MULTIPLE IMPUTATION AND QUANTILE REGRESSION METHODS FOR BIOMARKER DATA SUBJECT TO DETECTION LIMITS

by

MinJae Lee

BS in Statistics, Sookmyung Women's University, Korea, 2000

BS in Computer Science, Sookmyung Women's University, Korea, 2000

MS in Statistics, Sookmyung Women's University, Korea, 2004

Submitted to the Graduate Faculty of
the Department of Biostatistics in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS

This dissertation was presented

by

MinJae Lee

It was defended on

August 31, 2010

and approved by

Lan Kong, Ph.D., Assistant Professor,

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Lisa Weissfeld, Ph.D., 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

Sachin Yende, M.D., M.S., Assistant Professor,

Department of Critical Care Medicine, School of Medicine, University of Pittsburgh

Dissertation Director: **Lan Kong, Ph.D.**, Assistant Professor,

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

MULTIPLE IMPUTATION AND QUANTILE REGRESSION METHODS FOR BIOMARKER DATA SUBJECT TO DETECTION LIMITS

MinJae Lee, PhD

University of Pittsburgh, 2010

Biomarkers are increasingly used in biomedical studies to better understand the natural history and development of a disease, identify the patients at high-risk and guide the therapeutic strategies for intervention. However, the measurement of these markers is often limited by the sensitivity of the given assay, resulting in data that are censored either at the lower limit or upper limit of detection. Ignoring censoring issue in any analysis may lead to the biased results.

For a regression analysis where multiple censored biomarkers are included as predictors, we develop multiple imputation methods based on Gibbs sampling approach. The simulation study shows that our method significantly reduces the estimation bias as compared to the other simple imputation methods when the correlation between markers is high or the censoring proportion is high.

The likelihood based mean regression for repeatedly measured biomarkers often assume a multivariate normal distribution that may not hold for biomarker data even after transformations. We consider a robust alternative, median regression, for censored longitudinal data. We develop an estimating equation approach that can incorporate the serial correlations between repeated measurements. We conduct simulation studies to evaluate the proposed estimators and compare median regression model with the mixed models under various specifications of distributions and covariance structures.

Missing data is a common problem with longitudinal study. Under the assumptions that the missing pattern is monotonic and the missingness may only depend on the observed data, we propose a weighted estimating equation approach for the censored quantile regression models. The contribution of each individual to the estimating equation is weighted by the inverse probability of dropout at the given occasion. The resultant regression estimators are consistent when the dropout process is correctly specified. The performance of our estimating procedure is evaluated via simulation study.

We illustrate all the proposed methods using the biomarker data of the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. Appropriate handling of censored data in biomarker analysis is of public health importance because it will improve the understanding of the biological mechanisms of the underlying disease and aid in the successful development of future effective treatments.

Keywords: Left-censored data; Detection limits; Multiple imputation; Gibbs sampler; Median regression; Quantile regression; Longitudinal data; Drop-out.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Biomarkers are increasingly used in biomedical studies for diagnosis and prognosis of acute and chronic diseases, gaining insight of treatment effectiveness and establishing the potential disease pathways to guide the future treatment targets. The accuracy of biomarker measurements is very important for making valid and reliable conclusions of the findings. However, the biomarker data are subject to various sources of measurement errors. This includes error associated with specimen collection, processing, and storage; laboratory error (both within and between batches); and variability in the biomarker levels over time within an individual. Left-censoring, which is due to the limits of detection (LOD), is also a common source of error that may not be noticed in the analysis stage of biomarker data. LOD is the lowest concentration of analyte in a sample that can be detected under the stated experimental conditions. Reportable or quantifiable measurements are only those above LOD. For example, left-censoring occurs in the assessment of viral RNA in patients infected with the human immunodeficiency virus (HIV) (Hughes [29]) , antibody concentration in blood serum (Moulton et al. [46]), and the interleukin-10 (IL10) and IL6 in community-acquired pneumonia (CAP) patients (Kellum et al. [35]). Despite the improvement in assay sensitivity, left-censoring remains a critical issue in some studies because the ultra sensitive assays may be too expensive to use in a large cohort study or the biomarker concentrations are much lower than expected. Left-censoring problem results in many challenges in the statistical analysis of biomarker data. Ignoring the censored observations or replacing them with an arbitrary constant will introduce biased results in the analysis. It is imperative to develop statistical methods to address this issue appropriately and efficiently.

