ANALYSIS OF NON-IGNORABLE MISSING AND LEFT-CENSORED LONGITUDINAL BIOMARKER DATA

by

Abdus Sattar

B.Sc. in Statistics, Jahangirnagar University, Bangladash, 1992

M.Sc. in Statistics, Jahangirnagar University, Bangladesh, 1993

M.S. in Statistics, Texas A& M University, 2003

M.S. in Mathematics, Texas A&M University, 2004

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Abdus Sattar

It was defended on

Sept 29, 2009

and approved by

Lisa Weissfeld, Ph.D., Professor & Associate Chair, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Chung-Chou Ho Chang, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Jong-Hyeon Jeong, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Lan Kong, Ph.D., Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Mark Unruh, M.D., Assistant Professor, Renal Electrolyte Division, Department of medicine, School of Medicine, University of Pittsburgh

Dissertation Director: Lisa Weissfeld, Ph.D., Professor & Associate Chair, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

ANALYSIS OF NON-IGNORABLE MISSING AND LEFT-CENSORED LONGITUDINAL BIOMARKER DATA

Abdus Sattar, PhD

University of Pittsburgh, 2009

In a longitudinal study of biomarker data collected during a hospital stay, observations may be missing due to administrative reasons, the death of the subject or the subject's discharge from the hospital, resulting in non-ignorable missing data. Standard likelihood-based methods for the analysis of longitudinal data, e.g, mixed models, do not include a mechanism that accounts for the different reasons for missingness. Rather than specifying a full likelihood function for the observed and missing data, we have proposed a weighted pseudo likelihood (WPL) method. Using this method a model can be built based on available data by accounting for the unobserved data via weights which are then treated as nuisance parameters in the model. The WPL method accounts for the nuisance parameters in the computation of the variances of parameter estimates. The performance of the proposed method has been compared with a number of widely used methods. The WPL method is illustrated using an example from the Genetic and Inflammatory Marker of Sepsis (GenIMS) study. A simulation study has been conducted to study the properties of the proposed method and the results are competitive with the widely used methods.

In the second part, our goal is to address the problem of analyzing left-censored longitudinally measured biomarker data when subjects are lost due to the above mentioned reasons. We propose to analyze one such biomarker, IL-6, obtained from the GenIMS study, using a weighted random effects Tobit (WRT) model. We have compared the results of the WRT model with the random effects Tobit model. The simulation study shows that the WRT model estimates are approximately unbiased. The correct standard error has been computed using asymptotic pseudo likelihood theory. The use of multiple weights across the panel improves the estimate and produces smaller root mean square error. Therefore, the WRT model with multiple weights across panels is the recommended model for analyzing non-ignorable missing and left-censored biomarker longitudinal data.

Model selection is an extremely important part of the analysis of any data set. As illustrated in these analyses, conclusions, which can directly impact public health, depend heavily on the data analytic approach.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The primary goal of a longitudinal study is to characterize the change in the response variable during the study period and to measure the effects of factors on the response variable. But many longitudinal studies suffer from the problem of missing data. The presence of missing data has many implications in longitudinal data analysis including unbalanced design, loss of information, bias and hence misleading inferences about the change in the mean response (Fitzmaurice et al., 2004). To obtain a valid inference, the reasons for any missing data, known as the *missing data mechanism*, must be considered with proper care. The missing data mechanism is said to be missing completely at random (MCAR) if the probability of missing a response is unrelated to either the unobserved or the observed responses. If the probability of missing a response depends on the observed response values but does not depend on the unobserved response values then it is called missing at random (MAR). The missing data mechanisms, MCAR and MAR, are also known as *ignorable* because it is not necessary to model the missing data process as a part of the likelihood based analysis. The missing data mechanism is said to be missing not at random (MNAR) if the probability of missing is associated with the unobserved response values that should have been obtained. This process is often referred to as *non-ignorable* missingness due to the fact that the missing data mechanism must be considered to make a valid inference about the distribution of the responses (Little and Rubin, 2002, Fitzmaurice et al, 2004, Allison, 2002). In a longitudinal study the term dropout refers to the situation where a response at a particular time being missing, implies that all the subsequent follow-up responses are also missing (Fitzmaurice et al, 2004, Little and Rubin, 2002). In the Genetic and Inflammatory Markers of Sepsis (GenIMS) study (Kellum, et al. 2007), biomarkers were measured daily on many of the hospitalized subjects for a period of one week or longer. In the GenIMS study (details in

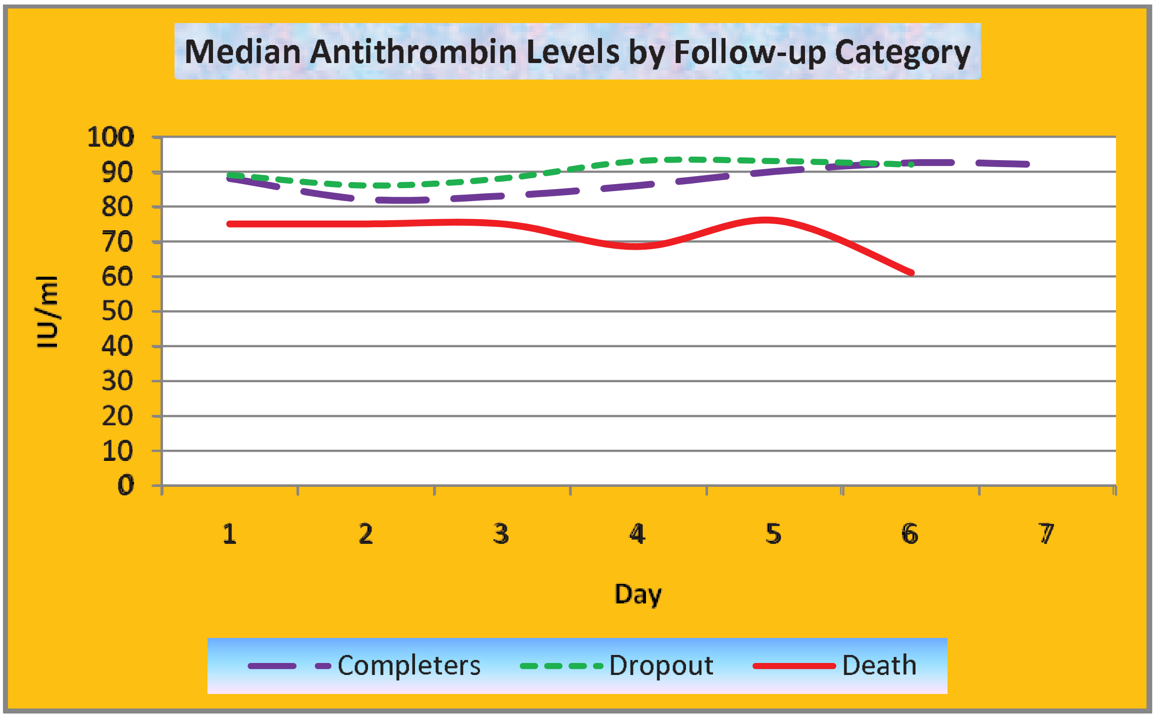

Figure 1.1: Median Antithrombin levels for a period of seven days by the follow-up category.

Chapter 3), the data are missing due to death, discharge from the hospital and administrative reasons. Figure 1.1 presents the median levels and trend of a biomarker, anti-thrombin, over the first seven days. Anti-thrombin is a small serum protein that interfere with coagulation cascade in the blood and the deficiency of this protein is associated with increasing risk of developing blood clot. The subjects who have dropped-out (669 subjects) during the study period had a higher level of anti-thrombin, followed by completers (330) and deaths (19). This figure illustrates that dropout and death led to substantial loss of information.

There are a number of approaches for analyzing longitudinal data with different types of missingness. If the missingness is MCAR then analyzing only the complete subject's information is known as a complete-case analysis. Due to the removal of subjects from the study, the reduced sample size often leads to inefficient estimates and reduced statistical power. A similar approach for handling missing longitudinal data is the analysis of available data. Though this approach covers more data when compared to the complete case analysis, statistical methods used for analyzing available data will produce biased estimates unless the the missingness is MCAR. A traditional method for handling missing data is the imputation method in which the missing data are replaced by the imputed data and standard statistical methods are then used for analyzing the full data set. The issue with the imputation method is how to obtain a valid data set for the missing data. An alternative approach for handling missing longitudinal data is to weight observed data (Robins et al., 1995, Fitzmaurice, et al., 2004, Demirtas, 2004, Dufouil et al., 2004). The attraction of these methods is that once the data set has been constructed, standard methods for analyzing longitudinal data such as weighted generalized estimating equations (WGEE) and mixed models can be applied.

In longitudinal studies, the missingness can be due to any of the following reasons: death of a subject, withdrawal from the study, or loss to follow-up. These differential reasons for missingness in longitudinal data analyses present a challenge to the statistical analyst. The losses of data can result in biased estimates and a loss in precision. The loss in precision is proportional to the amount of missing data. In addition, the effect depends on the association between the observed data and the missing data. Improper adjustment or no adjustment for missing data in a regression analysis can result in biased estimates of parameters and lead to erroneous inferences (Hogan et al., 2004). There are some likelihood based methods such as selection models, mixture models and shared parameter models that can be used for analyzing non-ignorable missing longitudinal data (details in Chapter 2). In these modeling approaches identification of parameters is problematic and implementation of these methods is not trivial. In addition, none of the likelihood based approaches adjust the likelihood function for differential reasons for missingness. To account for the differential reasons for missingness, we will compute the probabilities of observing a response for a subject belonging to a missingness category and invert this probability to obtain the weights. This basic idea of weighting adjustments to reduce the bias in estimation is common in sample survey for finite population randomization inference (Horvitz and Thompson, 1952). These derived weights will be used in weighted pseudo-likelihood (WPL) methods for analyzing differential reasons for missing longitudinal biomarker data (Lawless et al., 1999). The proposed WPL methods will be relatively easy to implement using standard statistical software, and provide an extension to the currently available methods for analyzing differential missingness.

In a longitudinal study of a biomarker with differential reasons for missingness, some measurements of the biomarker may also be censored. In the GenIMS study there are measurements that are left-censored for a panel of biomarkers. The left-censored data are characteristic of many bioassays due to the inherent limit of quantification in the assays. In the GenIMS study the censoring of the biomarker measurements occurred when the level of the biomarker was below the detection limit of the assay. Moreover, there are missing measurements that occurred primarily when the subjects in the study were discharged from the hospital or died. In addition to differential reasons for dropout and death, left-censoring leads to another level of complication in analyzing longitudinally measured biomarker data.

There are few methods for analyzing left-censored longitudinal data. Tobit regression is one of the classical approaches for analyzing left-censored data (Tobin, 1958, Amemiya, 1984). A semi-parametric estimator was derived using a fixed effect tobit model for panel data (Honore, 1992). A tobit-based variance components method has been developed to account for the censoring process in a variance component analysis (Epstein et al., 2003). Standard linear mixed models are used by omitting the censored data or imputing a fixed value. The most widely used methods for left-censored longitudinal data are imputing the quantification limit (Keet et al.,1997), using half of this limit (O'Brien et al.,1998) or the use of random imputation procedures (Paxton et al.,1997). Omitting censored data clearly results in a loss of information and the statistical properties of imputing a value are unclear (Beal, 2001). All of these ad-hoc methods for left-censored longitudinal data produce biased estimates and incorrect standard errors (Ghebregiorgis, 2008). A more efficient approach to multiple imputation has been proposed for analyzing censored longitudinal data using a linear mixed model (Hugues, 1999, Jacqmin-Gadda et al.,2000). In the case of left-censored and informative dropout longitudinal data, a maximum likelihood method has been developed to estimate parameters and standard errors (SE) were computed from a numerically derived observed information matrix (Lyles et al., 2000). All of these approaches are based on the full likelihood method. Using the full likelihood, the estimation of parameters and computation of the SE involve a series of multiple integration, numeric and algebraic complexities. When the rate of censoring is high, the integration becomes prohibitive and estimates are unstable for more than two random effects (Ghebregiorgis, 2008).

In summary, this research focus on addressing two statistical issues for analyzing longitudinally measured biomarker data.

First. The first issue is the non-ignorable missingness due to the differential reasons for dropout, and death. We are proposing to extend the pseudo likelihood method to the weighted pseudo likelihood (WPL) method for analyzing longitudinally measured biomarker data. In this new method weights will be used for the adjustment of the missing data and considered as nuisance parameters in the analysis. The consistent estimate of the variance covariance matrix of the parameters of interest will be computed by considering the fact that there are an infinite number of nuisance parameters used in the estimation process.

Second. The second issue is the left-censoring along with the non-ignorable missingness in analyzing longitudinal biomarker data. We are proposing to extend the theory of Tobit regression for the left-censored data to develop a Weighted Random Effects Tobit regression model using WPL theory for non-ignorable missing and left-censored longitudinal biomarker data. Again, the effect of infinitely many nuisance parameters (weights) in the estimation process will be taken into account and various censoring process will be compared to find a best one for analyzing left-censored and non-ignorable missing data. Correct standard errors of estimates will be computed using WPL theory.

The performance of the fitted models will be compared with a number of widely used models. So far to our knowledge no one has utilized the WPL theory in analyzing non-ignorable missing and/or left-censored longitudinal data.

2.0 A REVIEW OF LIKELIHOOD METHODS FOR ANALYZING NON-IGNORABLE MISSING AND LEFT-CENSORED LONGITUDINAL BIOMARKER DATA

2.1 LIKELIHOOD BASED METHODS FOR ANALYZING NON-IGNORABLE MISSING LONGITUDINAL DATA

The classical repeated measures design has been used for analyzing longitudinally measured continuous data. In this design, each subject is measured a fixed number of times under different conditions to compare the effect of (usually) treatments. In a repeated measures design the experimental conditions are the within-subject factors which are usually compared using within-subject contrasts. In this design, each subject acts as his or her own control and the estimate of the effects of the factor are free of any between subject variation in the outcome (Fitzmaurice, et al. 2004). This design is not suitable for the case where there are an unequal number of repeated measures and also has very strict assumptions on the variance covariance structure of the measures. Moreover, there is no method that allows for missing data using this design.

