by $-2\ell(\hat{\Theta}) + k + k \times log(n)$. Thus, the AIC for the Gaussian-Exponential model can then be given as,

$$\begin{aligned}AIC = 2k - 2\{&\Sigma_{i=1}^{m}[-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{i1} - x'_{i1}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_x)^2}{\sigma_y^2})] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_y^2) - \frac{1}{2}\frac{(y_{ij} - \gamma t_{ij} - y_{i(j-1)}\phi - x'_{(ij-1)}\psi - \boldsymbol{z}'_{iy}\boldsymbol{\beta}_y)^2}{\sigma_y^2}] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[(\alpha + \delta_i y_{i(j-1)}) - e^{\alpha+\delta_i y_i(j-1)}t_{ij})] \\ &+ \Sigma_{i=1}^{m}\Sigma_{j=2}^{n_i}[-\frac{1}{2}log(\sigma_x^2) - \frac{1}{2}\frac{(x_{ij} - x'_{i(j-1)}\eta - \boldsymbol{z}'_{ix}\boldsymbol{\beta}_x)^2}{\sigma_x^2}]\}.\end{aligned} \tag{4.7}$$

These model selection criteria applied on computing program in R with the likelihood ratio test.

Analysis of Bladder Cancer Study

For the purpose of illustration, we apply the proposed joint model to the longitudinal bladder cancer data discussed in chapter 3 in the Gaussian Exponential Model section, which have been analyzed widely in the literature (J. Sun, Park, Sun, and Zhao, 2005; J. Sun, Sun, and Liu, 2007; Y. Liang, Lu, and Ying, 2009; L. Sun, Song, and Zhou, 2011; Cai, Lu, and Zhang, 2012; Seo, 2015). This study was conducted by the Veterans Administration Cooperative Urological Research Group, where the clinical visit in month and the number of bladder tumors that occurred between clinical visits for 85 patients were recorded. The dataset contained the number of initial tumors before entering the study and the size of the largest initial tumor as a two baseline covariates. Note that, at the beginning of the study, all tumors were removed and, during the study, the recurrent tumors were also removed at each clinical visit and many patients had multiple recurrences of tumors. The patients were randomly assigned to the treatment groups: the placebo group includes 47 patients and the thiotepa group includes 38 patients. As can be seen in Figure 13, it is clear that the patients' visiting times varied, where the total follow-up is 53 months. Since the patients in the thiotepa group tended to come more often compared with the patients in the placebo group and the tumor recurrence rates were different, this made the time intervals become irregular across all subjects and indicated that the visiting process may be informative to the tumor recurrence rate. The goal here is to study the relationship between the possible

informative clinical visit times and tumor recurrence process with adjustment for the total number of observed tumors within the last 6 months as a time-dependent covariate and the effect of the following covariates: treatment group, number of the initial tumors, and the size of the largest initial tumor.

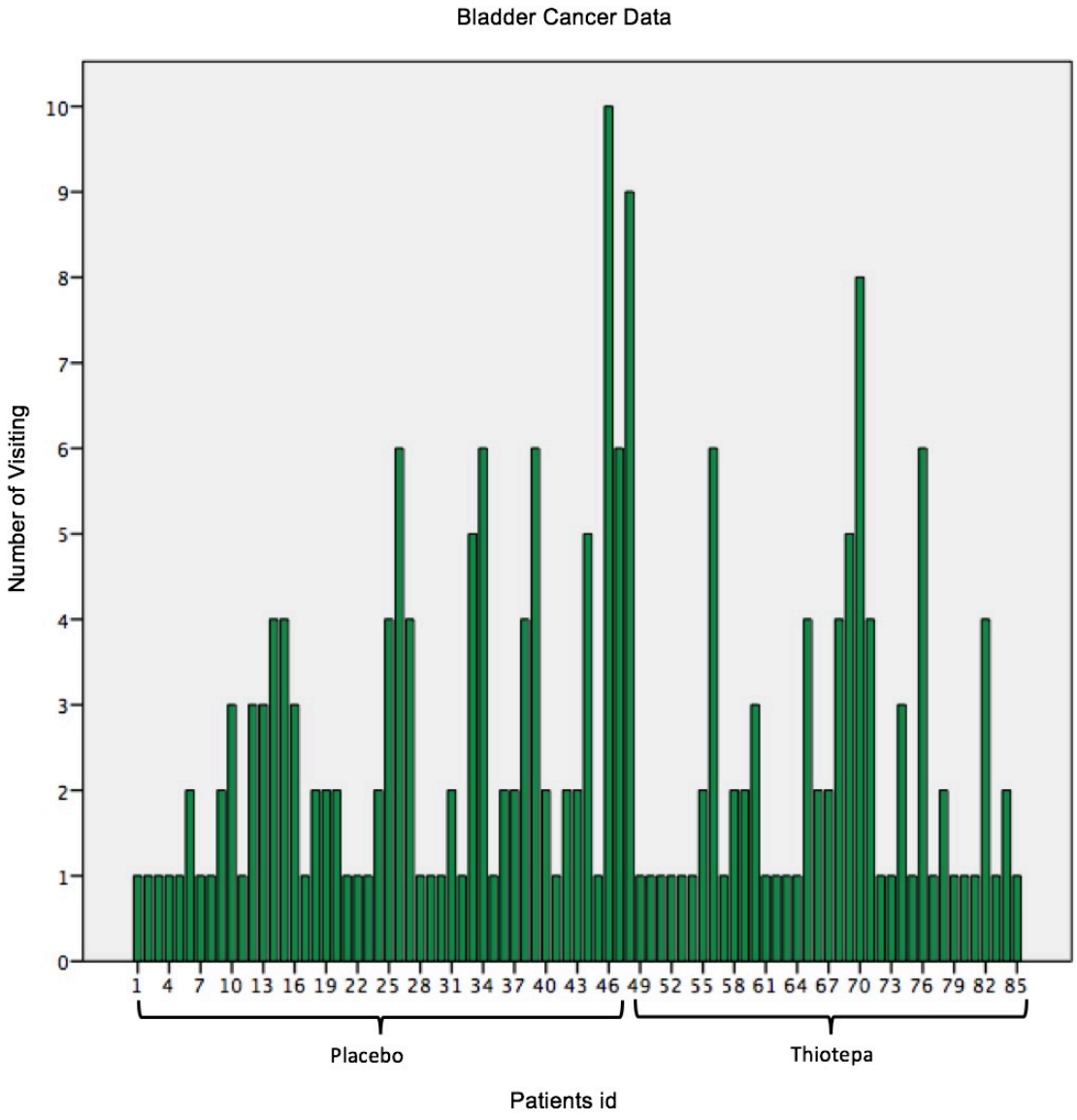

Figure 13. Bladder Cancer Data.