We encountered the left-censoring data from an inception cohort study of sepsis: the Genetic and Inflammatory Markers of Sepsis (GenIMS) study [35]. GenIMS study is a multicenter study of 2320 subjects from the emergency department (ED) with community-acquired pneumonia (CAP), the most common cause of sepsis. It was coordinated by the CRISMA Laboratory, Department of Critical Care Medicine at the University of Pittsburgh between 2001 and 2003. Sepsis is the leading cause of death in critically ill patients in the United States (Hotchkiss et al. [54]). Sepsis is defined as a systemic inflammatory response syndrome that occurs during infection (Bone et al. [7]). The frequency of sepsis is expected to increase given an aging population and increasing number of patients with chronic treatment-resistant infections. Better understanding of the pathophysiology of sepsis from the inflammatory, procoagulant and immunosuppressive aspects has contributed to the emergence of several therapeutic plans (Russell [59]). The treatments can be very effective if applied expeditiously, however rapid diagnosis of sepsis remains to be difficult. Many people are pressing to identify the important biomarkers for risk prediction of sepsis and the subsequent adverse outcomes. Meanwhile, gaining more insights of the roles of different pathways through more sophisticated modeling may reveal novel targets and new mechanisms of action that help to improve the current treatment. The main goals of GenIMS study are to better understand the natural history and development of sepsis, and to identify the biomarkers indicating the risk for severe sepsis, multiple organ failure, and death. A set of inflammatory and coagulation markers were evaluated repeatedly during the course of hospitalization. However, the assays used were not sensitive enough to measure low concentrations of some biomarkers, resulting in heavy left-censoring data. Figure 1 presents the censoring proportion of cytokines (TNF (tumor necrosis factor), IL6 (interleukin-6) and IL10) over the first seven days. Since left-censoring introduces informative missing data that are not ignorable, the traditional regression analyses are no longer valid.

The ad-hoc methods using either the LOD, or some other arbitrary value such as the LOD/2, or the LOD/$\sqrt{2}$, usually lead to biased results towards null. The simple "fill-in" imputation approaches based on certain distributional assumptions of the data eliminate the bias when the censoring is moderate ($< 30\%$), but introduce the biased variance estimates. The multiple imputation approaches can often provide valid statistical inference when the censoring proportion is not high ($< 50\%$) [32]. Although these simple methods have limitations, they are easy to use in practice when censored biomarkers are either considered

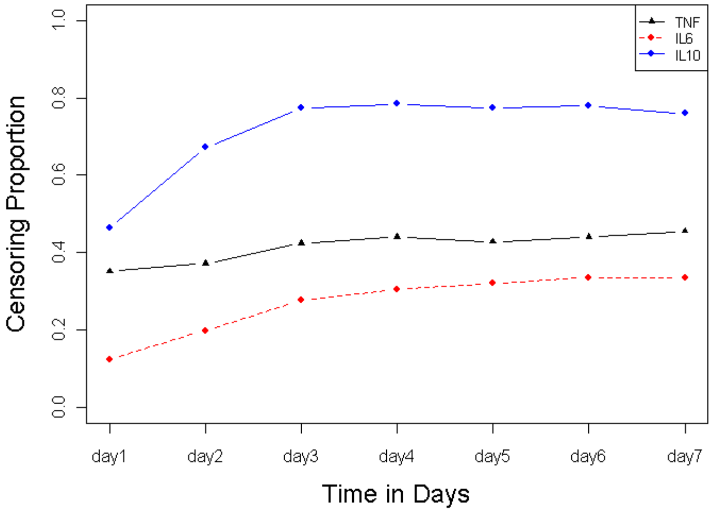

Figure 1: Daily Censoring Proportions of GenIMS Cytokines in the 1st week of Hospitalization

as predictors or dependent variables, also they aid in the graphic display of the raw data. More efficient statistical models have also been developed for handling the left-censoring data. To accommodate single censored covariates, Lynn [43] presented the likelihood-based approaches for linear and logistic models; Rigobon and Stoker [52] proposed a two-part regression model for both parametric and semiparametric models, where an index model was used to estimate the censored covariates. The same idea was adopted by D'Angelo and Weissfeld [10] to handle the censored covariates in the Cox model.

Much of the work for left-censored data has focused on censored response variables. The likelihood based approaches requires specification of the distribution of the response variables. The regression parameters can be estimated by maximizing a likelihood function contributed by both observed and censored observations. For example, the Normal distribution is assumed in the Tobit linear regression model (Tobin [72]) and mixed models for longitudinal censored data (Hughes [29]; Lyles et al. [42]; Jacqmin-Gadda et al. [31]; Wu [75]; Thiebaut et al. [70]). However the biomarker data are often highly skewed, even after log transformation. In this case, the quantile regression (QR) is more attractive than the least squares regression or any likelihood based method. Without imposing a full distribution on the response variable, QR provides a robust alternative and also allows one to look at covariate effects on various quantiles of a response variable. Most work on QR was done for independent data in econometrics. Powell [50] first considered the least absolute deviations (LAD) estimation which leads to the median regression estimator. LAD estimation method was later extended to more general quantiles (Powell [51]). Little work has been done for dependent observations until recently when Wang and Fygenson [73] proposed an inference procedure for longitudinal studies with application to a HIV/AIDS study.