The linear mixed model is widely used for modeling longitudinally measured continuous data. It is a generalization of the standard linear model and allows for flexibility. There are many choices that can be made for the variance covariance structure of the correlated data in this modeling. In this model the number of repeated measures can vary from individual to individual and missing data can be handled under the assumption of missing at random (MAR). If the data are missing not at random then one can not use the linear mixed model directly. When the probabilities of response depend on the unobserved level of biomarker data, then the missingness mechanism is known as not missing at random (NMAR) (Little and Rubin, 2002). In such a scenario, the standard likelihood based methods for analyzing longitudinal biomarker data do not include a mechanism for incorporating different reasons for loss to follow-up or death. When biomarker measurements are missing due to dropout or death, the two types of loss to follow-up are different and should not be combined (Dufouil et al., 2004, p.2215).

The majority of MNAR longitudinal data analysis techniques are based on a factorization of the joint distribution $f(Y, R|X)$, where Y is the full response data, X is the covariate and R indicates the missing data mechanism (Hogan et al. 2004, p. 1466). Likelihood based approaches such as the selection model, mixture model and shared parameter model are common for modeling non-ignorable missing longitudinal data. If the regression model is based on the joint distribution which is a product of the full data model, $f(y|x)$ and $f(r|y, x)$ then it is called a selection model. In selection modeling, the identification of model parameters depends on some unverifiable model assumptions. Generally this modeling technique requires specialized numerical routines for maximizing the likelihood function with the uncertainty of a well behaved likelihood function and consequently unstable estimation of the model parameters (Kenward, 1998). If the full data is modeled as a mixture over drop-out categories then it is a (pattern) mixture model ($f(y, r|x) = f(y|r, x)f(r|x)$). These models are under-identified and well suited for small percentages of missing observations (Little, R.J.A., 1994, Little, R.J.A., 1993, Little, R.J.A. and Wang, Y., 1996, Daniels, M.J. and Hogan, J.W. 2000). In modeling, if a latent random effect is being used to characterize the dependence between the response, Y, and the missing data indicator, R, then it is known as shared latent process model. With these models, the identification is heavily dependent on the distribution of the arbitrarily chosen shared random effects which affects the validity of the findings (Pulkstenis, et al., 1998).

In the full data modeling setting every observation is equally weighted. For modeling data with missing observations, weighting techniques have been used for semi-parametric regression modeling (Robins et al., 1994, 1995). The weighting procedure has been applied in analyzing many incomplete longitudinal data problems by Rotnitzky and Robins (1997), Lipsitz et al (1999), Lin (2003), Demirtas (2004), Dufouil et al (2004), Lin et al (2004), Ibrahim et al (2005). Weights are computed by inverting the probabilities of response. In a longitudinal study some subjects are more likely to complete the study than others. Therefore, our intention is to apply the weighting methodology in the setting of likelihood-based approaches. The pseudo-likelihood approach has been used for estimating parameters in generalized linear mixed models (Wolfinger and O'connell, 1993). Therefore, our goal is to estimate the parameters of a regression model for longitudinal data with differential reasons for loss to follow-up by extending the existing full likelihood approaches such as the mixed model and pseudo-likelihood methods. In this endeavor, we propose a model for longitudinal biomarker data with non-ignorable non-monotone missingness using a weighted pseudo-likelihood method. To our knowledge, these proposed estimation methods for analyzing longitudinal biomarker data with differential reasons for missingness do not appear in the literature.

For the pseudo-likelihood methods the weights are treated as the nuisance parameters. Nuisance parameters are the parameters which are not of direct inferential interest in the modeling. Our interest is to estimate the usual regression coefficients (parameters) of a general linear mixed effects model. A general method to eliminate the nuisance parameter is via the profile likelihood method but the estimates are biased when the number of nuisance parameters becomes large and the bias does not go away even for large samples, particularly when there are infinitely many nuisance parameters (Pawitan (2001), p. 274). Rather than being based on the observed data likelihood function, an estimate of the parameter of interest can be obtained if we use the observed response only and weight (nuisance parameters) their contribution to the likelihood function by inverting the probability of response (Lawless et al. 1999, p. 421). This approach will be called the weighted pseudo maximum likelihood method of estimation which basically uses the idea of the Horvitz-Thompson estimation procedure applied in unequal probability sampling (Thompson (2002), p.53).

2.2 LIKELIHOOD BASED METHODS FOR ANALYZING NON-IGNORABLE MISSING AND LEFT-CENSORED LONGITUDINAL DATA

In addition to non-ignorable missingness, analysis of censored longitudinal biomarker data is a challenge. A standard method for the analysis of censored data is Tobit regression (Tobin, 1958). Tobit regression has been extended to multivariate regression (Amemiya, 1984). Recently a Box-Cox transformation has been used for the analysis of left-censored cross sectional data (Han and Kronmal, 2004). In fitting linear mixed effect models, the Markov Chain Monte Carlo (MCMC) EM algorithm has been used to accommodate censoring in longitudinal data (Hughes, 1999). Lyles et al (2000) analyzed left-censored and informative dropout HIV data by maximizing a single likelihood function which has integrated the censoring and informative dropout process. They estimate the parameters from this complicated likelihood function and compute the standard errors using the observed information matrix directly. Linkage analysis of left-censored trait data has been based on a variance component tobit model (Epstein, et al. 2003). In this modeling approach, the standard generalized liner mixed model has been modified using the idea of Tobit model to accommodate the censored data. As used in these references, there are many issues in analyzing left-censored longitudinal data using a full likelihood. Beyond the algebraic and numeric intractability, it requires computation of a series of multiple integrals and becomes intractable for the case of a high rate of censoring. In addition, for more than two random effects the convergence of the estimates remains uncertain (Ghebregiorgis, 2008). As a remedy, the pseudo likelihood method has been used for analyzing multivariate longitudinal biomarker with left-censored data (Ghebregiorgis, 2008). But this method has not been developed for left-censored single longitudinal biomarker data with non-ignorable missingness.

The above mentioned standard and practiced methods for left-censored longitudinal data have no mechanism for incorporating differential reasons for loss to follow up, which can present problems. In this study we are proposing a weighted pseudo likelihood method for left-censored and non-ignorable missing longitudinal data. This method will be compared with the un-weighted random effect tobit model, and with the weighted random effect tobit model. The weights will be computed by inverting the probability of measuring a biomarker measurement from a subject using a multinomial logistic regression model. To the best of our knowledge, no one has addressed the problem of analyzing left-censored and non-ignorable missing longitudinal biomarker data using a weighting technique. In addition to the real data analysis, an extensive simulation will be performed to understand the variability in the inferences under different scenarios (or designs), percentages of missing and censoring processes.

3.0 ANALYSIS OF LONGITUDINAL BIOMARKER DATA WITH DROPOUT AND DEATH USING WEIGHTED PSEUDO LIKELIHOOD THEORY

3.1 INTRODUCTION

In longitudinal studies the data are collected over a period of time from each individual in the study resulting in missing data for a variety of reasons. A popular method for analyzing longitudinal data is linear mixed model which assumes that the missing data mechanism is missing at random (MAR). When the missing data mechanism is missing not at random (MNAR), then the standard linear mixed model cannot be used in analyzing these types of data. Though there are a few composite likelihood-based methods for analyzing non-ignorable missing longitudinal data, there is a lack of statistical methods for analyzing longitudinal data when the data are missing due to premature dropouts, deaths of the subjects, and administrative reasons, etc. In this work we are proposing a weighted pseudo likelihood (Lawless, et al. 1999) method for analyzing differential reasons for missing longitudinal continuous biomarker data in the linear mixed model frame work.

The motivation for this study comes from the Genetic and Inflammatory marker of sepsis study (GenIMS). The GenIMS study was a longitudinal cohort study of subjects with community acquired pneumonia that were recruited from 2001-2003 in 28 hospitals located in Pennsylvania, Connecticut, Michigan and Tennesse. One major goal of the GenIMS study was to understand the role of inflammatory markers in the progression of pneumonia to sepsis (Kellum, et al, 2007). In this study, one of the important biomarkers for understanding the mechanisms of progression to sepsis is anti-thrombin. Longitudinal anti-thrombin measurements were obtained for the first seven days in the study, but some of the measurements are missing due to subject discharge from the hospital or death within the first seven days. In addition, there was intermittent missingness in the measurements due to administrative and other reasons. Figure 1.1 presents the median level and the trends of the anti-thrombin biomarker data for the three groups of subjects over the seven day period. The subjects who have dropped out from the study (579 subjects) had the highest level of anti-thrombin, followed by the subjects who have completed the study (341 subjects) due to a full seven days of hospitalization with no missing data, and the subjects who died during the first seven days of hospitalization (19 subjects). This mechanism of missingness leads to non-ignorable missingness which can impact estimation and any inferences drawn from the data (Little and Rubin, 2002). This plot points to the differences in the anti-thrombin levels in these three groups of subjects and the need to account for these differences in the analysis as illustrated by the fact that subjects who died had the lowest anti-thrombin levels during the study.

Many longitudinal methods do not routinely incorporate missing data patterns into the analysis. The problem is further complicated in settings such as those presented here, where missing values occur due to reasons that may impact the outcome variable of interest. When measurements of biomarkers are missing due to dropout or death, the two types of loss to follow-up are different and should not be combined since the outcome may differ due to the reason for missingness (Dufouil et al., 2004). Likelihood based approaches such as the selection model, mixture model, and shared random effects model are common techniques for modeling non-ignorable missing longitudinal data (Rubin, 1977; Wu and Bailey, 1988; Little, 1994; Fallmann and Wu, 1995; Kenward, 1998; Hogan et al. 2004). Recently, joint modeling has also been proposed as another likelihood based method for the modeling of non-ignorable longitudinal missing data (Tsiatis and Davidian, 2004). A Pseudo-likelihood approach has also been used for the estimation of parameters in generalized linear mixed models (Wolfinger and O'connell, 1993).

In modeling settings where there are no missing data, each observation is equally weighted. For modeling data with missing observations, inverse probability weighting (IPW) techniques (Horvitz and Thomson, 1952) have been applied to semi-parametric regression modeling to give different weights to account for the probability of missingness (Robins et al., 1994, 1995). In the current literature, this weighting procedure has been applied for analyzing many incomplete longitudinal data problems (Rotnitzky and Robins, 1997; Lipsitz et al, 1999; Lin, 2003; Demirtas, 2004; Lin et al, 2004; Ibrahim et al, 2005). Our objective here is to estimate the parameters of a linear mixed model for longitudinal data with differential reasons for loss to follow-up by utilizing the weighted pseudo-likelihood (WPL) theory. We will estimate weights for each observation by inverting the probabilities of response. These estimated probabilities are proportional to the likelihood of measuring the values of the biomarker and computed using the logistic regression. Since most statistical packages contain options for accommodating weights, the estimated IPW can easily be placed into the log-likelihood function and hence the estimated quantities and inference will account for the non-ignorable missingness in the data. These weights will be used as an adjustment of the loss to dropout or death data. The price for incorporating these estimated weights into the likelihood function is that the number of nuisance parameters becomes large. The number of nuisance parameters is proportional to the number of occasions of measurement and the number of subjects in the study. So there is a need for statistical methods for estimating the parameters of interests in the presence of a large number of nuisance parameters in likelihood based inferences that include differential reasons for missing longitudinal data. In this endeavor our proposed WPL approach, which is an extension of the pseudo-likelihood approach (Lawless et al., 1999), will eases the numerical complexities. The wide use of longitudinal data modeling in many fields of application with the challenge of differential reasons for missingness is improved with these weighted estimation methods in the framework of standard statistical software.

Under the different underlying assumptions about the populations and pattern of missingness, we will compare the performance of the standard linear mixed model and the weighted linear mixed model with the proposed weighted linear mixed model fitted by the WPL theory. The methods are compared in terms of the bias, efficiency, root mean square error, and coverage probability. In the next section, we will describe likelihood based methods as well as our proposed methods for analyzing non-ignorable missing longitudinal biomarker data. In section 3.3 we will present the results from the analysis of GenIMS study data and in section 3.4 we will describe a simulation study to judge the performance of all estimators. In the section 3.5 will provide a discussion.

3.2 NOTATIONS AND MODEL FRAME WORK

Let Y_{ij} denote the measurement of a biomarker from the ith subject at the jth wave of measurements at time $t_{ij}, i = 1, 2, ..., N, j = 1, 2, ..., n_i$ and $X_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijp})'$ denote a $p \times 1$ vector of covariates associated with Y_{ij}. In vector notation, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ is the n_i-dimensional vector of biomarker measurements and $\mathbf{X}_i$ is the $n_i \times p$ matrix of covariates from the ith subject. However, in most longitudinal study the vector, $\mathbf{Y}_i$ is not always fully observed. Suppose that the observed and missing component of $\mathbf{Y}_i$ are denoted by $\mathbf{Y}_i^o$ and $\mathbf{Y}_i^m$ respectively. We define the missingness indicator vector $\mathbf{R}_i = (R_{i1}, R_{i2}, ..., R_{in_i})'$ where R_{ij} =0 if Y_{ij} is observed, =1 otherwise. Note, in this paper we will assume that $\mathbf{Y}_i$ and $\mathbf{Y}_{i'}$, $(i \neq i')$, are independent and the covariates vector X_{ij} is fully observed.