To analyze the dataset, the Gaussian-Exponential model was applied. For subject i, let y_{ij} stand for the natural logarithm of the number of tumors observed at time j plus 1 (to avoid 0). For the time-dependent covariate, let x_{ij} be the natural logarithm of the total number of observed tumors within the last 6 months, plus 1 (J. Sun, Park, Sun, and Zhao, 2005; Y. Liang, Lu, and Ying, 2009). In addition, for the other covariates, let z_{1i} be the treatment indicator (1, if the subject is from the placebo group; 0, if the subject is from the thiotepa group), let z_{2i} be the natural logarithm of the number of the initial tumors plus 1, and let z_{3i} be the natural logarithm of the size of the largest initial tumor plus 1. Moreover, for the informative time, t_{ij} stands for the difference between the current time of visit and the prior one to have an exponential distribution.

The complete bladder dataset is available in R under the survival package and in the Data book by Schmee et al. (1987). There are three data sets in R (*bladder1, bladder, and bladder2*) in that package. In this application *bladder1* was used since it contains all the variables required for the analysis, such as the number of tumors found at time j (rtumor), the month of the visit (stop), the treatments (treatment), initial number of tumors (number), and size of the largest initial tumor (size). *Bladder1* has the information for 118 subjects with three treatments, but this study analyzed the data for 85 subjects from the two treatments (placebo and thiotepa) because several studies consider the two groups (J. Sun, Park, Sun, and Zhao, 2005; Y. Liang, Lu, and Ying, 2009; L. Sun, Song, and Zhou, 2011).

Table 6 presented several nested models that have been tested to select the best fitting model, the likelihood ratio test statistic, and the p-value for each model.

Table 6
Model Selection Criteria for the Gaussian-Exponential Model

Model	AIC	AICc	BIC	L R Test	P-value
rtumor = trt + number + size	698.32	702.66	727.64	4.40	0.19
rtumor = trt + number	696.60	700.22	723.47	0.28	0.59
rtumor = trt + size	697.24	700.86	724.11	0.96	0.33
rtumor = trt	695.78	698.75	720.21	1.45	0.48

rtumor: number of tumors, trt: treatments, number: initial number of tumors, and size: size of the largest initial tumor.

Each model is compared to the first model as the full model, while the first one is compared to intercept model only. The best fitting model according to the information criteria, AIC, AICc, and BIC, is the one with the following predictors: treatment, prior outcome, prior time-dependent covariate, current time-dependent covariate, and current time. The final model equation given as,

$$y_{ij} = \beta_0 + \beta_1(\textit{treatment}) + \phi(\textit{prior outcome}) + \eta(\textit{prior time-dependent covariate}) + \psi(\textit{current time-dependent covariate}) + \gamma(\textit{current time}) + \alpha + \delta + \epsilon_{ij} \quad (4.8)$$

This model's output in Appendix A shows that the independent variable treatment has a non-significant effect on the number of tumors (β_1= 0.1481 with p-value of 0.07). However, the prior outcome (ϕ = -0.2859 with p-value of 0.0003), prior time-dependent covariate (η = 0.78 with p-value of 0.0001), current time-dependent

covariate (ψ = 0.7090 with p-value of 0.0001), and current time (γ= -0.0125 with p-value of 0.0116) have a significant effect on the number of tumors.This application shows how the informative time and time-dependent covariate are important variables that can be clearly affect the longitudinal outcomes. Finally, the researcher found a similar result in this analysis using the proposed model on the same data that have been studied by J. Sun et al. (2005); L. Sun et al. (2011). Appendix A presented the full output for each model in Table 6 with the information criteria and the likelihood ratio test.

CHAPTER V

CONCLUSIONS

In traditional longitudinal methods, it is often assumed that observation time is fixed and not informative and treat the time-vary covariates as time-independent covariates. Such assumptions, however, is often violated in practice. Since one of the longitudinal study advantage, is the possibility of gathering longitudinal response and covariates information in each time point and determine if the changes in covariates affect or related to the changes in the response. There are few studies cover this type of situation, where the time is informative, the time is irregular, and there is the time-vary covariates via latent variables.

In this dissertation, a joint model was developed to jointly model a longitudinal outcome with informative time and time-dependent covariates. The distribution for the three processes are normal for the outcome that contained 10 parameters, normal for the time-dependent covariate that contained 5 parameters, and exponential for the informative time that contained 2 parameters. The parameters were estimated using the maximum likelihood method for all three processes of the joint model. There are 6 different parameter schemes to observe how the results change with different parameter values and two different design structure (balanced and unbalanced).

The findings showed that the multivariate normality test of the maximum likelihood estimators was met as the number of observations became larger. Since the maximum likelihood estimators showed normality, the likelihood ratio test can be used for model comparisons. The criteria used to compare the models were AIC, AICc, and BIC. Moreover, an R program was created to handle the developed joint model. For illustration and evaluation of the model, the bladder cancer data were analyzed using R.

Monte Carlo simulations were used to evaluate the performance of the joint model. Chapter IV presented the simulation analyses and results. R- 3.2.5 was utilized to run the simulation. Based on the simulation, the asymptotic multivariate normality test was applied in both balanced and unbalanced designs. Moreover, the multivariate normality was tested on specific designs with specific numbers of variables which are three categorical variables with three levels for each and three continues variables. Furthermore, the multivariate normality test of the maximum likelihood estimators was met mostly when the sample size was 54 subjects and above. Therefore, the likelihood ratio test and the model selection criteria can be applied to the developed joint model when the assumption of the normality is met. Thus, it is recommended to apply the joint model with informative time and time-dependent covariate when the sample size is greater than 54 subjects and with the specified number of variables even though the concept should be generalizable with different number of variables.

The maximum likelihood estimator was obtained by R, where the log likelihood function contains four terms. The time-dependent covariate term was

assumed to be conditionally dependent on the prior time-dependent covariate. Whereas, the time-independent covariate was as part of the outcome term. The outcome term was assumed to be conditionally dependent on the prior outcome, the prior time-dependent covariate, current time, and covariates. Finally, the time term was assumed to be conditionally dependent on the prior outcome.

In the real data example, the joint model shows that the informative time and time-dependent covariate are highly significate. These two variables appear to be important components in the model that can give better decision about the response. The researcher found that the results of the application provide similar results that other researchers found using different methods. Thus, the proposed method can add some contribution to the analysis of the longitudinal data.

Limitations and Suggestions for Future Research

The presented joint model with informative time and time-dependent covariates was tested under certain conditions of simulation schemes. In addition, it is assumed in this study that the response variable depends only on the past response as well as the current time and some time-independent covariates, and the prior time-dependent covariate. Moreover, the model was built based on the assumption of independence between time, the time-dependent covariate, and time-independent covariates. In addition, the time-dependent covariate was assumed to be as an external covariate. Furthermore, it is assumed that time is informative and follows an exponential distribution. In conclusion, the presented model can be applied when these assumptions are met and it has not tested for different conditions.