As multiple biomarkers were measured over time from different potential pathways in GenIMS study, left-censoring data from multiple markers results in many new challenges in statistical analysis. This dissertation will focus on the following three statistical issues:

1. **Multiple Censored Predictors**

   In order to identify the important predictors of sepsis or mortality at 90 days, multiple cytokine and coagulation markers along with clinical predictors are considered in the logistic regression model. However, while several markers had various amounts of censoring data, the existing methods are difficult to use due to computational intensity. We will propose a multiple imputation (MI) approach for this problem that can account for the correlations between markers. Our approach is based on the use of Tobit regression and Gibbs sampling.

2. **Longitudinal Censored Responses**

   Mixed models and GEE methods are two popular approaches for analyzing longitudinal data. But only mixed models have been extended for the left-censoring responses as mentioned earlier. Mixed models may lead to biased results when the covariance structure of responses is misspecified. GEE methods overcome this problem, but require data to be missing completely at random. This is a stronger assumption than that required for the mixed model. The QR model is close to the GEE approach in that the estimating equations are formed to obtain the regression coefficient estimators and both are robust to misspecification of variance structure. Extending the approach of Wang and Fygenson [73], we will develop the estimating procedures to modeling the censored marker data while accounting for serial correlations between repeated measurements.

3. **Missing Data due to Dropouts**

   The biomarker data of GenIMS study were collected during the course of hospitalization. The missing data arose due to death or discharge early and administrative errors. The dominant missing pattern is monotone missing associated with dropout. The dropouts here refer to the cases that the patients had marker measurements up to a certain day and then no more measurements were collected afterwards because they were discharged early or died in the hospital. In GenIMS study, the subjects who had a lower level of IL6 appeared more likely to drop out. The inverse probability weighed GEE approach of Robins et al. [53] has been commonly used to handle the informative missing data for mean regression models. Assuming monotone missing data patterns and MAR mechanism, Lipsitz et al. [36] and Yi and He [77] adopted this approach to quantile regression models for uncensored data. The basic idea of this approach is that an individual's contribution to the traditional estimating equations is weighted by the inverse probability of dropout at the given occasion. We will apply the same weighting technique to the censored quantile regression estimating equation.

The organization of this dissertation is as follows. First, we review the literature work on missing data problem, existing methods for left-censored data and quantile regression methods in Chapter 2. Then we propose statistical methods to address the above three issues. In chapter 3, we present the multiple imputation methods based on gibbs sampling for multiple censored predictors. In chapter 4, we develop the median regression model to handle the left-censored longitudinal responses. In chapter 5, we propose the weighted censored quantile regression to incorporate the missing data due to dropouts. In chapter 6, we summarize the dissertation work and discuss some future work.

2.0 LITERATURE REVIEW

Left-censoring data is a special form of missing data. Rubin and Little [39] developed a terminology for different missing values processes. Three missing data mechanisms defined by Little and Rubin are: MCAR (missing completely at random), MAR (missing at random), and NMAR (not missing at random). There are several reasons why the data may be missing; they may be missing because equipment malfunctioned, the weather was terrible, or people got sick, or the data were not entered correctly. When we say that data are missing completely at random, we mean that the probability that an observation (X_i) is missing is unrelated to the value of X_i or to the value of any other variables. Often data are not missing completely at random, but they may be classifiable as missing at random (MAR). The data can be considered as missing at random if the data meet the requirement that missingness does not depend on the value of X_i after controlling for another variable. If data are not missing at random or completely at random then they are classed as Not Missing at Random (NMAR). For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random. Left-censoring data due to LOD are non-ignorable missing data, i.e., NMAR. Usually the statistical inference for NMAR data is complicated and the knowledge of missing mechanism is critical for valid estimation and inference. Left-censoring is a fixed censoring that is different from random censoring as seen in the survival analysis. The missing mechanism is clearly known and thus the modeling is less difficult than the general NMAR setup. The statistical framework for handling missing data can be adopted to left-censoring data with appropriate adjustment. In this chapter, we will first describe some methods to handle NMAR data in missing data literature, then we will review the existing methods for left-censoring data including naive approaches and efficient statistical approaches. We will also discuss briefly the statistical methods for handling missing data due to dropouts.

2.1 STATISTICAL METHODS FOR MISSING DATA

Statistical inference from missing data has been a researched topic over the last two decades. Since most statistical methods were derived for fully observed data, the impact of missing values is an issue. Missing values can occur on response variables (outcomes) and on covariates (predictors). The goal of a statistician still remains the same with or without missing values and they are required to draw valid and efficient inferences about its population of interest. Rubin and Little [59] developed a terminology for different missing values processes. Three missing data mechanisms defined by Little and Rubin are: MCAR, MAR, and NMAR. In this study, we focus on the methods for NMAR data.