Little and Rubin (2002, p. 118) defined the joint density of the full data $(\mathbf{Y}_i, \mathbf{R}_i)$ as,

$$f(\mathbf{y}_i, \mathbf{r}_i | X_i, Z_i, \boldsymbol{\gamma}, \boldsymbol{\psi}) \tag{3.1}$$

where X_i and Z_i are design matrices for fixed and random effects respectively, and ($\boldsymbol{\gamma}$, $\boldsymbol{\psi}$) is the parameter space for this joint density. Let $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ denote the parameter vectors associated with X_i and Z_i respectively, then $\boldsymbol{\gamma}$=($\boldsymbol{\beta}$,$\boldsymbol{\alpha}$) and $\boldsymbol{\psi}$ characterizes the observed response and missingness process respectively. Replacing $\mathbf{Y}_i$ by $(\mathbf{Y}_i^o, \mathbf{Y}_i^m)$ the full data density can be written as,

$$f(\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{r}_i | X_i, Z_i, \boldsymbol{\gamma}, \boldsymbol{\psi}) \tag{3.2}$$

Using the full data density (2), the full likelihood function of the parameter space ($\boldsymbol{\gamma}$, $\boldsymbol{\psi}$) can be written as

$$L^*(\boldsymbol{\gamma}, \boldsymbol{\psi}) = \prod_{i=1}^{N} f(\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{r}_i | X_i, Z_i, \boldsymbol{\gamma}, \boldsymbol{\psi}) \tag{3.3}$$

Since parameter estimation and inference are based on the observed data, so the full likelihood function (3) is proportional to following observed data likelihood:

$$\begin{aligned}L(\boldsymbol{\gamma},\boldsymbol{\psi}) &= \prod_{i=1}^{N} f(\mathbf{y}_i^o,\mathbf{r}_i|X_i,Z_i,\boldsymbol{\gamma},\boldsymbol{\psi})\\ &= \prod_{i=1}^{N}\int f(\mathbf{y}_i,\mathbf{r}_i|X_i,Z_i,\boldsymbol{\gamma},\boldsymbol{\psi})d\mathbf{y}_i^m\\ &= \prod_{i=1}^{N}\int f(\mathbf{y}_i^o,\mathbf{y}_i^m|X_i,Z_i,\boldsymbol{\gamma})f(\mathbf{r}_i|\mathbf{y}_i^o,\mathbf{y}_i^m,X_i,\boldsymbol{\psi})d\mathbf{y}_i^m\end{aligned} \tag{3.4}$$

where the limits of integration are over the values of unobserved biomarker, $\mathbf{Y}_i^m$.

If the distribution of observed response and probability of missingness can be specified correctly, then the maximum likelihood estimates can be obtained using likelihood equation (4) and the asymptotic covariance matrix of the estimates can be obtained by inverting the observed Fisher information matrix (Little and Rubin (2002), p. 315). The full specification of the likelihood function (4) involves the identification of the distribution of $f(\mathbf{y}_i|X_i,Z_i,\boldsymbol{\gamma})$ and multinomial distribution of the probability of missingness, $f(\mathbf{r}_i|\mathbf{y}_i,X_i,Z_i,\boldsymbol{\psi})$. Estimation of the parameter of interest, $\boldsymbol{\gamma}$, involves the estimation of many nuisance parameters $\boldsymbol{\psi}$. The number of nuisance parameters increases as to the number of subjects and measurement time increases. A general method to eliminate the nuisance parameter is via the profile likelihood method but the estimates are biased when the number of nuisance parameters become large and the bias persists for large samples, particularly when there are infinitely many nuisance parameters (Pawitan (2001), p. 274). Rather than basing the estimation on the observed data likelihood function (4), an estimate of $\boldsymbol{\gamma}$ can be obtained if we use the observed response and weight (nuisance parameters) their contribution to the likelihood function by the inverse of the probability of observing the response (Lawless et al. 1999, p. 421). This approach will be denoted as the weighted pseudo maximum likelihood method (WPL) which applies the idea of the Horvitz-Thompson estimation procedure for unequal probability sampling to this problem (Thompson (2002), p.53). Therefore, an alternative to deal with the complicated observed data likelihood (4) would be the following weighted pseudo loglikelihood,

$$l(\boldsymbol{\gamma},\boldsymbol{\omega}_i) = \boldsymbol{\omega}_i\mathtt{log}f(\mathbf{y}_i^o|X_i,Z_i,\boldsymbol{\gamma}) \tag{3.5}$$

where $\boldsymbol{\omega}_i = \boldsymbol{\pi}_i^{-1} = (\sum_{h=1}^{H} p_{ih}\delta_{ih})^{-1}$; p_{ih} is the probability of observing the biomarker from the ith subject belongs in dropout category h, and $\delta_{ih} = 1$ if the ith subject belongs to category h, 0 otherwise. We assume that the estimate of p_{ih} will be obtained by a method that produces a consistent estimate and, not necessarily by the method of maximum likelihood. Inference for pseudo-likelihood estimates is based on the asymptotic theory and accounts for the extra variability introduced by the use of nuisance parameters estimates. Now the score function of the parameter of interest, $\boldsymbol{\gamma}$, is

$$S(\boldsymbol{\gamma}, \boldsymbol{\omega}) = \sum_{i=1}^{N} R_i \boldsymbol{\omega}_i \frac{\partial}{\partial \boldsymbol{\gamma}} logf(y_i^o|X_i, Z_i, \boldsymbol{\gamma}). \tag{3.6}$$

The solution of $S(\boldsymbol{\gamma}, \hat{\omega}_i)$=0 will provide a unique Weighted Pseudo Maximum Likelihood Estimates(WPL) of $\boldsymbol{\gamma}$. The asymptotic covariance matrix of WPL $\hat{\boldsymbol{\gamma}}$ has the following form (Lawless et al. 1999, p. 426),

$$Var(\widehat{\gamma}) = \imath_{11}^{-1}(\jmath_{11} - \imath_{12}\imath_{22}^{-1}\imath_{12}^{T})\imath_{11}^{-T} \tag{3.7}$$

where

$$\imath_{11}(\gamma, \omega) = E\left[\sum_{i=1}^{N} -\hat{\omega}_i \frac{\partial^2}{\partial\gamma\partial\gamma'} logf(y_i^o|X_i, Z_i; \gamma)|_{\gamma=\hat{\gamma}}\right] \tag{3.8}$$

$$\imath_{12}(\gamma, \omega) = E\left[\sum_{i=1}^{N} \left\{S^*(\omega_i)\frac{\partial}{\partial\gamma}\mathtt{log}f(y_i^o|X_i, Z_i; \gamma)\right\}|_{\gamma=\hat{\gamma}, \omega_i=\hat{\omega}_i}\right] \tag{3.9}$$

$$\imath_{22}(\omega) = E\left[\sum_{i=1}^{N} I^*(\omega_i)|_{\omega_i=\hat{\omega}_i}\right] \tag{3.10}$$

$$\jmath_{11} = \mathtt{Var}(S) = \sum_{i=1}^{N} S_i S_i^T \tag{3.11}$$

$$= \sum_{i=1}^{N} R_i\{(\sum_{h=1}^{H} p_{ih}\delta_{ih})^{-1}\frac{\partial}{\partial\gamma}\mathtt{log}f(y_i^o|X_i, Z_i, \gamma)\}\{(\sum_{h=1}^{H} p_{ih}\delta_{ih})^{-1}\frac{\partial}{\partial\gamma}\mathtt{log}f(y_i^o|X_i, Z_i, \gamma)\}^T$$

and $S^*(\omega_i)$ and $I^*(\omega_i)$ are the score function and Fisher informatiton for the nuisancee parameters ω_i respectively. Since differential missingness or dropout is a categorical variable, a multinomial logistic regression model and hence a likelihood function for the parameters of the model will be developed to estimate ω_i, $S^*(\omega_i)$ and $I^*(\omega_i)$.

Let us define a categorical variable D which represent three categories of patients such as completers, dropout and death, coded as 0, 1 and 2 respectively. For simplicity in notations, the multinomial logistic regression model will be formed by taking one covariate (observed response, y_i^o) and a constant term, denoted by the vector, $\mathbf{x}^* = (1, y_i^o)'$, of length 2. This one covariate model can easibly be extended to the p covariate model. Now, taking completers(0) as the reference category, dropout(1) and death(2) as comparison categories, the multinomial logistic regression model (Agresti(2004, p.268)), can be written as

$$log\left(\frac{\pi_{hi}}{\pi_{0i}}\right) = g_h(\mathbf{x}_i^*) = \lambda_{h0} + \lambda_{h1} y_i^o, h = 1, 2$$
$$= (\mathbf{x}_i^*)'\theta_h$$

It follows that the conditional probability given the covariate vector in the three category model can be obtained by the following formula:

$$\pi_{hi} = \frac{exp(g_h(\mathbf{x}_i^*))}{\sum_{h=0}^{2} exp(g_h(\mathbf{x}_i^*))} \tag{3.12}$$

where the vector $\theta_0 = \mathbf{0}$ and $g_0(\mathbf{x}^*) = \mathbf{0}$. Each probability is a function of the vector of 4 parameters $\theta = ({\theta_1}', {\theta_2}')$. According to the outline in Hosmer and Lemeshow (2000, p. 262), for the development of the likelihood function of the parameter vector θ, we define three(3) dummy variables that coded as 0 or 1 to indicate the group membership of an observation. These variables are coded as follows: if D=0 then $D_0 = 1$, $D_1 = 0$ and $D_2 = 0$; if D=1 then $D_0 = 0$, $D_1 = 1$ and $D_2 = 0$; if D=2 then $D_0 = 0$, $D_1 = 0$ and $D_2 = 1$. Note, $\sum_{h=0}^{2} D_h = 1$ and these dummy variables are introduced for the construction of the likelihood function and are not used in the actual multinomail logistic regression analysis. Using this notation, the likelihood function for the parameter vector θ is

$$L(\theta) = \prod_{i=1}^{N} \left[\pi_0({\mathbf{x}_i}^*)^{d_{0i}} \pi_1({\mathbf{x}_i}^*)^{d_{1i}} \pi_2({\mathbf{x}_i}^*)^{d_{2i}}\right] \tag{3.13}$$

Putting $d_{0i} = 1 - d_{1i} - d_{2i}$ for each i and taking the logarithm, the log-likelihood function can be expressed as

$$logL(\theta) = l(\theta) = \sum_{i=1}^{N} \left[d_{1i} g_1({\mathbf{x}_i}^*) + d_{2i} g_2({\mathbf{x}_i}^*) - log\left\{1 + exp(g_1({\mathbf{x}_i}^*)) + exp(g_2({\mathbf{x}_i}^*))\right\}\right] \tag{3.14}$$

Taking the derivative of the log-likelihood function with respect to each of the unknown parameters of θ, the general form of the score function is

$$S(\theta) = \frac{\partial}{\partial \lambda_{hk}} l(\theta) = \sum_{i=1}^{N} x_{ki}(d_{hi} - \pi_{hi}) \tag{3.15}$$

for h=1,2; k=0,1 (subscripts for covariates) with $x_{0i} = 1$ and $x_{1i} = y_i^o$ for each subject. Equating the score to zero an iterative solution of θ can be obtained from these equations. The information matrix for $\hat{\theta}$ can be computed by taking second partial derivatives of the above log-likelihood function:

$$\frac{\partial^2}{\partial \lambda_{hk} \partial \lambda_{hk'}} l(\theta) = - \sum_{i=1}^{N} x_{k'i} \pi_{hi}(1 - \pi_{hi}) \tag{3.16}$$

and

$$\frac{\partial^2}{\partial \lambda_{hk} \partial \lambda_{h'k'}} l(\theta) = \sum_{i=1}^{N} x_{k'i} x_{ki'} \pi_{hi} \pi_{h'i} \tag{3.17}$$

for $h \neq h'$=1,2 and $k \neq k'$=0,1. By negating these two sets of equations and evaluating at $\hat{\theta}$ a 4×4 observed information matrix $I(\hat{\theta})$ can be obtained. The ultimate purpose of the above derivations is to derive the score function and Fisher information matrix of the function of parameter vector θ. The score function and Fisher information of ω_i which is a function of the parameter vector θ can be derive in the following way:

$$S^*(\omega_i) = \frac{\partial}{\partial \omega_i} logL(\theta) = \frac{\partial \theta}{\partial \omega_i} \frac{\partial}{\partial \theta} logL(\theta) = \left[\left(\frac{\partial \omega_i}{\partial \theta} \right)^{-1} \right] S(\theta) \tag{3.18}$$

and

$$I^*(\omega_i) = varS^*(\omega_i) = \left[\left(\frac{\partial \omega_i}{\partial \theta} \right)^{-1} \right]' I(\theta) \left[\left(\frac{\partial \omega_i}{\partial \theta} \right)^{-1} \right] \tag{3.19}$$

where

$$\begin{aligned} \frac{\partial \omega_i}{\partial \theta} = \frac{\partial}{\partial \lambda_{hk}} & \left[\frac{\sum_{h=0}^{2} exp(g_h(\mathbf{x}_i^*))}{exp(g_h(\mathbf{x}_i^*))} \right] \\ &= x_{ki}(1 - \omega_{hi}) \end{aligned}$$

Now we will derive the score and Fisher information of the parameter vector $\boldsymbol{\gamma}$ to compute $\imath_{11}$ and $\imath_{12}$ respectively. A popular model for fitting inherently unbalanced longitudinal continuous biomarker data is the linear mixed model(LMM). A LMM for the ith subject can be written in the following form(Laird and Ware 1982):

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{3.20}$$

and the vector $\mathbf{Y}_i$ is distributed as multivariate normal with the following specification:

$$\mathbf{Y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i) \tag{3.21}$$

$$\mathbf{b}_i \sim N(0, \mathbf{D}) \tag{3.22}$$

where $\mathbf{X}_i$ is the $n_i \times p$ design matrix for the fixed effects, $\boldsymbol{\beta}$; $\mathbf{Z}_i$ is the $n_i \times q$ design matrix for the random effects, $\mathbf{b}_i$; $\mathbf{D} = Cov(\mathbf{b}_i)$ and $\boldsymbol{\Sigma}_i$=$Cov(\mathbf{e}_i)$ are the covariance matrices of the random effects and errors respectively.