In the proposed model, the asymptotic multivariate normality of the parameter estimates was shown through simulation without digging into the mathematical theories behind it which can be verified in future research. The joint model examined in this study has several assumptions as discussed before which can be relaxed in the future to extend the model. One of the assumption is that the current response is conditionally dependent on the past response value as well as the past time-dependent covariate value. Depending on the research interests this assumption can be extended to not only the most recent prior outcome and time-dependent covariate but to more steps back of these two variables. Additionally, the distribution of time is assumed to be exponential in the current study and that distribution can be different based on the research design and the parameter estimates then can be obtained based on that distribution. Moreover, time and the independent variables are assumed to be independent of each other. In future research, one may incorporate some additional term to express the relationship between them if they are dependent. Also, this study considers the outcome variable to follow a normal distribution, a further extension can be pursued with other distributions such as Bernoulli, Poisson, and Gamma. In addition, the researcher treats the time-dependent covariate as an exogenous covariate, and following normal distribution, future research may use different distribution and as an endogenous time-dependent covariate. Furthermore, although the proposed joint model is examined in only a single outcome variable, this joint model should be able to handle multiple responses by taking the correlation among them into account.

Finally, prediction of future values is recommended as it is an important objective of regression model besides the estimation and the testing.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University press.

Bronsert, M. R. (2009). *A joint model of a longitudinal process and informative time schedule data* (Unpublished doctoral dissertation). University of Northern Colorado.

Cai, N., Lu, W., and Zhang, H. H. (2012). Time-varying latent effect model for longitudinal data with informative observation times. *Biometrics*, *68*(4), 1093–1102.

Chen, Q., May, R. C., Ibrahim, J. G., Chu, H., and Cole, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in medicine*, *33*(26), 4560–4576.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York, NY: Springer.

Dawber, T. R. (1980). *The framingham study: the epidemiology of atherosclerotic disease*. Cambridge, Mass: Harvard University Press.

De Leeuw, J., and Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, *11*(1), 57–85.

Diggle, P. (2002). *Analysis of longitudinal data* (2nd ed., Vol. 25.). Oxford, NY: Oxford University Press.

Fitzmaurice, G. M. (2008). *Longitudinal data analysis.* Boca Raton, FL: Chapman and Hall/CRC.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitudinal analysis.* Hoboken, N.J: Wiley-Interscience.

Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(4), 691–704.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed., Vol. 3.). London: Hodder Arnold.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*(358), 320–338.

Hedeker, D. R., and Gibbons, R. D. (2006). *Longitudinal data analysis.* Hoboken, N.J: Wiley-Interscience.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480.

Huang, C.-Y., Wang, M.-C., and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika*, *93*(4), 763–775.

Jennrich, R. I., and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*(4), 805-820.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lalonde, T. L., Nguyen, A. Q., Yin, J., Irimata, K., and Wilson, J. R. (2013). Modeling correlated binary outcomes with time-dependent covariates. *Journal of Data Science*, *11*, 715–38.

Lee, Y., and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(4), 619-678.

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Liang, Y., Lu, W., and Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*, *65*(2), 377–384.

Lin, Y.-K. (2011). *Hypothesis testing for the gaussian-exponential longitudinal model* (Unpublished doctoral dissertation). University of Northern Colorado.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, *58*(3), 621–630.

Long, M. T., and Fox, C. S. (2016). The framingham heart study 67 years of discovery in metabolic disease. *Nature Reviews Endocrinology*, *12*(3), 177–183.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*(4), 817–827.

Mecklin, C. J., and Mundfrom, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, *72*(1), 123–138.

Nash, J. C., Varadhan, R., et al. (2011). Unifying optimization algorithms to aid software system users: optimx for r. *Journal of Statistical Software*, *43*(9), 1–14.

Neuhaus, J. D., John M and Kalbfleisch. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, *54*(2), 638–645.

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Pinheiro, J., and Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Posada, D., and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, *53*(5), 793–808.

Qiu, F., Stein, C. M., Elston, R. C., (TBRU), T. R. U., and for the Tuberculosis Research Unit (TBRU). (2016). Joint modeling of longitudinal data and discrete-time survival outcome. *Statistical Methods in Medical Research*, *25*(4), 1512-1526.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed., Vol. 1.). Thousand Oaks, CA: Sage Publications.

Rohde, C. A. (1991). *Introduction to probability and mathematical statistics*. Boston, MA: Duxbury.

Rohde, C. A. (2014). *Introductory statistical inference with the likelihood function* (2014th ed.). Cham: Springer. doi: 10.1007/978-3-319-10461-4

Ryu, D., Sinha, D., Mallick, B., Lipsitz, S. R., and Lipshultz, S. E. (2007). Longitudinal studies with outcome-dependent follow-up: Models and bayesian regression. *Journal of the American Statistical Association*, *102*(479), 952-961.

Schmee, J., Andrews, D. F., and Herzberg, A. M. (1987). Data: A collection of problems from many fields for the student and research worker. *Technometrics*, *29*(1), 120.

Seo, J. (2015). *Joint models of longitudinal outcomes and informative time* (Unpublished doctoral dissertation). University of Northern Colorado.

Song, X., Mu, X., and Sun, L. (2012). Regression analysis of longitudinal data with time-dependent covariates and informative observation times. *Scandinavian Journal of Statistics*, *39*(2), 248–258.

Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

Sullivan Pepe, M., and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation*, *23*(4), 939–951.

Sun, J., Park, D.-H., Sun, L., and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*, *100*(471), 882–889.

Sun, J., Sun, L., and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association*, *102*(480), 1397–1406.

Sun, L., Song, X., and Zhou, J. (2011). Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times. *Journal of Statistical Planning and Inference*, *141*(8), 2902–2919.

Twisk, J. W. (2003). *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge, NY: Cambridge University Press.

Wu, L. (2010). *Mixed effects models for complex data*. Boca Raton, FL: Taylor and Francis Group, LLC.

APPENDIX A

OUTPUT OF THE GAUSSIAN-EXPONENTIAL MODEL

Model 1: rtumor = treatment+number+size

Coefficients	Estimate	Std..error	t.value	Pr...t.
Intercept	0.1385	0.1726	0.8025	0.4223
treatment	0.1768	0.0857	2.0631	0.0391
number	0.0888	0.0925	0.9598	0.3371
size	-0.0548	0.1041	-0.5263	0.5987
prior.outcome	-0.2881	0.0786	-3.6654	0.0002
prior.time-dependent	0.7846	0.0461	17.0245	0.0001
current.time-dependent	0.6897	0.0746	9.2482	0.0001
current.time	-0.0123	0.0050	-2.4861	0.0129
alpha	-2.7533	0.2160	-12.7491	0.0001
delta	0.4267	0.1624	2.6274	0.0086

AIC	AICc	BIC	LogLik	LogLik p-value
698.3283	702.6617	727.6401	4.657	0.1987

Model 2: rtumor = treatment+number

Coefficients	Estimate	Std..error	t.value	Pr...t.
Intercept	0.0769	0.1269	0.6059	0.5446
treatment	0.1738	0.0856	2.0305	0.0423
number	0.0986	0.0906	1.0884	0.2764
prior.outcome	-0.2852	0.0784	-3.6361	0.0003
prior.Time-dependent	0.7846	0.0461	17.0245	0.0001
current.Time-dependent	0.6871	0.0745	9.2278	0.0001
current.Time	-0.0126	0.0049	-2.5367	0.0112
alpha	-2.7533	0.2166	-12.7106	0.0001
delta	0.4267	0.1631	2.6172	0.0089
sigma_y	0.5712	0.0279	20.4450	0.0001
sigma_x	0.7940	0.0504	15.7482	0.0001