2.1.1 Missing Covariates

Rubin [57] developed a framework of inference from missing data that remains in use today. There are the approaches for the missing covariates in regression models (Little, JASA [38]) and Little [37] provided a review of methods that can handle missing covariates into six classes. The first class is the complete-case (CC) analysis and the second is the available-case (AC) methods. Available-case analysis approaches use the largest sets of available cases for each of parameters (Little and Rubin [39]). The problem of this AC analysis is that the estimated covariance matrix of the X's is not necessarily positive definite, which leads to inferior results compared to CC analysis for highly correlated data (Haitovsky [22]). The third method discussed by Little includes Lest Squares (LS) on imputed data methods. In this setting, the missing covariates are imputed and a regression of response variable on covariates is performed on the filled in data by ordinary least squares or weighted least squares regression. The imputation methods were unconditional and conditional mean imputation. The inference based on these methods lead to the biased and imprecise results. The fourth class is the class of maximum likelihood (ML) methods. In this method, a classical ML estimate for a model for the joint distribution of Y and X would be the multivariate normal with mean μ and covariance matrix Σ. Another method would be Expectation Maximization (EM) algorithm. The EM algorithm (Dempster, Laird and Rubin [12]) is a general iterative algorithm for ML estimation in data with missing values. The EM algorithm consists of Expectation (E-step) and Maximization (M-step). The E-step provides the conditional expectation of the complete data log-likelihood given the observed data and current estimated parameters and then substitutes these expectations for the missing data. In the M-step, we can have a maximum likelihood estimation of the parameters as if there were no missing data. The M-step gives the parameter estimates to maximize the complete data log-likelihood from the E-step. The fifth class in Little's paper was the Bayesian methods. In Bayesian inference, information about unknown parameters is expressed in the form of a posterior distribution. Markov Chain Monte Carlo (MCMC) has been applied as a method for exploring posterior distributions in Bayesian inference. In MCMC, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a common, stationary distribution. By repeatedly simulating steps of the chain, it simulates draws from the distribution of interest. The last class considered by Little has the multiple imputation (MI) methods. Rubin [58] introduced the idea of MI in which each missing value is replaced with M times using simulated values prior to analysis. This will produce M possible complete datasets that are the analyzed in the same manner as a complete dataset. For the inference, these results are then combined by simple arithmetic to obtain overall estimates along with their standard errors that reflect missing data uncertainty as well as the sample variation.

2.1.2 Missing Response Variables

If the probability of missingness is associated with the unobserved response values that should have been obtained, the missing data mechanism is said to be not missing at random (NMAR). 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 [40]; Fitzmaurice et al. [17]; Allison [1]). 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. [17]; Little and Rubin [40]). In this scenario, the standard likelihoods for analyzing longitudinal biomarker data do not include a mechanism for incorporating different reasons for loss to follow-up or death. When the 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. [13]). 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. [53]). The weighting procedure has been applied in analyzing many incompleted longitudinal data problems by Rotnitzky and Robins [55], Lipsitz et al. [36], Lin, Demirtas [11], Dufouil et al. [13], Ibrahim et al. [30]. Weights are computed by inverting the probabilities of response. In a longitudinal study some subjects are more likely to complete the study than others. The pseudo-likelihood approach has been used for estimating parameters in generalized linear mixed models (Wolfinger and O'connell [74]) and also non-ignorable missing covariate in cox model was proposed by Herring and Ibrahim [26].

2.2 EXISTING METHODS FOR LEFT-CENSORING DATA

Several approaches have been proposed in the statistics literature for the analysis of longitudinal left-censored data and all approaches differ in sophistication when handling the truncated values.

2.2.1 Naive and Imputation Approaches

Naive approaches are to use only observed data or replace censored observations with a single value i.e., Limit of Detection (LOD) (Keet et al. [34]), LOD/2 or LOD/$\sqrt{2}$ (O'Brien et al.; Hornung and Reed [28]). If the distribution of the measurement data is known, then an alternative strategy replaces values below the detection limit with expected values of the missing measurements, conditional on being less than the detection limit (Garland et al. [18]; Gleit [21]). Calculation of the conditional expected value requires the investigator to either know or estimate parameters of the measurement distribution. The substitution schemes are simple but because a single value represents all measurements below the detection limit, parameter estimates and their variances are likely biased. this limitation led to a single-impute "fill-in" method (Helsel [25]; Moschandreas et al. [44, 45]). An investigator first characterizes the form of the distribution and estimates its parameters and then assigns randomly sampled values below the detection limit from the estimated distribution. Fill-in values along with measured values above the detection limit are then used in analyses. The fill-in method did not include complex modeling of regression factors. In addition, although the fill-in approach assigned random values from an appropriate distribution, it did not account for the variability of the imputation process, because the inserted values are not real data. For the imputation of the values of the detection limit, single imputation and multiple imputation (Lubin et al. [32]; De Roos et al. [4]) have been considered. Because different methods are needed to handle censored dependent variables and censored covariates, there are different approaches for these cases.