From (20) the marginal distribution of $\mathbf{Y}_i$ is normal with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_i$=$\boldsymbol{\Sigma}_i$+ $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$, so the log-likelihood function of the fixed parameters $\boldsymbol{\beta}$ is

$$\mathtt{log}L(\boldsymbol{\beta}) = -\frac{N}{2}\mathtt{log}2\pi - \frac{1}{2}\mathtt{log}|\mathbf{V}| - \frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{3.23}$$

By differentiating the log-likelihood function (23) with respect to the parameter vector of $\boldsymbol{\beta}$ one and two times we can compute the score and observed Fisher information respectively:

$$\frac{\partial\mathtt{log}L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i'\mathbf{V}^{-1}\mathbf{Y}_i \tag{3.24}$$

$$\frac{\partial^2\mathtt{log}L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \mathbf{X}_i'\mathbf{V}^{-1}\mathbf{X}_i \tag{3.25}$$

3.3 APPLICATION: ANALYSIS OF ANTITHROMBIN BIOMARKER DATA FROM GENIMS STUDY

The GenIMS study was a large cohort study of patients with community acquired pneumonia followed over time. In this study, a series of biomarkers were measured daily on a subset of hospitalized patients for seven days. These biomarkers assessed the potential pathways of inflammation and coagulation related to the development of sepsis. The primary objective of the GenIMS study was to identify the potential biomarkers of sepsis. There were 2320 patients enrolled through the emergency departments in 28 hospitals in PA, CT, MI, and TN (2001-2003). From the pool of 2320 subjects we have found 939 subjects had at least some biomarkers measurements as well as covariates information from day 1 to day 7. The data set for our analysis consisted of 341 (36.3%) patients with seven full days of data, 579 (61.7%) patients who were discharged before the full seven days, and 19 (2.0%) patients who died during the first 7 days. The outcome variable for our analysis is the longitudinally measured anti-thrombin biomarker. The median level of the anti-thrombin biomarker for patients who dropped-out was higher followed by completers, and subjects who died (Figure 1). Higher levels of anti-thrombin indicate better health condition of the subject and hence their discharge from the hospital which resulted in drop out from the study. Patients, who died during the study, and obviously their biomarker measurements, are missing. This missingness mechanism of the biomarker, due to better health condition and death, is missing not at random (MNAR) and suggests taking this into account during the analysis.

We are applying our proposed WPL method for analyzing differential reasons for missingness in the anti-thrombin longitudinal biomarker data. We will compute the weights by inverting the probability of observing anti-thrombin biomarker data. Details on the IPW technique have been described in section 2. These weights will account for the differential reasons for missingness and are treated as nuisance parameters in the estimation process. Inferences for the parameter of interests in the presence of infinite number of nuisance parameters are made using the asymptotic PL theory. We are comparing the WPL method with the standard and weighted linear mixed models. In fitting the standard linear mixed model it is assumed that the missingness is missing at random. In our anti-thrombin analysis, we are considering time, time2 and statin use as fixed effects and the intercept as a random effect in the linear mixed model.

Table 3.1 presents the results from the analysis of differential reasons for missing longitudinal anti-thrombin biomarker data using WPL method. Weighted estimates of the parameters are larger compared to the estimates obtained from the SMM. Also the SE of the estimate from the weighted models are larger than the corresponding SMM estimate. Z-values for the quadratic term Day2 and Statin use by the WPL method are substantially different from the other three methods. According to the WPL method, the Day2 term should be dropped from the model and statin use is marginally significant. Note that the standard error of the estimate of WPL is the largest among the four methods followed by WMME, WMM and SMM. Again, the WPL accounts the nuisance parameters (weights) in the computation of the variance of the estimates.

3.4 SIMULATION STUDY FOR NON-IGNORABLE MISSING LONGITUDINAL BIOMARKER DATA

To evaluate the performance of the proposed weighted pseudo likelihood (WPL) methods, an extensive simulation study was conducted. Using theis approach we have compared the following models:
i) SMM: Standard linear mixed model with the SE of the estimates computed using the Fisher information
ii) WMM: Weighted linear mixed model with the SE of the estimates computed using the Fisher information
iii)WMME: Weighted linear mixed model with the SE of the estimates computed using the sandwich estimator
iv) WPL: Weighted linear mixed model with the SE of the estimates computed using asymptotic pseudo-likelihood (PL) theory.

For our simulation study, we have generated the anti-thrombin biomarker data from

Table 3.1: Analysis of non-ignorable missing longitudinal anti-thrombin biomarker data using weighted pseudo likelihood method

Variable	*Model*	*Coef.*	*Std.Err.*	*Z-statistic*	*p-value*
Intercept	SMM	4.441	0.011	399.630	<.0001
	WMM	4.473	0.012	385.920	<.0001
	WMME	4.473	0.025	182.500	<.0001
	WPL	4.473	0.123	36.376	<.0001
Day	SMM	-0.020	0.005	-4.000	<.0001
	WMM	-0.034	0.005	-6.470	<.0001
	WMME	-0.034	0.015	-2.210	0.027
	WPL	-0.034	0.019	-1.830	0.034
Day^2	SMM	0.005	0.001	7.580	<.0001
	WMM	0.006	0.001	8.130	<.0001
	WMME	0.006	0.002	2.770	0.006
	WPL	0.006	0.009	0.640	0.261
Statin Use	SMM	0.039	0.018	2.210	0.027
	WMM	0.043	0.018	2.420	0.016
	WMME	0.043	0.016	2.630	0.009
	WPL	0.043	0.032	1.360	0.087

a multivariate normal distribution with a specified mean vector and variance-covariance matrix. The mean vector and covariance matrix were obtained from the GenIMS data. The mean vector of the MVN distribution was as follows:

$$\mu_{jh} = \alpha + \beta_0 \; day_j + \beta_1 \; Statin \;\; Use$$

where Statin Use is a binary variable indicating whether patients were using statins prior to hospitalization; j(day) = 1,...,7 and h (dropout categories) =0, 1, or 2. The covariance of the MVN distribution has been drawn from the GenIMS study. The true values of the parameters were set to β_0=0.1 and β_1=0.56. The number of subjects (sample size, N) considered in the simulation study are 200, 500, and 1000 with a follow-up period of seven days. One thousand iterations were performed in each simulation study. There are three designs or scenarios have been considered for the simulation. In the first design, D1, 30% of the subjects have complete biomarker measurements at each of the seven days (completers) , 60% of the subjects have dropped out from the study (dropouts) , and 10% of the subjects died (death) in the study period. For the second design D2, there are 60% completers, 30% dropouts and 10% deaths. Third design, D3, consists of 70% completers and 30% dropouts. The generated anti-thrombin data were set to missing from the categories of dropouts and deaths at a rate of 10%, 20% and 30% at each wave of measurement to create missing data in the simulation study.

After generating the longitudinal biomarker data for seven days, we have created a dropout categorical variable representing the three categories of subjects in the study: completers, dropouts, deaths. Then we fitted the following multinomial logistic regression model with a generalized logit function to compute the probabilities of observing anti-thrombin biomarker data:

$$log(\frac{\phi_{hi}}{\phi_{0i}}) = g_h(x_i^*) = \lambda_{h0} + \lambda_{h1} \; day_{ij} + \lambda_{h2} \; Statin \;\; Use_i + \lambda_{h3} \; Antithrombin_{i1}$$

where i=1,2,...,N; $1 \leq j \leq 7$ and h=1, 2. Note, baseline anti-thrombin ($antithrombin_{i1}$) has been used in this logistic regression model. Using equations (12), (18), and (19), as derived in section 2, we have computed the weights, the score, and the Fisher information to obtain SE of the WPL estimates. The linear mixed model (20) has been fitted by considering time (days) and statin use as fixed effects, and by including the intercept as a random effect in the model. Maximum likelihood estimates of the fixed effects parameters are compared in terms of the bias, standard error (SE), root mean square error (RMSE), and coverage probability.

Table 3.2 presents the results from the analysis of simulated non-ignorable missing longitudinal anti-thrombin biomarker data with 30% completers, 60% dropouts, and 10% deaths. Results are presented for sample sizes of 200, 500, and 1000 subjects with 10% and 30% of the outcomes subject to missing values. Biases of the estimates from the standard linear mixed model, and weighted linear mixed models are minimal and negligible. There is no pattern in the bias due to the sample size and percentage of missingness. The SE of the estimates from the WPL model is the largest among the four SEs considered for the comparison. Consequently, the RMSE and the coverage probabilities of the WPL estimate are also the largest among the four estimates. The RMSE decreases with the increase in the sample size but there is no pattern that can be observed for the percentages of missing observations at each wave of measurement.

By comparing the all of the simulation results presented in Tables 3.2 - 3.4, we can report that bias depends on the number of subjects with incomplete measurements due to dropouts and deaths. Biases of the estimates are the smallest for design D3, where only 30% of the subjects have incomplete measurements due to the dropouts. The SE of the WPL is either a compromise or a competitor to the SE of the estimates of WMME. Similar observations can be found in terms of RMSE. In most cases, the RMSE increases with the increase in the percentages of missing observations but it decreases with the increases in the number of subjects. Also the RMSE of the WPL estimator is either a compromise or a competitor to the RMSE of the estimates of WMME. In terms of coverage probability, the WMM has poor performance while the WPL performs as expected.

Table 3.2: Analysis of differential reasons for missing longitudinal Anti-thrombin biomarker data with 30% completers, 60% dropouts, and 10% deaths(True value of the parameters: $\beta_0(time)$=0.1 and $\beta_1(StatinUse)$=0.56)

		Sample Size, N=200				Sample Size, N=500				Sample Size, N=1000			
		10% missing		30% missing		10% missing		30% missing		10% missing		30% missing	
Statistics	Model	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$
Bias*100	SMM	-0.024	0.043	0.016	0.046	0.004	-0.006	0.002	-0.012	0.000	-0.007	0.002	0.042
	WMM	-0.024	0.039	-0.003	0.047	-0.021	0.032	-0.007	-0.004	-0.033	0.042	-0.002	0.036
	WMME	-0.024	0.039	-0.003	0.047	-0.021	0.032	-0.007	-0.004	-0.033	0.042	-0.002	0.036
	WPL	-0.024	0.039	-0.003	0.047	-0.021	0.032	-0.007	-0.004	-0.033	0.042	-0.002	0.036
SE*100	SMM	0.200	0.963	0.187	0.973	0.127	0.619	0.148	0.624	0.090	0.438	0.105	0.442
	WMM	0.198	0.965	0.198	0.979	0.126	0.622	0.146	0.631	0.089	0.441	0.103	0.447
	WMME	0.319	1.038	0.295	1.028	0.207	0.672	0.220	0.660	0.148	0.480	0.157	0.471
	WPL	0.368	1.288	0.402	1.236	0.234	0.823	0.320	0.785	0.166	0.580	0.227	0.558
RMSE*100	SMM	0.202	0.964	0.188	0.974	0.127	0.619	0.148	0.624	0.090	0.439	0.105	0.444
	WMM	0.200	0.966	0.198	0.980	0.128	0.623	0.146	0.631	0.095	0.443	0.104	0.449
	WMME	0.320	1.039	0.295	1.029	0.208	0.673	0.220	0.660	0.151	0.482	0.157	0.472
	WPL	0.369	1.289	0.402	1.236	0.235	0.824	0.320	0.785	0.169	0.582	0.227	0.560
95%CP	SMM	0.85	0.95	0.88	0.95	0.86	0.92	0.85	0.92	0.86	0.94	0.84	0.94
	WMM	0.75	0.92	0.81	0.95	0.80	0.89	0.78	0.91	0.76	0.95	0.82	0.95
	WMME	0.94	0.95	0.96	0.96	0.95	0.92	0.95	0.93	0.96	0.97	0.97	0.96
	WPL	0.96	0.99	1.00	0.99	0.96	0.97	1.00	0.97	0.97	0.99	1.00	0.98

Table 3.3: Analysis of differential reasons for missing longitudinal Anti-thrombin biomarker data with 60% completers, 30% dropouts, and 10% deaths (True value of the parameters: $\beta_0(time)$=0.1 and $\beta_1(StatinUse)$=0.56)

		Sample Size, N=200				Sample Size, N=500				Sample Size, N=1000			
		10% missing		30% missing		10% missing		30% missing		10% missing		30% missing	
Statistics	Model	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$
Bias*100	SMM	0.015	-0.078	0.016	-0.072	0.002	-0.074	-0.002	-0.031	-0.008	-0.018	-0.006	-0.009
	WMM	-0.013	-0.029	-0.003	-0.043	-0.013	-0.053	-0.009	-0.030	-0.043	0.043	-0.011	-0.005
	WMME	-0.013	-0.029	-0.003	-0.043	-0.013	-0.053	-0.009	-0.030	-0.043	0.043	-0.011	-0.005
	WPL	-0.013	-0.029	-0.003	-0.043	-0.013	-0.053	-0.009	-0.030	-0.043	0.043	-0.011	-0.005
SE*100	SMM	0.177	0.970	0.187	0.976	0.112	0.614	0.119	0.617	0.079	0.435	0.084	0.436
	WMM	0.185	0.978	0.198	0.983	0.117	0.619	0.125	0.622	0.083	0.438	0.089	0.439
	WMME	0.308	1.066	0.295	1.030	0.196	0.679	0.189	0.658	0.139	0.480	0.135	0.465
	WPL	0.302	1.229	0.402	1.192	0.192	0.777	0.256	0.757	0.136	0.551	0.180	0.535
RMSE*100	SMM	0.177	0.973	0.188	0.979	0.112	0.618	0.119	0.618	0.079	0.435	0.084	0.436
	WMM	0.186	0.978	0.198	0.984	0.118	0.621	0.126	0.622	0.093	0.440	0.089	0.439
	WMME	0.308	1.066	0.295	1.031	0.197	0.681	0.190	0.659	0.146	0.482	0.135	0.465
	WPL	0.302	1.229	0.402	1.193	0.192	0.778	0.256	0.758	0.143	0.552	0.181	0.535
95%CP	SMM	0.85	0.94	0.86	0.95	0.87	0.94	0.87	0.95	0.872	0.95	0.86	0.95
	WMM	0.77	0.91	0.79	0.95	0.76	0.91	0.78	0.93	0.742	0.95	0.82	0.94
	WMME	0.94	0.93	0.94	0.95	0.95	0.94	0.94	0.95	0.942	0.96	0.94	0.95
	WPL	0.95	0.96	0.99	0.98	0.93	0.97	0.99	0.97	0.944	0.98	0.99	0.98