AIC	AICc	BIC	LogLik	LogLik p-value
696.6053	700.2217	723.4744	0.2768	0.5988

Model 3: rtumor = treatment+size

Coefficients	Estimate	Std..error	t.value	Pr...t.
Intercept	0.2503	0.1276	1.9616	0.0498
treatment	0.1557	0.0830	1.8758	0.0607
size	-0.0750	0.1021	-0.7341	0.4629
prior.outcome	-0.2897	0.0787	-3.6791	0.0002
prior.Time-dependent	0.7846	0.0461	17.0235	0.0001
current.Time-dependent	0.7096	0.0718	9.8834	0.0001
current.Time	-0.0122	0.0050	-2.4602	0.0139
alpha	-2.7533	0.2160	-12.7473	0.0001
delta	0.4267	0.1626	2.6247	0.0087
sigma_y	0.5721	0.0280	20.4442	0.0001
sigma_x	0.7940	0.0504	15.7482	0.0001

AIC	AICc	BIC	LogLik	LogLik p-value
697.2479	700.8644	724.1171	0.9586	0.3275

Model 4: rtumor = treatment

Coefficients	Estimate	Std..error	t.value	Pr...t.
Intercept	0.1801	0.0846	2.1295	0.0332
treatment	0.1481	0.0825	1.7953	0.0726
prior.outcome	-0.2859	0.0787	-3.6343	0.0003
prior.Time-dependent	0.7846	0.0461	17.0256	0.0001
current.Time-dependent	0.7090	0.0719	9.8634	0.0001
current.Time	-0.0125	0.0050	-2.5238	0.0116
alpha	-2.7533	0.2167	-12.7073	0.0001
delta	0.4267	0.1630	2.6177	0.0089
sigma_y	0.5728	0.0280	20.4450	0.0001
sigma_x	0.7940	0.0504	15.7482	0.0001

AIC	AICc	BIC	LogLik	LogLik p-value
695.7862	698.7591	720.2127	1.4578	0.4824

APPENDIX B

R PROGRAM FOR SIMULATIONS

```
########################################
####    Gaussian-Exponential Model ###
########################################


########################################
####              Packages            ###
########################################

install.packages('MVN')  # hzTest, roystonTest
install.packages('MASS')
install.packages('maxLik') #maxLik
install.packages('AlgDesign') # gen.factorial
install.packages('mefa') # provide rep(dat,times)

library(MVN)
library(MASS)
library(maxLik)
library(AlgDesign)
library(mefa)
```

```
#############################################################################
####                    Parameter Setting                                ###
#############################################################################

parameter = matrix(c(0.4,0.4,0.4,0.4,0.4,0.4,   #1:beta0_y
0.2,0.2,0.2,0.2,0.2,0.2,          #2:beta1_y
0.3,0.3,0.3,0.3,0.3,0.3,          #3:beta2_y
0.1,0.1,0.1,0.1,0.1,0.1,          #4:beta3_y
0.3,0.3,0.3,0.3,0.3,0.3,          #5:beta4_y
0.4,0.4,0.4,0.4,0.4,0.4,          #6:beta5_y
0.9,0.9,0.9,0.9,0.9,0.9,          #7:beta6_y

0.5,0.5,0.5,0.5,0.5,0.5,          #8:beta0_x
0.6,0.6,0.6,0.6,0.6,0.6,          #9:beta1_x
0.2,0.2,0.2,0.2,0.2,0.2,          #10:beta2_x
0.7,0.7,0.7,0.7,0.7,0.7,          #11:beta3_x

0.8,0.8,0.8,0.0,0.0,0.8,          #12:phi
0.7,0.7,0.7,0.0,0.0,0.7,          #13:Eta
0.6,0.6,0.6,0.0,0.0,0.6,          #14:Psi
0.1,0.1,0.1,0.1,0.1,0.1,          #15:gamma
2,1,2,1,2,1,                      #16:alpha
0.01,0.02,0.01,0.02,0.01,0.02,   #17:delta
1,1,2,2,0.5,0.5,             #18:sigma_y
1,1,2,2,0.5,0.5),nrow=6)  #19:sigma_x
```

```
##############################################################################
### create design matrix (Z) with two catandtwo cont vars   ###
##############################################################################

design.z=function(level=c(3,3),m=18,p=2){
catg=gen.factorial(levels=level,center=FALSE,factors='all')
ext=rep(catg,m/(prod(level)))
des=model.matrix(~.,data=ext) #'~.' is supported by {AlgDesign}
cont=data.frame(matrix(NA,nrow=m,ncol=p))
for (i in 1:p){
cont[i]=rnorm(m)
}
zmatrix=as.matrix(cbind(des,cont))
colnames(zmatrix)<-c("Intercept","Z12","Z13","Z22","Z23","Z1","Z2")
zmatrix
}
#design.z()
#design(level=c(3,3),m=18,p=2)
```

```
#########################################################################
### create design matrix (X) with one cont andone cat var   ###
#########################################################################

design.x=function(m=18){
cat_cont <- data.frame(X=rep(c("1","2","3")), Xc=rnorm(m))
Xmatrix<- model.matrix(~X+ Xc , data=cat_cont)
Xmatrix
}
#design.x()

#########################################################################
### Create Data:  c('outcome','TDC', 'time','subject')      ###
#########################################################################
outcome<- function(m=m,num=num,parm=parm){

n1=num[1]
n2=num[2]
ndesign = matrix(c(rep(n1,m/2),rep(n2,m/2)),byrow=T)
nn=cumsum(c(1,ndesign[-length(ndesign)]))
raw= matrix(NA,sum(ndesign),4)          #Null matrix
colnames(raw) <- c("Y","X","T","S")    #set the column names
rownames(raw) <-rownames(raw,do.NULL=FALSE,prefix="Obs.")
mu_y = zmatrix %*% parm[1:7]
mu_x = Xmatrix %*% parm[8:11]
```

```
raw[nn,1]= mu_y + rnorm(m)*parm[18]
raw[nn,2]= mu_x + rnorm(m)*parm[19]
raw[nn,3] = rexp(m)

for (i in 1:m){
for (j in 2:ndesign[i]){

yjmin1 = raw[nn[i] - 1 + j - 1,1] # is first obs for y
xjmin1 = raw[nn[i] - 1 + j - 1,2] # is  first obs for x "TDC"

raw[nn[i] - 1 + j,3]=rexp(1)*exp(parm[16] +parm[17] * yjmin1)
raw[nn[i] - 1 + j,2] =mu_x[i] + xjmin1 * parm[14] +
                      rnorm(1)*parm[19]
raw[nn[i]- 1 + j,1] =mu_y[i] + yjmin1 * parm[12] +
raw[nn[i] - 1 + j,2]*parm[13]+raw[nn[i]-1+j,3]*parm[15]+
                                rnorm(1)*parm[18]
raw[nn[i],4]=i
raw[nn[i]-1+j,4]=i
} #j
}#i
result=list(raw=raw,nn=nn,ndesign=ndesign)
result
} #outcome
#outcome(m=m,num=num,parm=parm)
```

```
#############################################################################
###                    Log-Likelihood Function                            ###
#############################################################################

loglikfn<- function(parms){

y1=y[nn,1] #initial obs for every subjects
x1=y[nn,2]#initial obs of TDC for every subjects

f1=sum(-0.5 * log(parms[18]^2)-0.5* (y1-x1*parms[14]-
zmatrix %*% parms[1:7])^2/parms[18]^2)
f2=0;f3=0 ;f4=0

for (i in 1:m){
yi=y[(y[,4]==i), 1] # all obs for ith subject
xi=y[(y[,4]==i), 2] # all TDC  for ith subject
ti=y[(y[,4]==i), 3] # all time points for ith subject
yi1=yi[-ndesign[i]] #previous obs
xi1=xi[-ndesign[i]] #previous obs of TDC
tti=ti[-1] #current time
yi2=yi[-1] #current obs
xi2=xi[-1] #current obs of TDC

f2=sum(-0.5 * log(parms[18]^2)-0.5*(yi2-parms[15]*tti-
parms[12]*yi1-parms[14]*xi1-
          zmatrix[i,]%*%parms[1:7])^2/parms[18]^2)+f2
f3=sum(parms[16]+parms[17]*yi1-
```

```
                exp(parms[16]+parms[17]*yi1)*tti)+f3
f4=sum(-0.5 * log(parms[19]^2)-0.5 *(xi2-parms[13]*xi1-
                Xmatrix[i,]%*%parms[8:11])^2/parms[19]^2)+f4

} #i
(m+f1+f2+f3+f4)
}  # loglike
##############################################################################
###                          Simulation                                    ###
##############################################################################

Pschem = 1 # parameter setting, 1 to 6
m=54 # Sample size (18,36,54,90,180)
num =matrix(c(10,5))#design structure(10,10),(5,3),(10,5),(20,6)
rep=2000 #number of replications
out = matrix(NA,rep,ncol(parameter))
parm = parameter[Pschem,]
zmatrix=design.z(level=c(3,3),m=m,p=2)
Xmatrix=design.x(m=m)
strt<-Sys.time()
for (r in 1:rep){
#compute some info to be used in optimization
result=outcome(m=m,num=num,parm=parm)
y=result$raw
nn=result$nn
ndesign=result$ndesign
mle=maxLik(logLik = loglikfn, start = parm)
```

```
diff=coef(mle)-parm
out[r,]=sqrt(sum(ndesign))*diff/summary(mle)$estimate[,2]
}
print(Sys.time()-strt)
print(c("TimeR=",TimeR))
print(c("m=",m))
print(c("Pschem=",Pschem))
print(c("rep=",rep))
print(c("num=",num))
summary(mle)
hzTest(out[,c((1:7),12,14,15)])    #Outcome process
hzTest(out[,c((8:11),13)])         #TDC process
hzTest(out[,c(16:17)])             #Time process
```