2.2.2 Likelihood Based Approaches

Especially, for the censored dependent variables, there are many likelihood based approaches where the distribution is fully specified. A standard method for the analysis of censored data is Tobit regression. Tobit regression based on normal assumption (Gilbert [20]; Persson & Rootzen [49]; Tobin [72]). Tobit regression has been extended to multivariate regression (Amemiya [2]). Recently a Box-Cox transformation has been used for the analysis of left-censored cross sectional data (Han and Kronmal [23]). Linkage analysis of left-censored trait data has been based on the variance component of the Tobit model (Epstein et al. [15]). In this model, the traditional variance component model has been modified to accommodate the censored data by random effects. The mixed models for the longitudinal censored data have been proposed. Hughes [29] modified the usual EM estimation procedure for the mixed effects model to account for left-censoring. The method uses Monte Carlo procedure to provide a general solution can be used with left-censored data and since the expectation step of the EM algorithm is intractable, the Gibbs sampler is used to implement a Monte Carlo expectation step in the EM algorithm. Jacqmin-Gadda et al. [31] proposed an approach of direct maximization of the likelihood without the EM or the Monte Carlo methods where maximization is based on the Marquardt algorithm. The Lyles et al. [42] approach was based on a hierarchical formulation of the likelihood, and estimation was carried out by direct maximization of the likelihood using built-in algorithms. Lyles et al. [42] analyzed left-censored longitudinal data with informative dropout HIV data by maximizing a single likelihood function which has integrated the censoring and informative dropout process. They estimated the parameters from this complicated likelihood function and compute the standard errors using the observed information matrix directly. Thiebaut et al. [70] considered joint modeling for bivariate longitudinal data. To accommodate single censored covariates, Lynn [43] presented the likelihood-based approaches for linear and logistic models, two-step linear regression model with index model used to estimate censored/selected covariates (Rigobon and Stoker [52]) and left-censored covariates in cox model (D'Angelo and Weissfeld [10]) have been proposed. 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.

2.2.3 Quantile Regression Approaches

In addition to these likelihood based approaches, the quantile regression approaches have been developed for left-censoring data. The quantile regression (QR) methods have been well developed for independent and longitudinal data when there is no censoring issue. For fixed censoring, i.e., the observations are censored at a fixed constant, and most work on QR was done for independent data in econometrics. Powell [50] first considered the least absolute deviations (LAD) estimation which lead to the median regression estimator. LAD estimation method was later extended to more general quantiles [51]. The censored quantile regression is very appealing for analyzing economic data due to its robustness to nonnormality or heteroscedasticity. However, the computation of censored QR estimators and associated variance estimators is challenging because the objective function is not convex and an unknown density function needs to be estimated.

No work has been done for dependent observations until recently Wang and Fygenson [73] proposed an inference procedure for longitudinal studies with application to a HIV/AIDS study. A simple quantile rank score test was developed to test for the treatment effect, while the regression coefficient of treatment effect was not directly estimated. This approach avoided the computation of the complex variance estimator. In the observational study like GenIMS, the point estimates of all the covariate effects are of equal importance. The methods focusing on the estimating procedures are not fully developed yet.

2.3 ANALYSIS OF LONGITUDINAL DATA WITH DROPOUTS

2.3.1 Likelihood Based Approaches

We often encounter missing data due to dropouts in longitudinal studies. Generalized estimating equations (GEEs) can be used for estimating the parameters of marginal models in longitudinal studies and provide consistent estimates when the missingness is independent of both the observed and missing data (MCAR). If missingness may be related to the observed responses but conditionally independent of the missing data (MAR), GEE methods are no longer valid. However, mixed models can handle data that are MAR. When there are missing data that are not ignorable (NMAR), mixed models will result in biased estimates. For general missing patterns, the selection model and mixture model are commonly used for modeling non-ignorable missing longitudinal data. The selection model is based on the joint distribution which is a product of the complete data model. The interest of the selection model is in parameter which is under the hypothesized complete data. If the full data is modeled as a mixture over dropout categories then it is called a pattern mixture model. In the pattern mixture model, the parameter conditional on the missing data pattern is the primary interest. These models are under-identified and well suited for small percentages of missing observations.

If data are missing due to dropout, joint modeling is a common strategy to handle informative dropout data (NMAR). Such models have been considered for longitudinal censored data. Lyles [42] implemented a maximum likelihood procedure to estimate initial HIV RNA levels and slopes within a population, compare these parameters across subgroups of HIV-infected women and illustrate the importance of appropriate treatment of left-censoring and informative dropouts. Thiebaut [70] propose a likelihood inference for a parametric joint model including a bivariate linear mixed model for the two markers and a log normal survival model for the time to dropout. Gao and Thiebaut [66] considered the situation when the longitudinal outcomes are also subject to non-ignorable missing in addition to truncation. A shared random effect parameter model is presented where the missing data mechanism depends on the random effects used to model the longitudinal outcomes.