Table 3.4: Analysis of differential reasons for missing longitudinal Anti-thrombin biomarker data with 70% completers and 30% dropouts (True value of the parameters: $\beta_0(time)$=0.1 and $\beta_1(StatinUse$=0.56)

		Sample Size, N=200				Sample Size, N=500				Sample Size, N=1000			
		10% missing		30% missing		10% missing		30% missing		10% missing		30% missing	
Statistics	Model	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$	$\hat{\beta}_t$	$\hat{\beta}_s$
Bias*100	SMM	-0.015	-0.007	-0.027	-0.031	0.011	-0.049	0.005	-0.023	0.011	0.005	-0.002	0.018
	WMM	-0.004	-0.024	-0.015	-0.045	0.014	-0.053	0.006	-0.025	0.014	0.001	0.000	0.016
	WMME	-0.004	-0.024	-0.015	-0.045	0.014	-0.053	0.006	-0.025	0.014	0.001	0.000	0.016
	WPL	-0.004	-0.024	-0.015	-0.045	0.014	-0.053	0.006	-0.025	0.014	0.001	0.000	0.016
SE*100	SMM	0.238	0.982	0.363	1.009	0.151	0.619	0.229	0.639	0.107	0.438	0.162	0.452
	WMM	0.238	0.977	0.363	1.001	0.151	0.618	0.229	0.638	0.107	0.437	0.162	0.451
	WMME	0.324	1.010	0.480	1.036	0.205	0.644	0.304	0.662	0.145	0.456	0.216	0.469
	WPL	0.354	1.134	0.492	1.068	0.220	0.690	0.305	0.651	0.155	0.487	0.216	0.460
RMSE*100	SMM	0.239	0.982	0.364	1.010	0.151	0.621	0.229	0.639	0.108	0.438	0.162	0.452
	WMM	0.238	0.977	0.363	1.002	0.152	0.620	0.229	0.638	0.108	0.437	0.162	0.451
	WMME	0.324	1.011	0.481	1.037	0.206	0.647	0.304	0.662	0.146	0.456	0.216	0.469
	WPL	0.354	1.135	0.493	1.069	0.220	0.692	0.305	0.652	0.156	0.487	0.216	0.460
95%CP	SMM	0.87	0.95	0.86	0.95	0.86	0.96	0.84	0.94	0.85	0.94	0.86	0.93
	WMM	0.89	0.95	0.85	0.95	0.86	0.96	0.84	0.94	0.84	0.94	0.86	0.93
	WMME	0.97	0.95	0.95	0.95	0.95	0.96	0.93	0.95	0.95	0.94	0.96	0.95
	WPL	0.98	0.98	0.96	0.96	0.97	0.97	0.94	0.95	0.96	0.96	0.96	0.94

3.5 DISCUSSION ON THE FINDINGS OF NON-IGNORABLE MISSING LONGITUDINAL BIOMARKER DATA ANALYSIS

For longitudinal biomarker data with differential reasons for dropout and death, the correct specification of the full likelihood function and estimation of parameters requires infinite dimensional integrations. Weighted pseudo likelihood (WPL) method has been proposed for these types of longitudinal biomarker data. This WPL approach to the analysis of longitudinally measured biomarker data with differential reasons for dropout and death is a generalization of the weighted pseudo approach used by Lawless et al (1999). Weights were computed by inverting the probability of response which was originally used in differential sampling rate problems by Horvitz and Thompson. Recently Robins et al used IPW methods in the semi-parametric regression models. Our proposed methods can be implemented using existing statistical software. We have described and compared four models: SMM, WMM, WMME and WPL for analyzing longitudinal biomarker data with differential reasons for missing. Results of the WPL approach have been compared with the results of SMM, WMM and WMME via a real data analysis and a simulation study. The WPL approach, unlike the other weighted methods, accounts the fact that weights has been used as nuisance parameters (weights) in the estimation.

The simulation study suggests that the WPL approach performs reasonably well compared to the other standard and weighted approaches. It suffers from the fact that the estimation of the weights is taken into account when adjusting for the missingness, resulting in larger standard errors. The results obtained by WPL method are competitive with the results of WMME. For all methods the bias, SE and RMSE increase with the increase in missing observations. In terms of coverage probability the WMM is the worst performer and WPL is a competitor to the WMME.

4.0 ANALYSIS OF LEFT-CENSORED AND NON-IGNORABLE MISSING LONGITUDINAL BIOMARKER DATA USING WEIGHTED PSEUDO LIKELIHOOD THEORY

4.1 INTRODUCTION

Missing data is a persistent problem in longitudinal studies, presenting challenges at the analysis stage and resulting in a need for methodology to address these issues. The data may be missing due to subject drop out or death, failure to collect a subset of the data at follow up for administrative reasons, or missing due to censoring or truncation. This loss of data can result in biased estimates and a loss in precision. In addition, the relationship between the outcome and predictors may vary depending on the reason for missingness and failure to account for this in an analysis can affect the results.

The modeling of missing data has been an ongoing issue in the Genetic and Inflammatory Markers of Sepsis Study (GenIMS). The GenIMS study is a longitudinal cohort study of subjects with community acquired pneumonia that were recruited from 2001-2003 in 28 hospitals located in Pennsylvania, Connecticut, Michigan and Tennesse. One major goal of this study was to understand the role of inflammatory markers in the progression of pneumonia to sepsis (Kellum, et al, 2007). The biomarkers were collected daily throughout the first seven days of hospitalization and may be missing due to death in the hospital during the first seven days, discharge from the hospital before day 7 or for administrative reasons. In these settings it is likely that the pattern of any given biomarker will differ depending on the reason for the missing data. One of the markers of greatest interest in GenIMS was the pro-inflammatory marker IL-6; however, the measurement of IL-6 was limited due to the sensitivity of the assay resulting in left censoring of the measure at the lower limit of detection. Figure 4.1 presents plots of IL-6 over the seven days of measurement for three groups of subjects; those with complete observations over days 1 - 7, those who have incomplete observations due to hospital discharge and those who have incomplete observations due to death during the first seven days of hospitalization. This plot points to the differences in these groups and the need to account for this in the analysis as illustrated by the fact that subjects who died had the highest IL-6 levels during the study.

Many methods exist for the handling of missing data in longitudinal studies when the data are missing at random (MAR) and missing not at random (MNAR) including those based on imputation (Rubin, 1976; Rubin, 1996, Schafer, 1997) and those based on weighting (Robins, Rotnitzky and Zhao, 1995; Rotnitzky and Robins, 1997; Lipsitz, Ibrahim and Zhao, 1999; Lawless, Kalbfleisch and Wild, 1999; Dufouil, Brayne and Clayton, 2004; Fitzmaurice, et al., 2005; Ibrahim et al., 2005). Using the method of weighting, greater weight is placed on those observations that are less likely to be observed. The attraction of both imputation and weighting is that standard methods can be used for analysis. There are also likelihood-based methods, including selection, pattern mixture, and shared parameter models, (Little and Rubin, 2002; Hogan et al., 2004) for analyzing MNAR or non-ignorable missing longitudinal data. However, these likelihood based methods do not allow for differential reasons for missingness and censored observations to be included in the analysis. In addition, the estimation of the weights has not been studied in these likelihood based methods.

The inverse probability weighting (IPW) method has been used for the differential sampling rate problem to account for the fact that the data are not obtained from a random sample(Horvitz and Thompson, 1952). The idea of the IPW method is that if the probability of selecting a unit is π_i then the total π_i^{-1} units in the population should be used in the estimation. Recently, the IPW method has been used for handling dropouts or an under-represented response profile in non- or semi-parametric models (Robins and Rotnitzky, 1995). "The underlying idea is to base estimation on the observed responses but weight them to account for the probability of remaining in the study (Fitzmaurice et al, 2004)".

Since left-censored data arise in a variety of applications, there are many methods available to account for left-censoring in the outcome, with the tobit regression model being one of the first models developed for this problem (Tobin, 1958). Other approaches include the

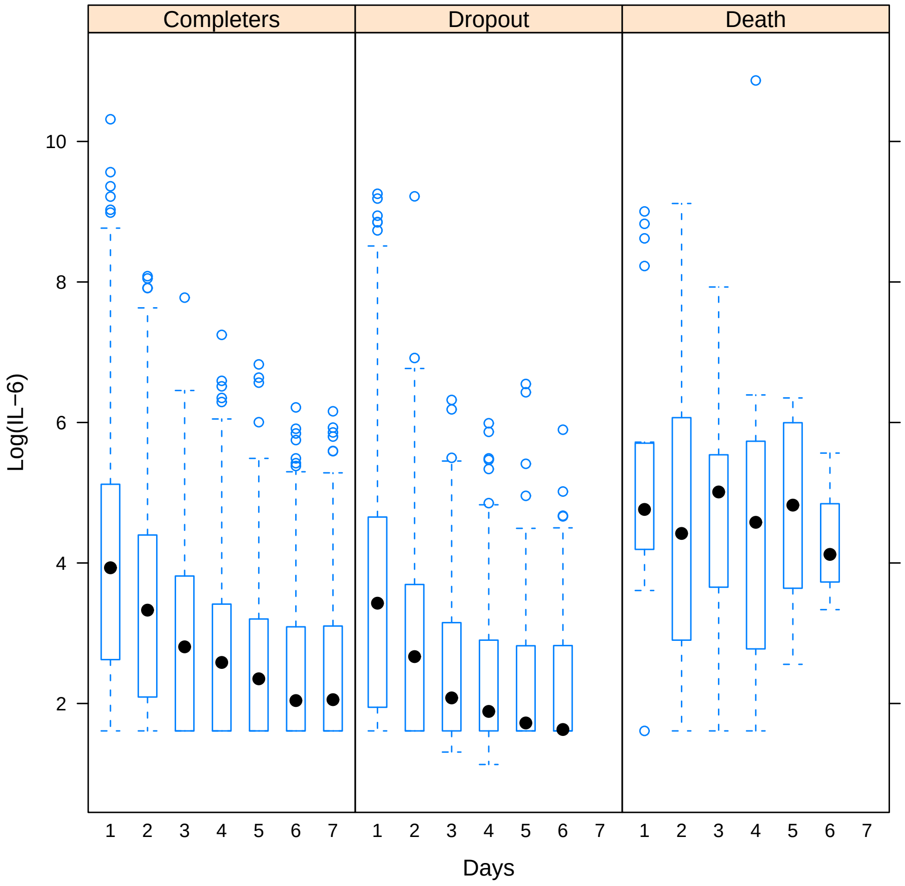

Figure 4.1: Comparison of IL-6 levels over 7 days period by the follow-up category.

use of variance components (Epstein et al, 2003), imputation of the quantification limit (Keet et al, 1997), using half of the lower limit of detection (O'Brien et al, 1998) and the use of random imputation procedures (Paxton et al, 1997). A more efficient approach to multiple imputation has also been proposed for analyzing censored longitudinal data using a linear mixed model (Hugues, 1999, Jacqmin-Gadda et al, 2000).

The goal of this work is to address the problem of differentially missing longitudinal data when the outcome is subject to left-censoring. We are proposing to extend the theory of tobit regression and the random effects tobit model (Tobin, 1958, Epstein et al, 2003) to develop a weighted random effects tobit (WRT) model for analyzing non-ignorable missing and left-censored longitudinal biomarker data. The performance of the WRT model will be compared with the random effects tobit (RT) model as well as weighted linear mixed models (Laird and Ware, 1982). In this setting, weighted linear mixed models (WMM) will be fit by replacing the censored values with the half of the detection limit and a randomly imputed value. In sample survey theory the weights have been assumed to be fixed and known. Here the IPW are computed from the observed data and hence their sampling variability will be taken into account in the inference (Little and Rubin, 2002, p. 53). We will also consider the IPW as nuisance parameters in the WRT model and use pseudo likelihood (PL) theory to account for the uncertainty associated with the estimation of an infinite number of nuisance parameters (Gong and Samaniego, 1981).

In section 4.2 we will describe the random effects tobit model and weighted random effects tobit model for analyzing non-ignorable missing and left-censored longitudinal biomarker data. In section 4.3 we will present an analysis of the IL-6 biomarker data from the GenIMS study. In section 4.4 we will demonstrate the performance of the proposed model using a simulation study. In the section 4.5 we will offer a discussion on the findings.

4.2 NOTATION AND WEIGHTED RANDOM EFFECTS TOBIT MODEL FORMULATION

Let Y_{ij} denote the measurement of a biomarker from the ith subject on the jth day of measurement at time $t_{ij}, i = 1, 2, ..., N, j = 1, 2, ..., n_i$ and $X_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijp})'$ denote a $p \times 1$ vector of covariates associated with Y_{ij}. In vector notation, $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ is the n_i-dimensional vector of biomarker measurements and $\mathbf{X}_i$ is the $n_i \times p$ matrix of covariates for the ith subject. Suppose the observed, censored and missing components of $\mathbf{Y}_i$ are denoted by $\mathbf{Y}_i^o$, $\mathbf{Y}_i^c$ and $\mathbf{Y}_i^m$ respectively. Define the missingness indicator vector as $\mathbf{R}_i = (R_{i1}, R_{i2}, ..., R_{in_i})'$, where R_{ij} =0 if Y_{ij} is observed, and =1 otherwise. Note, in this paper we will assume that $\mathbf{Y}_i$ and $\mathbf{Y}_{i'}$, $(i \neq i')$, are independent and that the covariate vector X_{ij} is fully observed.