APPENDIX C

R PROGRAM FOR THE APPLICATION

```
#########################################
####            Packages             ###
#########################################
install.packages("dplyr")
install.packages("MuMIn")
install.packages('MVN')  # hzTest, roystonTest
install.packages('MASS')
install.packages('maxLik') #maxLik
install.packages('AlgDesign') # gen.factorial
install.packages('mefa') # provide rep(dat,times)

library(survival)
library(MuMIn)
library(dplyr)
library(MVN)
library(MASS)
library(maxLik)
library(AlgDesign)
library(mefa)
```

```
#############################################################
##  Read the date after prepare it for the analysis##
#############################################################

y <- read.csv(file="", header=TRUE, sep=",")
y1obs <-y %>% group_by(y[,1]) %>% mutate(rank=row_number())%>% filter(ra
zmatrix<-y1obs[,5:8]
m=85
zmatrix=as.matrix(zmatrix1[,1:4])
zmatrix2=as.matrix(zmatrix1[,c(1,2,3)])
zmatrix3=as.matrix(zmatrix1[,c(1,2,4)])
zmatrix4=as.matrix(zmatrix1[,1:2])
#########################################################################
####                  Parameter Setting                              ###
#########################################################################
parameter = matrix(c(0.4, #1:beta0 inter
0.2, #2:beta1   Xt1
0.3, #3:beta2    Xn2
0.1, #4:beta3   Xs3
0.8  , #5:phi
0.7  , #6:Eta
0.6  , #7:Psi
0.1  , #8:gamma
2    , #9:alpha
0.01  , #10:delta
0.5    , #11:sigma_y
1   ),nrow = 1) #12:sigma_x
```

```
parameter2=matrix(parameter[c(1,2,3,5,6,7,8,9,10,11,12)],nrow=1)
parameter3=matrix(parameter[c(1,2,4,5,6,7,8,9,10,11,12)],nrow=1)
parameter4=matrix(parameter[c(1,2,5,6,7,8,9,10,11,12)],nrow=1)
###################################
#1#  rtumor=treatment+number+size   #
###################################
strt<-Sys.time()
loglikfn<- function(parms){
y1obs <-y %>% group_by(y[,1]) %>%
          mutate(rank=row_number())%>% filter(rank==1)
y1=y1obs[,2]
x1= y1obs[,3]
f1=sum(-0.5 * log(parms[11]^2)-0.5*(y1-x1*parms[7]-
          zmatrix[,c(1,2,3)] %*% parms[1:4])^2/parms[11]^2)
f2=0;f3=0 ;f4=0
for (i in 1:m){
yi=y[(y[,1]==i), 2] # all obs for ith subject
xi=y[(y[,1]==i), 3] # all TDC  for ith subject
ti=y[(y[,1]==i), 4] # all time points for ith subject
yi1=yi[-length(yi)]  #previous obs
xi1=xi[-length(xi)] #previous obs of TDC
tti=ti[-1] #current time
yi2=yi[-1] #current obs
xi2=xi[-1] #current obs of TDC
f2=sum(-0.5 * log(parms[11]^2)-0.5*(yi2-parms[8]*tti-
  parms[5]*yi1- parms[7]*xi1-zmatrix[i,] %*%
    parms[1:4])^2/parms[11]^2)+f2
```

```
f3=sum(parms[9]+parms[10]*yi1-exp(parms[9]+
        parms[10]*yi1)*tti)+f3
f4=sum(-0.5 * log(parms[12]^2)-
          0.5 *(xi2-parms[6]*xi1)^2/parms[12]^2)+f4
} #i
(m+f1+f2+f3+f4)
}  # loglike
loglikfn (parms=parameter)
mle=maxLik(logLik = loglikfn, start = parameter)
summary(mle)
print(Sys.time()-strt)
Coefficients<- c("Intercept","treatment","number", "size",
"prior.outcome", "prior.time-dependent",
"current.time-dependent","current.time",
"alpha","delta","sigma_y","sigma_x")
summary(mle)$estimate
M<-data.frame(Coefficients,summary(mle)$estimate)
parm <- summary(mle)
est <- summary(mle)$estimate
AIC <- AIC(mle)
AICc <- AIC+2*parm$NActivePar*
        (parm$NActivePar+1)/ (m-parm$NActivePar-1)
BIC <- -2*parm$loglik+parm$NActivePar*log(m)
ratio <- 2*(logLik(mle)- logLik(mle_inter))
dffull <- summary(mle)$NActivePar
dfred <- summary(mle_inter)$NActivePar
dfchi <- dffull-dfred
```

```
Pr <- 1-pchisq(ratio,dfchi)
R1=list(Model= rtumor ~ treatment+number+size,Coefficients =
      M, AIC = AIC, AICc = AICc, BIC = BIC,
          LogLik = ratio, LogLikPval = Pr)



          ##################################
          #2# rtumor=treatment+number  #
          ##################################
          strt<-Sys.time()
          loglikfn2<- function(parms){
  y1obs <-y %>% group_by(y[,1]) %>%
  mutate(rank=row_number())%>% filter(rank==1)
  y1=y1obs[,2]
  x1= y1obs[,3]
          f1=sum(-0.5 * log(parms[10]^2)-
          0.5*(y1-x1*parms[6]-zmatrix2 %*%
          parms[1:3])^2/parms[10]^2)
          f2=0;f3=0 ;f4=0
          for (i in 1:m){
          yi=y[(y[,1]==i), 2] # all obs for ith subject
          xi=y[(y[,1]==i), 3] # all TDC  for ith subject
          ti=y[(y[,1]==i), 4] # all time points for ith subject
          yi1=yi[-length(yi)]  #previous obs
          xi1=xi[-length(xi)] #previous obs of TDC
          tti=ti[-1] #current time
```

```
yi2=yi[-1] #current obs
xi2=xi[-1] #current obs of TDC
f2=sum(-0.5 * log(parms[10]^2)-
0.5*(yi2-parms[7]*tti-parms[4]*yi1-parms[6]*
xi1-zmatrix2[i,] %*% parms[1:3])^2/parms[10]^2)+f2
f3=sum(parms[8]+parms[9]*
yi1-exp(parms[8]+
parms[9]*yi1)*tti)+f3
f4=sum(-0.5 * log(parms[11]^2)-
0.5 *(xi2-parms[5]*xi1)^2/parms[11]^2)+f4
} #i
(m+f1+f2+f3+f4)
}  # loglike
loglikfn2 (parms=parameter2)