The weighting techniques have been considered for semi-parametric regression modeling (Robins et al. [53]) and has been applied in analyzing many incomplete longitudinal data problems by Rotnitzky and Robins [55], Lipsitz et al. [36], Demirtas [11], Dufouil et al. [13], Lin et al. [27] and Ibrahim et al. [30]. Weights are computed by inverting the probabilities of being observed. In the longitudinal study, some subjects are more likely to complete the study than others. The pseudo-likelihood approach has been used for estimating parameters in generalized linear mixed models (Wolfinger and O'connell [74]).

2.3.2 Weighting Techniques for Quantile Regression Approaches

Like the GEE method, standard estimating functions of quantile regression models result in consistent estimators when the data are missing completely at random. For missing data due to dropouts and under MAR mechanisms, Lipsitz et al. [36] and Yi and He [77] adopted the inverse probability weighted GEE approach to quantile regression models for uncensored data. Lipsitz et al. [36] mainly focused on the application of quantile regression methods using the independent working assumption of covariance structure to longitudinal responses with dropout and discussed methods for estimating the parameters of marginal models, in which the median or any other quantile of an individual's response at time t is modeled as a function of a time trend and a set of covariates. Yi and He [77] considered median regression models to analyze a longitudinal data set arising from a controlled trial of HIV disease and they incorporated the covariance structure in the estimation procedures and established the asymptotic properties for the resultant estimators. The basic idea of this weighting approach is that an individual's contribution to the traditional estimating equations is weighted by the inverse probability of dropout at the given occasion, i.e., the conventional estimating equations for the quantile regression parameters are weighted inversely proportionally to the probability of dropout. This approach requires the process generating the missing data to be estimable but makes no assumptions about the distribution of the responses other than those imposed by the quantile regression model. This method yields consistent estimates of the quantile regression parameters provided that the model for dropout has been correctly specified. There are many ways to handle informative dropouts in the mixed models or GEE approaches when data are not censored, but limited work was done for quantile regression models [36, 76]. Lyles et al. [42] and Thiebaut et al. [70] studied the mixed models for left-censoring data with informative dropouts. The corresponding method for QR is still lacking.

3.0 MULTIPLE IMPUTATION APPROACHES FOR THE CENSORED COVARIATES

Increasingly used in biomedical studies for the diagnosis and prognosis of acute and chronic diseases, biomarkers provide insight into the effectiveness of treatments and potential pathways that can be used to guide future treatment targets. The measurement of these markers is often limited by the sensitivity of the given assay, resulting in data that are censored either at the lower or upper limit of detection. For the Genetic and Inflammatory Markers of Sepsis (GenIMS) study, many different biomarkers were measured to examine the effect of different pathways on the development of sepsis. In this study, the left-censoring of several important inflammatory markers has led to the need for statistical methods that can incorporate this censoring into any analysis of the biomarker data. This paper focuses on the development of multiple imputation (MI) methods for the inclusion of multiple left-censored biomarkers in a logistic regression analysis. A multivariate normal distribution is assumed to account for the correlations between biomarkers. The Gibbs sampler is used for estimation of the distributional parameters and imputation of the censored markers. The proposed methods are evaluated and compared with some simple imputation methods through simulation. A data set of inflammatory and coagulation markers from the GenIMS study is used for illustration.

3.1 INTRODUCTION

Biomarkers are now a key component of many biomedical studies, providing insight into potential treatment targets and disease pathways. They also provide key information that can be used to inform the diagnosis and prognosis of both acute and chronic diseases. While biomarkers provide useful information, they are also subject to many different limitations due to numerous sources of measurement errors and difficulty in collecting the actual specimens. In addition to the error associated with specimen collecting/processing and laboratory error, left-censoring due to the limits of detection (LOD) is also a common source of error that needs to be addressed at the analysis stage of biomarker data. Left-censoring can be addressed through the use of assays that are more sensitive; however, this is often not feasible due to the cost and time constraints of many studies. Well-known examples of left-censoring occur in the assessment of viral RNA in patients infected with the human immunodeficiency virus (HIV) [29], the antibody response to vaccine in blood serum [46], and the biomarkers interleukin-10 (IL10) and interleukin-6 (IL6) that were collected in the Genetic and Inflammatory Markers of Sepsis (GenIMS) study [35]. In the GenIMS study, the motivating example for this work, a set of inflammatory and coagulation markers were evaluated repeatedly during the course of hospitalization. These markers were measured daily during the first week of hospitalization and less frequently after day 7. The assays used to measure the concentration of the biomarkers in GenIMS were not sensitive enough to detect levels of the molecule at the low end of normal, resulting in moderate to heavy left-censoring of the biomarker data. Since left-censoring generates informative missing data that are not ignorable, the traditional methods of statistical analysis are not optimal and may be invalid. Thus statistical methods that can be applied to left-censored biomarker data are needed.