Suppose we observe $Y_{ij} = Y_{ij}^o$ only if $Y_{ij}^o > c$ (a constant) and $Y_{ij} = c$ if $Y_{ij}^o \leq c$. In this scenario, we have censored observations, since we do not observe any Y_{ij} that is less than c. For the observations where $Y_{ij} = c$ all we know is that $Y_{ij}^o \leq c$, i.e., $Pr(Y_{ij} = c) = Pr(Y_{ij}^o \leq c)$. Under the assumption that the missing data mechanism is missing at random (MAR), the left-censored longitudinal biomarker (e.g., IL-6) data can be analyzed utilizing the following tobit model (Tobin, 1958, Amemiya, 1974):

$$\mathbf{Y}_i^o = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \qquad if \mathbf{Y}_i^o > c, \tag{4.1}$$

where $\boldsymbol{\epsilon}_i$ is the usual error vector and is assumed to be distributed as a multivariate normal with mean vector zero and variance covariance matrix $\boldsymbol{\Sigma}$. To estimate the parameters in model (1) we can use the maximum likelihood procedure. For the two sets of observations: (i) $Y_{ij} = Y_{ij}^o$ with $Y_{ij}^o > c$ we can write the density function as $\phi[(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})/\boldsymbol{\Sigma}]$ where $\phi(.)$ is the pdf of the standard multivariate normal distribution, and (ii) $Y_{ij} = c$ with $Y_{ij}^o \leq c$ having the following probability

$$Pr[\mathbf{Y}_i = c] = Pr[\mathbf{Y}_i^o \leq c] = Pr[\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \leq c] = Pr[\boldsymbol{\epsilon}_i \leq c - \mathbf{X}_i\boldsymbol{\beta}] = \Phi(c^o), \tag{4.2}$$

where $\Phi(.)$ is the cumulative density of the multivariate normal distribution and $c^o = (c - \mathbf{X}_i\boldsymbol{\beta})/\boldsymbol{\Sigma}$. From these probability specifications, the likelihood function for the parameters associated with model (1) is given by:

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{\mathbf{Y}_i > c} (2\pi)^{-\frac{N_1}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) \prod_{\mathbf{Y}_i \leq c^o} \Phi((c - \mathbf{X}_i\boldsymbol{\beta})/\boldsymbol{\Sigma}). \quad (4.3)$$

Without loss of generality we can assume that the first N_1 subjects have $Y_{ij} = Y_{ij}^0$ and that the remaining $N_0 = N - N_1$ have $Y_{ij} = c$. So the log-likelihood function can be written as

$$ln\mathsf{L} = -\frac{N_1}{2} ln(|2\pi\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{N_1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{i=N_1+1}^{N} ln[\Phi((c - \mathbf{X}_i\boldsymbol{\beta})/\boldsymbol{\Sigma})]. \quad (4.4)$$

Epstein et al. (2003) have developed a tobit variance components method in the linear mixed model (LMM) frame work, a popular model for fitting inherently unbalanced longitudinal continuous biomarker data. A LMM for the ith subject can be written in the following form (Laird and Ware 1982):

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad (4.5)$$

where the vector $\mathbf{Y}_i$ is distributed as multivariate normal with the following specification:

$$\mathbf{Y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i) \quad (4.6)$$

$$\mathbf{b}_i \sim N(0, \mathbf{D}), \quad (4.7)$$

where $\mathbf{X}_i$ is the $n_i \times p$ design matrix for the fixed effects, $\boldsymbol{\beta}$; $\mathbf{Z}_i$ is the $n_i \times q$ design matrix for the random effects, $\mathbf{b}_i$; and $\mathbf{D} = Cov(\mathbf{b}_i)$ and $\boldsymbol{\Sigma}_i$=$Cov(\mathbf{e}_i)$ are the covariance matrices of the random effects and errors respectively.

If there is no left-censoring in the measurements, then from (5), the marginal distribution of $\mathbf{Y}_i$ is normal with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_i$=$\boldsymbol{\Sigma}_i$+ $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$. So inferences for the fixed parameters $\boldsymbol{\beta}$ can be based on the following log-likelihood function:

$$\texttt{log}L(\boldsymbol{\beta}) = -\frac{N}{2}\texttt{log}2\pi - \frac{1}{2}\texttt{log}|\mathbf{V}| - \frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \quad (4.8)$$

If there is left-censoring in the biomarker measurements, the likelihood function for the random effects tobit (RT) model (Epstein et al, 2003) can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\eta}_i) = \prod_{\mathbf{Y}_i > c} f(\mathbf{y}_i^o | X_i, Z_i, \boldsymbol{\beta}, \boldsymbol{\eta}_i) \prod_{\mathbf{Y}_i \leq c} Pr(\mathbf{y}_i^c < c | X_i, Z_i, \boldsymbol{\beta}, \boldsymbol{\eta}_i), \tag{4.9}$$

where $\boldsymbol{\eta}_i$ denotes the i^{th} component of $\mathbf{V}_i$. Now, the conditional distribution of $\mathbf{Y}_i^c | \mathbf{Y}_i^o$ is multivariate normal (Johnson and Wichern, 2001):

$$\mathbf{Y}_i^c | \mathbf{Y}_i^o \sim N(\boldsymbol{\mu}_i^{c|o}, \mathbf{V}_i^{c|o}), \tag{4.10}$$

where the mean vector and covariance matrix can be written as

$$\begin{aligned} \boldsymbol{\mu}_i^{c|o} &= \mathbf{X}_i^c \boldsymbol{\beta} + \boldsymbol{\eta}_i^{co} \boldsymbol{\eta}_i^{oo-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \\ \mathbf{V}_i^{c|o} &= \boldsymbol{\eta}_i^c - \boldsymbol{\eta}_i^{co} \boldsymbol{\eta}_i^{oo-1} \boldsymbol{\eta}_i^{oc}. \end{aligned}$$

Using the above quantities, the likelihood function (9) can be re-written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\eta}_i) = & \prod_{\mathbf{Y}_i > c} \frac{1}{2\pi |\boldsymbol{\eta}^{oo}|^{1/2}} e^{-\frac{1}{2} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})^T \boldsymbol{\eta}_i^{oo-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})} \\ & \times \prod_{\mathbf{Y}_i \leq c} \int_{-\infty}^{c_{i1}} \int_{-\infty}^{c_{i2}} \cdots \int_{-\infty}^{c_{in_i}} \frac{1}{2\pi |\boldsymbol{\eta}^{c|o}|^{1/2}} e^{-\frac{1}{2} (\mathbf{u} - \boldsymbol{\mu}_i^{c|o})^T \boldsymbol{\eta}_i^{c|o^{-1}} (\mathbf{u} - \boldsymbol{\mu}_i^{c|o})} du. \end{aligned} \tag{4.11}$$

So the log likelihood function is

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\eta}_i) = & \sum_{\mathbf{Y}_i > c} [-log(2\pi) - \frac{1}{2} log |\boldsymbol{\eta}_i^{oo}| - \frac{1}{2} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})^T \boldsymbol{\eta}_i^{oo-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})] \\ & + log \prod_{\mathbf{Y}_i \leq c} \int_{-\infty}^{c_{i1}} \int_{-\infty}^{c_{i2}} \cdots \int_{-\infty}^{c_{in_i}} \frac{1}{2\pi |\boldsymbol{\eta}^{c|o}|^{1/2}} e^{-\frac{1}{2} (\mathbf{u} - \boldsymbol{\mu}_i^{c|o})^T \boldsymbol{\eta}_i^{c|o^{-1}} (\mathbf{u} - \boldsymbol{\mu}_i^{c|o})} du. \end{aligned} \tag{4.12}$$

When applying the pseudo maximum likelihood method, the likelihood function is maximized for the parameters of interest and all other parameters are treated as nuisance parameters (Gong and Samaniego, 1981). These nuisance parameters are replaced by their consistent estimates in the likelihood function. Treating $\boldsymbol{\eta}$ as a nuisance parameter vector, the pseudo log-likelihood and the corresponding score functions for the parameter vector $\boldsymbol{\beta}$ are as follows:

$$l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = l(\boldsymbol{\beta}, \boldsymbol{\eta})_{|\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}} \tag{4.13}$$

$$S(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}). \tag{4.14}$$

Equations (13) and (14) define the log-likelihood function and the score function corresponding to the RT model, respectively. If the missing data are ignorable, then the inferences can be based on the RT model's pseudo log-likelihood function (13) and its score equation (14). For the left-censored IL-6 longitudinal biomarker data subject to non-ignorable missingness, the likelihood function is complex (Little and Rubin, 2002). To simply this likelihood, the pseudo likelihood method is implemented with weighting to account for missing data (Lawless et al., 1999). These weights are incorporated by multiplying the pseudo log-likelihood function (eq. 13) and the score function (eq. 14) of the RT model by weights. Now, the weighted pseudo log-likelihood function and weighted score function of the RT model can be defined as,

$$l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\omega}}) = \sum_{i=1}^{N} R_i \hat{\boldsymbol{\omega}}_i l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}) \tag{4.15}$$

$$S(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\omega}}) = \sum_{i=1}^{N} R_i \hat{\boldsymbol{\omega}}_i \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}). \tag{4.16}$$

where $\boldsymbol{\omega}_i = \boldsymbol{\pi}_i^{-1} = (\sum_{h=1}^{H} p_{ih}\delta_{ih})^{-1}$; p_{ih} is the probability of observing the biomarker from the ith subject in dropout category h with $\delta_{ih} = 1$ if the ith subject belongs to category h and 0 otherwise. We assume that the estimate of p_{ih} will be obtained by a plausible estimation procedure and need not be obtained by the method of maximum likelihood. The probability of observing a biomarker is modeled with a generalized logit model (Hosmer and Lemeshow, 2000). The generalized logistic regression model will be fitted with covariate vector, $\mathbf{x}_i$, observed response vector prior to time t_j, $\mathbf{y}_{ij}^o$=(Y_{i1}^o, ..., $\mathrm{Y}_{i,j-1}^o$), and unobserved response, $\mathbf{y}_{ij}^m$ as described by Hogan and Laird (1997) in p. 263. Now, taking completers(0) as the reference category, and dropout(1) and death(2) as comparison categories, the logistic regression model can be written as

$$log\left(\frac{\pi_{hi}}{\pi_{0i}}\right) = g_h(\mathbf{x}_i^*) = \lambda_{h0} + \lambda_{h1}'\mathbf{x}_i + \lambda_{h2}'\mathbf{y}_{ij}^o + \lambda_{h3}\mathbf{y}_{ij}^m \qquad h = 1, 2. \tag{4.17}$$

It follows that the conditional probability given the covariate vector $\mathbf{x}_i^*$ in this three category model can be obtained by the following formula:

$$\pi_{hi} = \frac{exp(g_h(\mathbf{x}_i^*))}{\sum_{h=0}^{2} exp(g_h(\mathbf{x}_i^*))}. \tag{4.18}$$

The inference for the pseudo-likelihood estimates obtained from equation (16) is based on asymptotic theory. The PL asymptotic SE of the estimates accounts for the extra variability due to the use of the estimated weights, in the estimation process (Gong and Samaniego, 1981). The asymptotic covariance matrix for the parameter of interest will be obtained using the following formula (Lawless et al. 1999, p. 427):

$$Var(\hat{\boldsymbol{\beta}}) \cong A_{11} + A_{11}^{-1}\tilde{V}A_{11}^{-T}, \tag{4.19}$$

where

$$\tilde{A}_{11} = -\sum_{i=1}^{N} R_i\hat{\omega}_i \frac{\partial^2}{\partial\beta\partial\beta'} l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})|_{\beta=\hat{\beta}}. \tag{4.20}$$

Letting, $\tilde{\xi}_i = \frac{\partial}{\partial\beta} l(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}})$ and $\bar{\tilde{\xi}}^{(h)} = \frac{1}{n_h}\sum_{i\in D_h}\tilde{\xi}_i$, $\tilde{V}$ can be defined as

$$\tilde{V} = \sum_{i=1}^{N}(\boldsymbol{\omega}_i^2 - \boldsymbol{\omega}_i)\sum_{i\in D_h}(\tilde{\xi}_i - \bar{\tilde{\xi}}^{(h)})(\tilde{\xi}_i - \bar{\tilde{\xi}}^{(h)})^T. \tag{4.21}$$

Equations (15) and (16) are the log-likelihood function and score equation of the proposed weighted random effects tobit (WRT) model, respectively. The SE of the WRT model's parameter estimate will be computed using formulas (19) - (21).

For comparison, we will fit a weighted linear mixed model (WMM) to the non-ignorable missing and left-censored longitudinal data. In the WMM, the weights will be used as a remedy for the non-ignorable missingness, and the censored values will be replaced by half of the detection limit or a randomly imputed value. The score equation for the parameters of interest from the WMM will be as follows,

$$\frac{\partial \texttt{log} L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\mathbf{X}_i'\mathbf{V}_\omega^{-1}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{X}_i'\mathbf{V}_\omega^{-1}\mathbf{Y}_i, \tag{4.22}$$

where $Cov(\mathbf{Y}) = \mathbf{V}_\omega = \mathbf{ZGZ}' + \mathbf{LRL}$ with $\mathbf{L} = diag(\boldsymbol{\omega}^{-\frac{1}{2}})$. The parameters of the WMM will be estimated by the score equation (22) and the SEs of the estimated parameters will be computed by the variance formula in (19)-(21).

We will compare the performance of the proposed WRT model with a number of different models. Each of these models is denoted with a subscript on both the left and right side. The subscripts on the left side are a and r denoting the asymptotic SE from the PL theory and the robust SE, respectively. The subscripts on the right side are *1*, m, *7*, h, and r denoting the use of a single weights across the wave of measurements, mis-specified weights, seven(multiple) weights, half of the detection limit, and a randomly imputed value in the model respectively. The models are,

(i) ${}_r$RT: denotes a random effects tobit model with the SEs of the estimates computed using the sandwich estimator.

(ii) ${}_r$WRT$_1$: denotes a weighted random effects tobit model with the SEs of the estimates computed using the sandwich estimator.

(iii) ${}_a$WRT$_1$: denotes a weighted random effects tobit model with the SEs of the estimates computed using the asymptotic PL theory.

(iv) ${}_a$WRT$_m$: denotes a weighted random effects tobit model with the SEs of the estimates computed using the asymptotic PL theory. In this model, a small perturbation has been applied to the observed probabilities (0.10 added to the observed probabilities) and hence the weights are mis-specified.

(v) ${}_a$WRT$_7$: denotes a weighted random effects tobit model with the SEs of the estimates computed using asymptotic PL theory. In this model, multiple (seven) weights are computed for the multiple waves (seven days) of the IL-6 measurements.