mle2=maxLik(logLik = loglikfn2, start = parameter2)
summary(mle2)
print(Sys.time()-strt)
Coefficients<- c("Intercept","treatment","number",
"prior.outcome", "prior.Time-dependent",
"current.Time-dependent",
"current.Time","alpha","delta","sigma_y","sigma_x")
summary(mle2)$estimate
M2<-data.frame(Coefficients,summary(mle2)$estimate)
parm <- summary(mle2)
est <- summary(mle2)$estimate
AIC <- AIC(mle2)
AICc <- AIC+2*parm$NActivePar*
```

```
(parm$NActivePar+1)/ (m-parm$NActivePar-1)
BIC <- -2*parm$loglik+parm$NActivePar*log(m)
ratio <- 2*(logLik(mle)- logLik(mle2))
dffull <- summary(mle)$NActivePar
dfred <- summary(mle2)$NActivePar
dfchi <- dffull-dfred
Pr <- 1-pchisq(ratio,dfchi)
R2=list(Model= rtumor ~ treatment+number, Coefficients =
M2, AIC = AIC, AICc = AICc, BIC = BIC,
LogLik = ratio, LogLikPval = Pr)
############################
#3# rtumor=treatment+size    #
############################
strt<-Sys.time()
loglikfn3<- function(parms){
y1obs <-y %>% group_by(y[,1]) %>%
mutate(rank=row_number())%>% filter(rank==1)
y1=y1obs[,2]
x1= y1obs[,3]
f1=sum(-0.5 * log(parms[10]^2)-
0.5*(y1-x1*parms[6]-
zmatrix3 %*% parms[1:3])^2/parms[10]^2)
f2=0;f3=0 ;f4=0
for (i in 1:m){
yi=y[(y[,1]==i), 2] # all obs for ith subject
xi=y[(y[,1]==i), 3] # all TDC  for ith subject
ti=y[(y[,1]==i), 4] # all time points for ith subject
```

```
yi1=yi[-length(yi)]  #previous obs
xi1=xi[-length(xi)] #previous obs of TDC
tti=ti[-1] #current time
yi2=yi[-1] #current obs
xi2=xi[-1] #current obs of TDC
f2=sum(-0.5 * log(parms[10]^2)-
0.5*(yi2-parms[7]*tti-parms[4]*yi1-parms[6]*xi1-
zmatrix3[i,] %*% parms[1:3])^2/parms[10]^2)+f2
f3=sum(parms[8]+parms[9]*yi1-
exp(parms[8]+parms[9]*yi1)*tti)+f3
f4=sum(-0.5 * log(parms[11]^2)-
0.5 *(xi2-parms[5]*xi1)^2/parms[11]^2)+f4
} #i
(m+f1+f2+f3+f4)
}  # loglike
loglikfn3 (parms=parameter3)
mle3=maxLik(logLik = loglikfn3, start = parameter3)
summary(mle3)
Coefficients<- c("Intercept","treatment", "size",
"prior.outcome","prior.Time-dependent",
"current.Time-dependent",
 "current.Time","alpha","delta","sigma_y","sigma_x")
summary(mle3)$estimate
M3<-data.frame(Coefficients,summary(mle3)$estimate)
      parm <- summary(mle3)
      est <- summary(mle3)$estimate
      AIC <- AIC(mle3)
```

```
AICc <- AIC+2*parm$NActivePar*
(parm$NActivePar+1)/ (m-parm$NActivePar-1)
BIC <- -2*parm$loglik+parm$NActivePar*log(m)
ratio <- 2*(logLik(mle)- logLik(mle3))
dffull <- summary(mle)$NActivePar
dfred <- summary(mle3)$NActivePar
dfchi <- dffull-dfred
Pr <- 1-pchisq(ratio,dfchi)
R2=list(Model= rtumor ~ treatment+size, Coefficients =
M3, AIC = AIC, AICc = AICc, BIC = BIC,
LogLik = ratio, LogLikPval = Pr)
#########################
#4# rtumor=treatment      #
#########################
strt<-Sys.time()
loglikfn4<- function(parms){
y1obs <-y %>% group_by(y[,1]) %>%
mutate(rank=row_number())%>% filter(rank==1)
y1=y1obs[,2]
x1= y1obs[,3]
f1=sum(-0.5 * log(parms[9]^2)-
0.5*(y1-x1*parms[5]-
zmatrix4 %*% parms[1:2])^2/parms[9]^2)
f2=0;f3=0 ;f4=0
for (i in 1:m){
yi=y[(y[,1]==i), 2] # all obs for ith subject
xi=y[(y[,1]==i), 3] # all TDC  for ith subject
```

```
        ti=y[(y[,1]==i), 4] # all time points for ith subject
        yi1=yi[-length(yi)]  #previous obs
        xi1=xi[-length(xi)] #previous obs of TDC
        tti=ti[-1] #current time
        yi2=yi[-1] #current obs
        xi2=xi[-1] #current obs of TDC
        f2=sum(-0.5 * log(parms[9]^2)-
        0.5*(yi2-parms[6]*tti-parms[3]*yi1-
        parms[5]*xi1-zmatrix4[i,] %*%
        parms[1:2])^2/parms[9]^2)+f2
        f3=sum(parms[7]+parms[8]*yi1-
        exp(parms[7]+
        parms[8]*yi1)*tti)+f3
        f4=sum(-0.5 * log(parms[10]^2)-
        0.5 *(xi2-parms[4]*xi1)^2/
        parms[10]^2)+f4
        } #i
        (m+f1+f2+f3+f4)
        }  # loglike
        loglikfn4 (parms=parameter4)
        mle4=maxLik(logLik = loglikfn3, start = parameter4)
        summary(mle4)
        print(Sys.time()-strt)
Coefficients<- c("Intercept","treatment",
"prior.outcome", "prior.Time-dependent",
"current.Time-dependent",
"current.Time","alpha","delta","sigma_y","sigma_x")
```

```
summary(mle4)$estimate
M4<-data.frame(Coefficients,summary(mle4)$estimate)
      parm <- summary(mle4)
      est <- summary(mle4)$estimate
      AIC <- AIC(mle4)
      AICc <- AIC+2*parm$NActivePar*
      (parm$NActivePar+1)/ (m-parm$NActivePar-1)
      BIC <- -2*parm$loglik+parm$NActivePar*log(m)
      ratio <- 2*(logLik(mle)- logLik(mle4))
      dffull <- summary(mle)$NActivePar
      dfred <- summary(mle4)$NActivePar
      dfchi <- dffull-dfred
      Pr <- 1-pchisq(ratio,dfchi)
      R4=list(Model= rtumor ~ treatment, Coefficients =
      M4, AIC = AIC, AICc = AICc, BIC = BIC,
      LogLik = ratio, LogLikPval = Pr)
            ##################
            # Results  #
            ##################
            R1
            R2
            R3
            R4
```