One common approach that has been applied to the analysis of left-censored data is the use of "fill-in" methods where the left-censored observation is replaced by the LOD, the LOD/2, or the LOD/$\sqrt{2}$ depending on the assumed shape of the left tail of the distribution. This approach is straightforward to implement, but leads to results that are biased towards the null hypothesis. The simple "fill-in" imputation approaches based on certain distributional assumptions of the data eliminate this bias when the censoring is moderate ($< 30\%$), but introduce bias into the associated variance estimates. Multiple imputation (MI) approaches [58] are useful and provide valid statistical inference when the censoring proportion is not high ($< 50\%$) [32]. The "fill-in" methods have some limitations but easy to use in practice and apply to either the outcome or the predictor in an analysis.

Most of methodological development for censored biomarkers has focused on the case where the biomarker is considered as the outcome variable. One commonly applied model for this setting is the Tobit linear regression model [72, 49]. This model has been extended to include a variance component model for the analysis of clustered left-censored outcome data [15]. Many other researchers have developed mixed models for longitudinal left-censored data [29, 31, 42, 75, 70]. For non-normal left-censored outcome data, a method based on the Box-Cox transformation has been proposed by Han and Kronmal [23]. Lyles et al. [41] focused on the estimation of the correlation coefficient when the data are left-censored.

There have been fewer proposals for methods that accommodate censored predictors. Lynn [43] developed likelihood-based methods that can be applied to linear and logistic models with a single censored covariate. Austin and Hoch [5] compared several approaches for censored covariates in a simulation study that was designed to evaluate the bias of estimates for a censored covariate in a linear regression model. Rigobon and Stoker [52] proposed a two-part regression model for both parametric and semi-parametric models, where an index model was used to estimate the censored covariates. The same idea was adopted by D'Angelo and Weissfeld [10] to handle censored covariates in the Cox model. The methods discussed above can be very useful for modeling when only one covariate is censored, but are not easily extended to the multiple covariate setting where the covariates may be correlated. As a result, there are no methods that are available for the inclusion of multiple censored covariates in a regression model.

The focus of this work is on the development and evaluation of potential methods for the setting of multiple censored covariates in a regression model. We propose the use of multiple imputations (MI) due to its simplicity and the fact that it has been shown to be a competitor to maximum likelihood method [43]. A simple MI approach has been used by De Roos et al. [4] to estimate the risk of non-Hodgkin's lymphoma associated with the level of organochlorine chemicals in plasma. However, they assumed a univariate distribution for each censored predictor without accounting for the potential correlation between predictors. The motivation for this work comes from the analysis of biomarker data collected in the GenIMS study, where multiple biomarkers of interest are measured with many of these biomarkers being left-censored. The analysis is similar to that of De Roos et al. [4] with the focus being the prediction of acute kidney injury (AKI) as a function of several, potentially correlated biomarkers. Thus we propose the use of MI methods that account for the potential correlation of the biomarkers by assuming a multivariate normal distribution. The method is based on the use of a Tobit regression model combined with a Gibbs sampler to estimate the distributional parameters and to impute the censored covariate information. In Section 3.2 we present the notation and the proposed methods. In section 3.3 we present the results of the simulation study that was used to compare the proposed methods. Section 3.4 presents the results from the motivating example using the GenIMS data to examine the role of inflammatory and coagulation markers in predicting acute kidney injury.

3.2 NOTATION AND METHODS

Let Z^*_{ij} be the concentration of the j-th biomarker for the i-th subject ($i = 1, \cdots, n; j = 1, \cdots, K$). Assume the biomarker vector $\boldsymbol{Z}^*_i = (Z^*_{i1}, Z^*_{i2}, \cdots, Z^*_{iK})^T$ follows a K-dimensional multivariate normal distribution $\text{MVN}_K(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ with a mean vector, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iK})^T$ and a common covariance matrix, $\boldsymbol{\Sigma}$. Then the j-th biomarker Z^*_{ij} is normally distributed, i.e., $Z^*_{ij} \sim \text{N}(\mu_{ij}, \sigma^2_j)$, where σ^2_j is the j-th diagonal element of $\boldsymbol{\Sigma}$. Suppose the means of biomarkers are related to a set of covariates through a linear regression model, i.e., $\mu_{ij} = \boldsymbol{X}_i\boldsymbol{\beta}_j$, where $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jp})^T$ is an unknown regression parameter vector and $\boldsymbol{X}_i = (x_{i1}, \cdots, x_{ip})$ is covariate vector for the i-th subject. When there is a lower limit of detection for the j-th biomarker, say d_j, Z^*_{ij} is a latent variable and we only observe

$$Z_{ij} = \begin{cases} Z^*_{ij} & \text{if} \quad Z^*_{ij} \geq d_j \\ \textit{censored} & \text{if} \quad Z^*_{ij} < d_j \end{cases}.$$

In the following, we first introduce the Gibbs sampling algorithm for a single censored marker and then extend it to the case of multiple censored markers where the correlations between them are accounted for in the MI procedure.