(vi) ${}_a$WMM$_h$: denotes a weighted linear mixed model with the SEs of the estimates computed using the asymptotic PL theory. In this model censored values are replaced by half of the detection limit.

(vii) ${}_a$WMM$_r$: denotes a weighted linear mixed model with the SEs of the estimates computed using the asymptotic PL theory. In this model censored values are replaced by randomly imputed values.

4.3 APPLICATION: ANALYSIS OF LEFT-CENSORED AND NON-IGNORABLE MISSING LONGITUDINAL IL-6 BIOMARKER DATA

The GenIMS study was a large cohort study that was designed to gain an understanding into the role of both genetic and inflammatory biomarkers in the development of sepsis. The study focused on recruiting patients with community acquired pneumonia (CAP) to insure a relatively homogenous group of subjects, since sepsis can result from multiple illnesses. A total of 2320 patients were enrolled into the study through the emergency departments in 28 hospitals in Pennsylvania, Connecticut, Michigan, and Tennessee (2001-2003). The focus of this analysis is on the biomarkers that were measured as part of the study. One marker of inflammation that was obtained was interleukin-6 (IL-6), which is thought to be a pro-inflammatory maker. The IL-6 measurements were left-censored due to the limit of quantification and were measured daily during the first seven days of the hospitalization. Thus the IL-6 data could be missing due to death, discharge from the hospital before day 7 or for administrative reasons. Figure 4.1 presents the plot of IL-6 over time and indicates that the level of this biomarker depends on the reasons of missingness. This leads to non-ignorable missingness and points to the need to take this into account in the analysis. In this analysis, we have 330 subjects with complete IL-6 measurements, 699 subjects with incomplete IL-6 measurements due to hospital discharge and 19 subjects with incomplete data due to death during the first 7 days of hospitalization. We are analyzing the IL-6 data with non-ignorable missingness using the proposed WRT method that has been described in Section 2. In this weighted analysis the computed weights are obtained for the observed response to account for the non-ignorable missingness. The probability of an IL-6 value being missing is computed from a multinomial logistic regression model with dropout category as the outcome and the following covariates: logarithm of IL-6, steroid use, pneumonia severity index (PSI), acute physiology and chronic health evaluation (APACHE). The weights are then computed by inverting the probabilities estimated from this model. For this example, we fitted all of the models that have been described in section 2. The SE of the estimates has been computed using either the sandwich estimator or the asymptotic estimator obtained from the PL theory.

Table 4.1: Analysis of left-censored and non-ignorable missing longitudinal IL-6 biomarker data obtained from the GenIMS study

Covariate	Model	Estimate	S.E.	z-value	p-value
Day	${}_r$RT	-0.4134	0.0186	-22.24	<.0001
	${}_r$WRT$_1$	-0.3580	0.0216	-16.60	<.0001
	${}_a$WRT$_1$	-0.3580	0.0171	-20.89	<.0001
	${}_a$WRT$_m$	-0.4102	0.0480	-8.55	<.0001
	${}_a$WRT$_7$	-0.4063	0.2221	-1.83	0.0673
	${}_a$WMM$_h$	-0.3316	0.0134	-24.71	<.0001
	${}_a$WMM$_r$	-0.3434	0.0143	-24.02	<.0001
Race	${}_r$RT	0.1167	0.1123	1.04	0.2988
	${}_r$WRT$_1$	0.1426	0.1171	1.22	0.2237
	${}_a$WRT$_1$	0.1426	0.1126	1.27	0.2041
	${}_a$WRT$_m$	0.1613	0.1431	1.13	0.2585
	${}_a$WRT$_7$	0.0823	0.1146	0.72	0.4715
	${}_a$WMM$_h$	0.1163	0.0898	1.30	0.1936
	${}_a$WMM$_r$	0.1194	0.0927	1.29	0.1971
APACHE	${}_r$RT	0.0424	0.0040	10.57	<.0001
	${}_r$WRT$_1$	0.0349	0.0044	7.92	<.0001
	${}_a$WRT$_1$	0.0349	0.0049	7.06	<.0001
	${}_a$WRT$_m$	0.0390	0.0067	5.81	<.0001
	${}_a$WRT$_7$	0.0413	0.0140	2.94	0.0029
	${}_a$WMM$_h$	0.0350	0.0038	9.13	<.0001
	${}_a$WMM$_r$	0.0354	0.0039	9.02	<.0001

Table 4.1 presents the results obtained from the weighted random effects tobit (WRT) analysis of IL-6 biomarker data. The parameter estimates obtained from the WRT models are different from the estimates obtained from the RT model. The parameter estimates of intercept, day and race obtained from the weighted random effects tobit models (${}_r\text{WRT}_1$ and ${}_a\text{WRT}_1$) are larger than the estimates obtained from the random effects tobit model. The corresponding SE of the WRT model's parameter estimates is also larger when compared to the SE of the ${}_r\text{RT}$ estimates. For most of the covariates in the weighted linear mixed models (${}_a\text{WMM}_h$ or ${}_a\text{WMM}_r$) which are fitted by replacing the censored values with half of the detection limit or a randomly imputed value, parameter estimates and SEs are relatively smaller. Small perturbations of the computed probabilities, resulting in mis-specified weights, have little effect on the parameter estimates and SEs. Generally, the ${}_a\text{WRT}_7$ model estimates are smaller and it's SEs are larger, hence the corresponding z-values of this model are smaller resulting in some covariates being insignificant when compared to the other methods.

4.4 SIMULATION STUDY FOR LEFT-CENSORED AND NON-IGNORABLE MISSING LONGITUDINAL BIOMARKER DATA

To compare the performance of the proposed weighted random effects tobit (WRT) model with the random effects tobit (RT) model, we conducted a simulation study that was designed to address the following issues: the handling of the missing data through the estimation of the weights used in the model, the definition of the censored outcome, the estimation of the variance of the estimators, and the sensitivity of the model to mis-specification of the weights. We used several different approaches to compare methods for the estimation of the weights and to examine the sensitivity of the estimates to potential mis-specification. We compared the ${}_r\text{WRT}_1$, ${}_a\text{WRT}_1$, ${}_a\text{WRT}_m$, and ${}_a\text{WRT}_7$ models, with the left subscripts r and a indicating standard errors based on the sandwich estimator and asymptotic pseudo likelihood theory respectively. The right subscripts are 1 indicating that only one weight is used across the wave of measurements (constant weights), m indicating that the weights are misspecified and, 7 indicating that the weights are estimated for each of the 7 time points. The impact of the estimation of the variance was examined via the models ${}_{r}\mathrm{WRT}_1$ and ${}_{a}\mathrm{WRT}_1$, which are both weighted random tobit models with r denoting the sandwich estimator and a denoting the asymptotic variance based on the PL theory. Finally, the impact of ignoring the left censoring of the outcome was examined by replacing censored values either by half of the lower limit of detection or by a randomly imputed value with the models ${}_{a}\mathrm{WMM}_h$ and ${}_{a}\mathrm{WMM}_r$, respectively.

We generated outcome data from a multivariate normal (MVN) distribution with a specified mean vector and variance-covariance matrix. The mean vector and covariance matrix were obtained from the GenIMS data and all simulations are based on this underlying covariance structure. Based on these assumptions, the mean vector has the following form,

$$\boldsymbol{\mu}_{jh} = \alpha + \beta_0 day_j + \beta_1 APACHE_h,$$

where APACHE denotes a severity of illness measure that was used in the GenIMS study, j (day) = 1,..., 7 and h(dropout categories) = 0, 1, or 2. The true values of the parameters were set to β_0=-0.1 and β_1=0.07. The outcome variable was then censored based on a rate of 10%, 25% and 40% for the simulation studies. We created the missing patterns for the data and estimated the weights based on the multinomial logistic regression model as described in section 2 and equation (17):

$$\begin{aligned} log\left(\frac{\pi_{hi}}{\pi_{0i}}\right) &= g_h(\mathbf{x}_i^*) = \lambda_{h0} + \lambda_{h1} day_j + \lambda_{h2} APACHE_i + \lambda_{h3(1)} IL\text{–}6^o_{(1)} + \\ &\cdots + \lambda_{h3,(j-1)} IL\text{–}6^o_{(j-1)} + \lambda_{hj} IL\text{–}6^m_j, \qquad h = 1, 2;\ j > 1. \end{aligned}$$

All simulations were run with three different sample sizes of 1000, 500 and 200 subjects and a follow up period of 7 days. The results presented are based on 500 samples and three different designs. The first design, D1, is constructed so that 60% of the subjects have complete observations, that is, data are present for all 7 days, and the remaining 40% of the observations are subjects to missing due to drop out. For the second design, D2, 60% of the subjects have complete data, 30% of the subjects are missing due to drop out and the remaining 10% have missing data due to death. Design 3, D3, consists of 30% of subjects with complete data, 60% of the subjects missing due to drop out and the remaining 10%

missing due to death. The missingness across waves of time points was generated in the following manner, 5% of subjects were removed starting from the 2nd wave at all waves. Once a subject's data is set to missing all data at the remaining time points is also missing.

Table 4.2: Analysis of simulated left-censored and non-ignorable missing longitudinal IL-6 biomarker data with 60% completers and 40% dropouts(True value of the parameters: β_0=-0.1 and β_1=0.07)

		Sample Size, N=200				Sample Size, N=1000			
Censored	Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$
10%	${}_r$RT	-0.0183	0.0085	0.0818	0.1141	-0.0916	0.0683	0.0119	0.1856
	${}_r$WRT$_1$	-0.0183	0.0082	0.0818	0.1137	-0.0912	0.0653	0.0127	0.1826
	${}_a$WRT$_1$	-0.0183	0.0082	0.1030	0.8099	-0.0912	0.0653	0.0122	0.1934
	${}_a$WRT$_m$	-0.0157	0.0108	0.1166	0.7814	-0.0912	0.0654	0.0122	0.1932
	${}_a$WRT$_7$	-0.0885	0.0637	0.0327	0.4038	-0.0894	0.0661	0.0135	0.1935
	${}_a$WMM$_h$	-0.0179	0.0084	0.0996	0.7884	-0.0899	0.0664	0.0127	0.1932
	${}_a$WMM$_r$	-0.0180	0.0083	0.1081	0.8376	-0.0898	0.0668	0.0138	0.1961
40%	${}_r$RT	-0.0133	0.0083	0.0870	0.1145	-0.0678	0.0548	0.0351	0.1753
	${}_r$WRT$_1$	-0.0122	0.0074	0.0880	0.1132	-0.0622	0.0496	0.0403	0.1694
	${}_a$WRT$_1$	-0.0122	0.0074	0.1288	0.7817	-0.0622	0.0496	0.0401	0.1794
	${}_a$WRT$_m$	-0.0123	0.0074	0.1287	0.7780	-0.0623	0.0496	0.0400	0.1792
	${}_a$WRT$_7$	-0.0589	0.0487	0.0692	0.4069	-0.0612	0.0495	0.0412	0.1792
	${}_a$WMM$_h$	-0.0116	0.0067	0.1039	0.5406	-0.0596	0.0460	0.0414	0.1636
	${}_a$WMM$_r$	-0.0117	0.0070	0.1130	0.6273	-0.0594	0.0467	0.0420	0.1685

Table 4.2 presents the results obtained from design D1 where 60% of the subjects have complete observations. Results are presented for sample sizes of 200 and 1000 subjects with 10% and 40% of the outcomes subject to censoring. Note that the bias of the estimates is heavily dependent on the sample size with the estimates for a sample size of 200 being severely biased while the results for a sample size of 1000 indicate that the estimates are much closer to the true values. The bias is also substantially larger when comparing the scenarios with 10% censoring to those with 40% censoring. In all cases the estimates for ${}_a\mathrm{WRT}_7$ are the least biased and for some of the settings these are the only estimates that are close to the true values. While the results for 25% censoring and a sample size of 500 are not presented, the overall pattern was the same. For a sample size of 200 the results for the RMSE varied widely across the methods. For estimation of the β_0 term, ${}_a\mathrm{WRT}_7$ had the smallest RMSE followed by ${}_r\mathrm{RT}$ and ${}_r\mathrm{WRT}_1$ while the results were flipped for the β_1 term. For a sample size of 1000, the methods were comparable across the two censoring scenarios with RMSE increasing as the percentages of censoring increased.

In Table 4.3 we have presented the simulation results for design D2, where 30% of the subjects are missing due to drop out and 10% have missing data due to death. For N=200 and 10% censored observations, biases associated with all of the parameter estimates are very high with the exception of ${}_a\mathrm{WRT}_7$ model parameters. Again, the ${}_a\mathrm{WRT}_7$ model parameter estimates improve consistently as the sample size increases and the censoring percentage decreases. In most cases, the RMSE for the parameter estimates associated with the model ${}_a\mathrm{WRT}_7$ are the smallest among all weighted estimates based on the asymptotic SE of the estimates obtained from the PL theory. Simulation results obtained from design D3 (not presented), where 60% subjects are missing due to drop out and 10% are missing due to death, are similar to those obtained from the other two designs. Again, the biases of the parameter estimates are minimal for a sample size of N=1000 and the biases are large for a sample size of N=200.

When comparing the estimates across the three simulation designs (D1 with 40% dropouts; D2 with 30% dropouts and 10% deaths; D3 with 60% dropouts and 10% deaths) we found that the biases of the estimate of β_1 generally increased as the percentages of missing observations increased, while this was not the case for the term β_0. We also found that the RMSE for the parameter estimates obtained from the largest sample sizes considered (N=1000) were very similar across the designs while the results obtained for the smaller sample sizes were more variable. Specifically for sample sizes of 200 and 500, the RMSE increased as the percentage of censoring increased. When comparing the methods, models that accounted for the censoring through use of tobit regression performed better than models where censored observations were replaced by a randomly imputed value or a half of the lower limit of detection. Overall, the weighted random effects tobit model with multiple weights produced the best estimates for small and moderate sample sizes, even with large percentages of missing and censored observations, making this the method of choice for small and moderate sample sizes.