APPENDIX D

BLADDER CANCER RECURRENCES DATA SET

Bladder1 Cancer Data

This analysis uses "rtumor" as a outcome, "treatment" as a grouping variable, the total number of tumors within last 6 months as a time-dependent covariate, and "number" and "size" as covariates. This dataset consists of 118 patients. However, the researcher consider two groups in the analysis so the total of patients becomes 85 .

1. id: Patient

2. treatment: Placebo,or thiotepa

3. number: Initial number of tumors (8 = 8 or more)

4. size: Size (cm) of the largest initial tumor

5. recur: Number of recurrences

6. start: The start time of each time interval

7. stop: The end time of each time interval

8. status: End of interval code, 0 = censored, 1 = recurrence, 2 = death from bladder disease, 3 = death other/unknown cause

9. rtumor: Number of tumors found at the time of a recurrence

10. rsize: Size of largest tumor at a recurrence

id treatment number size recur start stop status rtumor rsize enum

1 1 1 1 0 0 0 3 . . 1

2 1 1 3 0 0 1 3 . . 1

3 1 2 1 0 0 4 0 . . 1

4 1 1 1 0 0 7 0 . . 1

5 1 5 1 0 0 10 3 . . 1

6 1 4 1 1 0 6 1 1 1 1

6 1 4 1 1 6 10 3 . . 2

7 1 1 1 0 0 14 0 . . 1

8 1 1 1 0 0 18 0 . . 1

9 1 1 3 1 0 5 1 2 4 1

9 1 1 3 1 5 18 3 . . 2

10 1 1 1 2 0 12 1 2 2 1

10 1 1 1 2 12 16 1 3 . 2

10 1 1 1 2 16 18 3 . . 3

11 1 3 3 0 0 23 0 . . 1

12 1 1 3 2 0 10 1 6 1 1

12 1 1 3 2 10 15 1 3 1 2

12 1 1 3 2 15 23 0 . . 3

13 1 1 1 3 0 3 1 8 1 1

13 1 1 1 3 3 16 1 8 . 2

13 1 1 1 3 16 23 1 8 . 3

14 1 3 1 3 0 3 1 1 1 1

14 1 3 1 3 3 9 1 1 2 2

14 1 3 1 3 9 21 1 8 8 3

14 1 3 1 3 21 23 2 . . 4

15 1 2 3 4 0 7 1 8 2 1

15 1 2 3 4 7 10 1 7 1 2

15 1 2 3 4 10 16 1 5 3 3

15 1 2 3 4 16 24 1 7 . 4

16 1 1 1 3 0 3 1 1 1 1

16 1 1 1 3 3 15 1 1 . 2

16 1 1 1 3 15 25 1 3 . 3

17 1 1 2 0 0 26 0 . . 1

18 1 8 1 1 0 1 1 8 1 1

18 1 8 1 1 1 26 0 . . 2

19 1 1 4 2 0 2 1 4 1 1

19 1 1 4 2 2 26 1 8 . 2

20 1 1 2 1 0 25 1 3 . 1

20 1 1 2 1 25 28 0 . . 2

21 1 1 4 0 0 29 0 . . 1

22 1 1 2 0 0 29 0 . . 1

23 1 4 1 0 0 29 3 . . 1

24 1 1 6 2 0 28 1 2 1 1

24 1 1 6 2 28 30 1 1 1 2

25 1 1 5 3 0 2 1 4 1 1

25 1 1 5 3 2 17 1 2 1 2

25 1 1 5 3 17 22 1 4 . 3

25 1 1 5 3 22 30 0 . . 4

26 1 2 1 5 0 3 1 1 . 1

26 1 2 1 5 3 6 1 3 . 2

26 1 2 1 5 6 8 1 3 . 3

26 1 2 1 5 8 12 1 3 . 4

26 1 2 1 5 12 26 1 3 . 5

26 1 2 1 5 26 30 3 . . 6

27 1 1 3 3 0 12 1 2 . 1

27 1 1 3 3 12 15 1 3 1 2

27 1 1 3 3 15 24 1 1 . 3

27 1 1 3 3 24 31 0 . . 4

28 1 1 2 0 0 32 0 . . 1

29 1 2 1 0 0 34 3 . . 1

30 1 2 1 0 0 36 0 . . 1

31 1 3 1 1 0 29 1 8 1 1

31 1 3 1 1 29 36 0 . . 2

32 1 1 2 0 0 37 0 . . 1

33 1 4 1 4 0 9 1 8 1 1

33 1 4 1 4 9 17 1 2 1 2

33 1 4 1 4 17 22 1 5 . 3

33 1 4 1 4 22 24 1 1 . 4

33 1 4 1 4 24 40 0 . . 5

34 1 5 1 6 0 16 1 1 1 1

34 1 5 1 6 16 19 1 8 1 2

34 1 5 1 6 19 23 1 1 1 3

34 1 5 1 6 23 29 1 2 1 4

34 1 5 1 6 29 34 1 1 1 5

34 1 5 1 6 34 40 1 3 . 6

35 1 1 2 0 0 41 0 . . 1

36 1 1 1 1 0 3 1 3 1 1

36 1 1 1 1 3 43 0 . . 2

37 1 2 6 1 0 6 1 1 1 1

37 1 2 6 1 6 43 0 . . 2

38 1 2 1 3 0 3 1 5 1 1

38 1 2 1 3 3 6 1 3 1 2

38 1 2 1 3 6 9 1 4 1 3

38 1 2 1 3 9 44 0 . . 4

39 1 1 1 5 0 9 1 1 1 1