Table 4.3: Analysis of simulated left-censored and non-ignorable missing longitudinal IL-6 biomarker data with 60% completers, 30% dropouts and 10% deaths (True value of the parameters: β_1=-0.1 and β_2=0.07))

		Sample Size, N=200				Sample Size, N=1000			
Censored	Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$	$\hat{\beta}_0$	$\hat{\beta}_1$	$RMSE(\hat{\beta}_0)$	$RMSE(\hat{\beta}_1)$
10%	${}_r$RT	-0.0183	0.0182	0.0818	0.1225	-0.0912	0.0695	0.0121	0.1839
	${}_r$WRT$_1$	-0.0165	0.0173	0.0836	0.1216	-0.0826	0.0650	0.0194	0.1800
	${}_a$WRT$_1$	-0.0165	0.0173	0.0989	0.6109	-0.0826	0.0650	0.0195	0.1858
	${}_a$WRT$_m$	-0.0165	0.0172	0.0985	0.6009	-0.0827	0.0651	0.0194	0.1856
	${}_a$WRT$_7$	-0.0905	0.0889	0.0262	0.3027	-0.0906	0.0716	0.0130	0.1894
	${}_a$WMM$_h$	-0.0179	0.0179	0.0966	0.5728	-0.0895	0.0681	0.0135	0.1867
	${}_a$WMM$_r$	-0.0179	0.0179	0.1026	0.6050	-0.0895	0.0681	0.0145	0.1883
40%	${}_r$RT	-0.0141	0.0106	0.0861	0.1155	-0.0681	0.0534	0.0348	0.1710
	${}_r$WRT$_1$	-0.0132	0.0104	0.0870	0.1153	-0.0637	0.0525	0.0389	0.1701
	${}_a$WRT$_1$	-0.0132	0.0104	0.1179	0.4739	-0.0637	0.0525	0.0388	0.1717
	${}_a$WRT$_m$	-0.0133	0.0104	0.1171	0.4652	-0.0639	0.0527	0.0386	0.1717
	${}_a$WRT$_7$	-0.0667	0.0613	0.0489	0.2611	-0.0655	0.0507	0.0372	0.1698
	${}_a$WMM$_h$	-0.0123	0.0096	0.1016	0.3472	-0.0597	0.0463	0.0414	0.1579
	${}_a$WMM$_r$	-0.0124	0.0095	0.1079	0.3852	-0.0596	0.0466	0.0419	0.1606

4.5 DISCUSSION ON THE FINDINGS OF LEFT-CENSORED AND NON-IGNORABLE MISSING LONGITUDINAL BIOMARKER DATA ANALYSIS

Analyzing left-censored and non-ignorable missing longitudinal biomarker data is a challenge. For analyzing this type of data, we have proposed a weighted random effects tobit model. We have compared the proposed model with the random effects tobit model (RT), a standard model for analyzing left-censored longitudinal data. Using simulated data, four (4) WRT models have been fitted by considering various combinations of SE and weights computation. By replacing the censored values with a half of the detection limit or randomly imputed values, two linear mixed models have been fitted and compared with the WRT models. The weights are obtained by the IPW methodology, that is, the probabilities of observing a biomarker value have been computed and inverted. These weights are used with the observed data for recovering the contribution of missing observations in the analysis. In the estimation process, these weights have been treated as nuisance parameters. Our interest lies in the estimation of parameters of a weighted random effects tobit model. We have adjusted the effects of nuisance parameters in the estimation of the SEs of the parameter estimates of interest. Through the simulation study, we have compared the proposed WRT model with a random effects tobit model, mis-specified WRT model, and linear mixed models.

In the simulation study, we have simulated three (3) scenarios or designs with various percentages of missing and censored observations. Irrespective of the design, for small and moderate sample sizes, both estimates of the coefficients are biased unless they are estimated using the multiple weights model. Using the multiple weights model produces the smallest bias in the estimation of parameters for both small and moderate sample sizes. All estimates are very close to each other for large sample sizes (N=1000). Biases of the estimates increase as the percentage of missingness increases. Biases of the estimates also depend on the percentages of censored observations. For heavily censored data, the estimates are badly biased though estimates obtained from using the multiple weights model are relatively less biased. Mixed models and WRT models based on single weights produce similar estimates. The mis-specified model which has been constructed based on a small perturbation of the computed probabilities, gives very similar estimates to those based on single weight WRT and mixed models. From this simulation study we have seen that the multiple weight model's (${}_a\text{WRT}_7$) estimates are the least biased across the designs, sample sizes, and percentages of missing and censored observations.

For analyzing non-ignorable missing and left-censored longitudinal continuous biomarkers, we extend the theory of random effects tobit model. We propose the use of a multiple weights random effects tobit model for settings where data are subject to missingness for different reasons. We have corrected the SEs using the PL theory for the use of a large number of nuisance parameters (weights) in the estimation process. The proposed model works well even in the setting of small to moderate sample sizes. In addition, the estimates have the smallest bias and RMSE for small percentages of missing and censored observations.

5.0 CONCLUSIONS

There are two issues in analyzing longitudinal biomarker data that have been addressed in this endeavor. The first issue is the analysis of longitudinal biomarker data with dropout and death. Though there are some likelihood based methods for analyzing non-ignorable missing biomarker data, these methods have some concern on the issue of identifiability along with their requirement of a rigorous computational approach for implementation. To avoid this issue and for easy implementation in the standard software, we proposed weighted pseudo likelihood (WPL) methods for analyzing non-ignorable missing longitudinally measured biomarker data. The proposed method has been compared with a number of methods. We have tested the method by analyzing a real data set obtained from the GenIMS study and performed a simulation study. A standard method for analyzing longitudinal data is the standard linear mixed (SMM) model. Though the SMM model fit gives smaller biases, SE estimates, and RMSE estimates, it does not account the fact that the data are non-ignorable missing. The weighted linear mixed model (WMM) has been fitted with the intention to capture the missing data. Using this method, the SE of the estimate was not corrected due to the uncertainty of estimating weights. For comparison, we have fitted the WMM with the robust SE estimate. This SE estimate does not take into consideration the fact that the nuisance parameters have been used in the estimation process either. The proposed WPL method corrects the computed SE estimate. The three weighted methods produce same biases. The SE and RMSE of the WPL method estimate are competitive to the SE and RMSE of the WMME method estimate. The coverage probabilities for the regression coefficients estimates by the WMME and WPL methods vary according to the designs. No method is consistently performing over the other method across the designs.

The second issue for analyzing longitudinal biomarker data is the left-censoring. In addition to the non-ignorable missingness, left-censoring gives another non-trivial challenge to the analyst. Though there are some likelihood based methods for analyzing left-censored and/or non-ignorable missing longitudinal data, none of these methods are based on weighting techniques. We proposed a weighted random effect tobit (WRT) model based on weighted pseudo likelihood theory. This proposed model has been compared with the un-weighted random effects tobit model. Also for the comparison, several WRT models have been fitted either by varying the SE estimation procedure or by varying replacement methods for the censored observations. The real (IL-6 biomarker) data analysis shows that the z-values differs among the weighted methods. Again, only the SE estimates based on the WPL method account for the effect of nuisance parameters in the estimation process. From the simulation study it can be inferred that the WRT model using WPL theory and censored values replaced by the detection limit produce smallest root mean square error (RMSE). Simulation study also indicates that, instead of using a single weight across the panels, use of multiple weights produce smallest RMSE. Therefore, for analyzing non-ignorable missing and left-censored longitudinal biomarker data, a WRT model based on WPL theory with censored values replaced by the detection limit and use of multiple weights would be recommended.

BIBLIOGRAPHY

[1] A. Agresti. *Categorical Data Analysis.* Wiley-Interscience, 2nd ed. edition, 2004.

[2] P.D. Allison. *Missing Data.* A SAGE University paper, 2002.

[3] T. Amemiya. Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica*, 42:999–1012, 1984.

[4] S. Beal. Ways to fit a pk model with some data below the quantification limit. *J. Pharmacokinet. Pharmacodyn.*, 28:481–504, 2001.

[5] M.J. Daniels and J.W. Hogan. Reparameterization of the pattern mixture model for sensitivity analysis under informative dropout. *Biometrics*, 66:12411248, 2000.

[6] H. Demirtas. Assessment of relative improvement due to weights within generalized estimating equations framework for incomplete clinical trials d. *Journal Of Biopharmaceutical Statistics*, 14:10851098, 2004.

[7] C. Dufouil, C. Brayne, and D. Clayton. Analysis of longitudinal studies with death and drop-out: a case study. *Statist. Med.*, 23:22152226, 2004.

[8] M.P. Epestin, X. Lin, and M. Boehnke. A tobit variance component methods for linkage analysis of censored trait data. *Am. J. Hum. Genn.*, 72:611–620, 2003.

[9] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. Applied longitudinal analysis. *Wiley.*, 2004.

[10] G.M. Fitzmaurice, S.R. Lipsitz, G. Molenberghs, and J.G. Ibrahim. A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *J. R. Statist. Soc. A*, 168:723735, 2005.

[11] G.S. Ghebregiorgis. *Modeling and analyzing multivariate longitudinal left-censored biomarker data.* Ph.d thesis, Arts and Science, University of Pittsburgh, 2008.

[12] G. Gong and F.J. Samaniego. Pseudo maximum likelihood estimation: theory and applications. *Annals of Statistics*, 9:861–869, 1981.

[13] C. Han and R. Kronmal. Box-cox transformation of left-censored data with application to the analysis of coronary artery calcifcation and pharmacokinetic data. *Statistics in Medicine*, 23:36713679, 2004.

[14] J.W. Hogan and N.M. Laird. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, 16:259–272, 1997.

[15] J.W. Hogan, J. Roy, and C. Korkontzelou. Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statist. Med.*, 23:14551497, 2004.

[16] B.E. Honore. Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometnca*, 60:533–565, 1992.

[17] D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *J. Amer. Stastist. Assoc.*, 47:663–685, 1952.

[18] D.W. Hosmer and L. Lemeshow. *Applied Logistic Regression.* Wiley, 2nd edition, 2000.

[19] J. Huges. Mixed effects models with censored data with applications to hiv rna levels. *Biometrics*, 55:625–629, 1999.

[20] J.G. Ibrahim, S.R. Chen, M. Lipsitz, and A.H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100:332–346, 2005.

[21] H. Jacqmin-Gadda, R. Thibaut, G. Chne, and D. Commenges. Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, 6:1:355368, 2000.

[22] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, 5th edition, 2002.

[23] I.P.M. Keet, M. Janssen, P. J. Veugelers, F. Miedema, M.R. Klein, J. Goudsmit, R.A. Coutinho, and F. Wolf. Longitudinal analysis of cd4 t cell counts, t cell reactivity, and human immunodeficiency virus type 1 rna levels in persons remaining aids-free despite cd4 cell counts .200 for .5 years. *The Journal of Infectious Diseases*, 176:66571, 1997.

[24] J.A. Kellum, L. Kong, M.P. Fink, LA. Weissfeld, D.M. Yealy, M.R. Pinsky, J. Fine, A. Krichevsky, R.L. Delude, and D.C. Angus. Understanding the inflammatory cytokine response in pneumonia and sepsis. *Crit Care Med.*, 35(4):1061–1067, 2007.

[25] M.G. Kenward. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statist. Med.*, 17:2723 2732, 1998.

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

[27] J.F. Lawless, J.D. Kalbfleisch, and C.J. Wild. Semiparametric methods for response-selective and missing data problems in regression. *J. R. Statist. Soc. B*, 61:413–438, 1999.

[28] D.Y. Lin. Regression analysis of incomplete medical cost data. *Statistics in Medicine*, 22:1181–1200, 2003.

[29] H. Lin, D. Scharfstein, and R. Rosenheck. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of Royal Statistical Society, B*, 66(3):791–813, 2004.

[30] S.R. Lipsitz, J.G. Ibrahim, and L.P. Zhao. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94:1147–1160, 1999.

[31] RJA. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125 134, 1993.

[32] RJA. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81:471 483, 1994.

[33] RJA. Little and D.B. Rubin. Statistical analysis with missing data. *New York, Wiley*, 2002.

[34] RJA. Little and Y. Wang. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, 52:98111, 1996.

[35] R.H Lyles, C.M. Lyles, and J.T. Douglus. Random regression model for human immunodeficiency virous ribonucleic acid data subject to left-censoring and informative dropouts. *Appl. Statist.*, 49:485–497, 2000.

[36] T.R. O'Brien, P.S. Rosenberg, F. Yellin, and J.J. Goedert. Longitudinal hiv-1 rna levels in a cohort of homosexual men. *J. AIDS.*, 18:155–161, 1998.

[37] Y. Pawitan. *In All Likelihood: Statistical Modeling and Inference Using Likelihood.* Oxford Science Publications, 2001.

[38] W. Paxton, R. Coombs, J. McElrath, M. Keefers, F. Sinangil, B. Williams, D. Chernok, J. Hughes, and L. Corey. Longitudinal analysis of quantitative virologic measures in hiv-1 infected individuals with greater than 400 cd4+ cells/microliter. *J. Infect. Dis.*, 175:274–254, 1997.

[39] E.P. Pulkstenis, T.R. Ten Have, and Landis J.R. Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication. *Journal of the American Statistical Association*, 93:438 450, 1998.

[40] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846866, 1994.

[41] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106 121, 1995.

[42] A. Rotnitzky and J. Robins. Analysis of semi-parametric regression models with non-ignorable non-respons. *Statistics in Medicine*, 16:81–102, 1997.

[43] J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London, 1997.

[44] C. Shen and L. Weissfeld. Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. *Biostatistics*, 6:333347, 2005.

[45] StataCorp. *Longitudinal/Panel data.* Stata, College Station, TX, relase 10 edition, 2007.

[46] S.K. Thompson. *Sampling.* Wiley-Interscience, 2nd ed. edition, 2002.

[47] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

[48] R Wolfinger and M. O'Connell. Generalized linear mixed models: A pseudo-likelihood approach. *J. Statist. Comput. Simul.*, 48:233–243, 1993.