39 1 1 1 5 9 11 1 3 1 2

39 1 1 1 5 11 20 1 1 1 3

39 1 1 1 5 20 26 1 4 1 4

39 1 1 1 5 26 30 1 3 1 5

39 1 1 1 5 30 45 3 . . 6

40 1 1 1 1 0 18 1 1 1 1

40 1 1 1 1 18 48 0 . . 2

41 1 1 3 0 0 49 0 . . 1

42 1 3 1 1 0 35 1 1 1 1

42 1 3 1 1 35 51 0 . . 2

43 1 1 7 1 0 17 1 1 1 1

43 1 1 7 1 17 53 0 . . 2

44 1 3 1 5 0 3 1 7 1 1

44 1 3 1 5 3 15 1 2 1 2

44 1 3 1 5 15 46 1 3 . 3

44 1 3 1 5 46 51 1 2 . 4

44 1 3 1 5 51 53 1 1 1 5

45 1 1 1 0 0 59 0 . . 1

46 1 3 2 9 0 2 1 1 3 1

46 1 3 2 9 2 15 1 3 1 2

46 1 3 2 9 15 24 1 4 1 3

46 1 3 2 9 24 30 1 3 2 4

46 1 3 2 9 30 34 1 4 1 5

46 1 3 2 9 34 39 1 1 . 6

46 1 3 2 9 39 43 1 1 . 7

46 1 3 2 9 43 49 1 1 . 8

46 1 3 2 9 49 52 1 1 . 9

46 1 3 2 9 52 61 0 . . 10

47 1 1 3 5 0 5 1 3 1 1

47 1 1 3 5 5 14 1 4 1 2

47 1 1 3 5 14 19 1 2 1 3

47 1 1 3 5 19 27 1 5 1 4

47 1 1 3 5 27 41 1 . . 5

47 1 1 3 5 41 64 0 . . 6

48 1 2 3 8 0 2 1 1 1 1

48 1 2 3 8 2 8 1 3 1 2

48 1 2 3 8 8 12 1 6 1 3

48 1 2 3 8 12 13 1 2 1 4

48 1 2 3 8 13 17 1 2 1 5

48 1 2 3 8 17 21 1 1 1 6

48 1 2 3 8 21 33 1 1 1 7

48 1 2 3 8 33 49 1 1 . 8

48 1 2 3 8 49 64 0 . . 9

49 0 1 3 0 0 1 0 . . 1

50 0 1 1 0 0 1 3 . . 1

51 0 8 1 1 0 5 1 8 1 1

52 0 1 2 0 0 9 0 . . 1

53 0 1 1 0 0 10 3 . . 1

54 0 1 1 0 0 13 0 . . 1

55 0 2 6 1 0 3 1 1 1 1

55 0 2 6 1 3 14 0 . . 2

56 0 5 3 5 0 1 1 5 2 1

56 0 5 3 5 1 3 1 2 1 2

56 0 5 3 5 3 5 1 5 1 3

56 0 5 3 5 5 7 1 2 1 4

56 0 5 3 5 7 10 1 2 1 5

56 0 5 3 5 10 17 3 . . 6

57 0 5 1 0 0 18 3 . . 1

58 0 1 3 1 0 17 1 2 1 1

58 0 1 3 1 17 18 3 . . 2

59 0 5 1 1 0 2 1 2 1 1

59 0 5 1 1 2 19 2 . . 2

60 0 1 1 2 0 17 1 1 1 1

60 0 1 1 2 17 19 1 1 1 2

60 0 1 1 2 19 21 3 . . 3

61 0 1 1 0 0 22 0 . . 1

62 0 1 3 0 0 25 0 . . 1

63 0 1 5 0 0 25 0 . . 1

64 0 1 1 0 0 25 0 . . 1

65 0 1 1 3 0 6 1 2 1 1

65 0 1 1 3 6 12 1 3 1 2

65 0 1 1 3 12 13 1 1 . 3

65 0 1 1 3 13 26 0 . . 4

66 0 1 1 1 0 6 1 1 1 1

66 0 1 1 1 6 27 0 . . 2

67 0 2 1 1 0 2 1 2 . 1

67 0 2 1 1 2 29 0 . . 2

68 0 8 3 2 0 26 1 3 . 1

68 0 8 3 2 26 35 1 3 . 2

68 0 8 3 2 35 36 0 . . 3

69 0 1 1 0 0 38 0 . . 1

70 0 1 1 4 0 22 1 2 1 1

70 0 1 1 4 22 23 1 1 1 2

70 0 1 1 4 23 27 1 2 1 3

70 0 1 1 4 27 32 1 3 . 4

70 0 1 1 4 32 39 0 . . 5

71 0 6 1 7 0 4 1 1 1 1

71 0 6 1 7 4 16 1 3 1 2

71 0 6 1 7 16 23 1 3 1 3

71 0 6 1 7 23 27 1 3 1 4

71 0 6 1 7 27 33 1 8 . 5

71 0 6 1 7 33 36 1 . . 6

71 0 6 1 7 36 37 1 8 1 7

71 0 6 1 7 37 39 3 . . 8

72 0 3 1 4 0 24 1 3 1 1

72 0 3 1 4 24 26 1 2 . 2

72 0 3 1 4 26 29 1 1 . 3

72 0 3 1 4 29 40 1 2 . 4

73 0 3 2 0 0 41 0 . . 1

74 0 1 1 0 0 41 3 . . 1

75 0 1 1 2 0 1 1 1 2 1

75 0 1 1 2 1 27 1 1 . 2

75 0 1 1 2 27 43 0 . . 3

76 0 1 1 0 0 44 0 . . 1

77 0 6 1 5 0 2 1 2 1 1

77 0 6 1 5 2 20 1 1 1 2

77 0 6 1 5 20 23 1 2 1 3

77 0 6 1 5 23 27 1 1 1 4

77 0 6 1 5 27 38 1 8 . 5

77 0 6 1 5 38 44 0 . . 6

78 0 1 2 0 0 45 0 . . 1

79 0 1 4 1 0 2 1 1 1 1

79 0 1 4 1 2 46 0 . . 2

79 0 1 4 0 0 46 3 . . 1

80 0 3 3 0 0 49 0 . . 1

81 0 1 1 0 0 50 0 . . 1

82 0 4 1 3 0 4 1 1 1 1

82 0 4 1 3 4 24 1 1 1 2

82 0 4 1 3 24 47 1 1 . 3

82 0 4 1 3 47 50 0 . . 4

83 0 3 4 0 0 54 0 . . 1

84 0 2 1 1 0 38 1 2 1 1

84 0 2 1 1 38 54 0 . . 2

85 0 1 3 0 0 59 3 . . 1