3.2.1 Multiple Imputation For Single Censored Marker

Following the previous notation and suppressing the subscript j, we denote Z_i as the observed marker concentration for the i-th subject. The corresponding latent variable $Z_i^* \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2)$ is subject to a detection limit d. Let C denote the set of censored observations and C' the uncensored set. Then a Tobit likelihood function for a parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ is

$$L(\boldsymbol{\theta}) = \prod_{i \in C} \Phi\left(d; \boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\right) \prod_{i \in C'} \phi\left(Z_i; \boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\right), \tag{3.1}$$

where Φ and ϕ are the cumulative distribution function and probability density function of a random variable from the $\mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2)$. Under the Bayesian framework, we can derive the posterior distribution of the parameter $\boldsymbol{\theta}$ by using the prior distribution of $\boldsymbol{\theta}$ and the above likelihood function (3.1). Usually informative prior distributions of conjugate form are assumed for $\boldsymbol{\beta}$ and σ^2. Specifically, we assume that the prior conditional distribution for $\boldsymbol{\beta}|\sigma^2$ is $\mathrm{MVN}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{B}_0^{-1})$ and the prior distribution for σ^2 is Inv-$\chi^2(\nu_0, \sigma_0^2)$, an inverse chi-square distribution. The hyperparameters $\boldsymbol{\beta}_0$ and $\boldsymbol{B}_0^{-1}$ are assumed to be a known constant vector and matrix; ν_0 and σ_0^2 are known positive constants.

The Gibbs sampling methodology requires the generation of a Markov chain from the posterior density. However, a Tobit likelihood function multiplied by the prior density is not easy to simplify into tractable posterior densities. In other words, the analytical niceties associated with the conjugate prior no longer hold for the likelihood function of censored data. Chib [63] applied the data augmentation schemes presented by Tanner and Wong [68] to handle the censoring problems within the Gibbs sampling framework. Let $\boldsymbol{Z_C} = \{Z_i < d, i \in C\}$ represent the censored observations and $\boldsymbol{Z_{C'}} = \{Z_i, i \in C'\}$ represent the uncensored observations. Once the parameter space is augmented by the latent data corresponding to the censored observations, the posterior density resulting from the complete data is much easier to simulate. We partition the latent data $\boldsymbol{Z}^*$ into $\boldsymbol{Z_C^*}$ and $\boldsymbol{Z_{C'}^*}$, corresponding to the observed data $\boldsymbol{Z_C}$ and $\boldsymbol{Z_{C'}}$. Then $\boldsymbol{Z_{C'}^*} = \boldsymbol{Z_{C'}}$, and $\boldsymbol{Z_C^*}$ needs to be simulated from the Bayesian predictive distribution. Given the latent data $\boldsymbol{Z}^*$, the conditional posterior distributions of $(\boldsymbol{\beta}, \sigma^2)$ are simply expressed by

$$f[\boldsymbol{\beta}|\boldsymbol{Z}^*, \sigma^2] = \mathrm{MVN}_p(\boldsymbol{\beta}_n, \sigma^2 \boldsymbol{B}_n^{-1}), \tag{3.2}$$

$$f[\sigma^2|\boldsymbol{Z}^*, \boldsymbol{\beta}] = \text{Inv-}\chi^2(\nu_n, \sigma_n^2), \tag{3.3}$$

where $\boldsymbol{\beta}_n = (\boldsymbol{B}_0 + \boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{B}_0\boldsymbol{\beta}_0 + \boldsymbol{X}^T\boldsymbol{Z}^*)$, $\boldsymbol{B}_n^{-1} = (\boldsymbol{B}_0 + \boldsymbol{X}^T\boldsymbol{X})^{-1}$, $\nu_n = \nu_0 + n$ and $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \{\boldsymbol{\beta}_0^T\boldsymbol{B}_0\boldsymbol{\beta}_0 + \boldsymbol{Z}^{*T}\boldsymbol{Z}^* - \boldsymbol{\beta}_n^T\boldsymbol{B}_n\boldsymbol{\beta}_n\}$. Our goal is to impute the censored values by taking independent draws from the distribution $f(\boldsymbol{Z}^*_{\boldsymbol{C}}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{Z})$, where $\boldsymbol{Z} = \{\boldsymbol{Z}_C, \boldsymbol{Z}_{C'}\}$. This is carried out by applying the data augmentation in the Gibbs sampling as follows:

1. Initialize $\boldsymbol{\beta}$ and σ^2 with the maximum-likelihood estimates of a Tobit model: $(\boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}})$
2. **Imputation Step** (update imputed values)
   Given $\beta^{(r)}$ and $\sigma^{2^{(r)}}$ at the r-th iteration,
   sample $Z_i^{*(r+1)}$ from $\mathrm{TruncNormal}_{(-\infty,d]}(\boldsymbol{X}_i\boldsymbol{\beta}^{(r)}, \sigma^{2^{(r)}})$ for the censored observation of subject i, where $\mathrm{TruncNormal}_{(a,b)}(\mu, \sigma^2)$ denotes the normal distribution density $\mathrm{N}(\mu, \sigma^2)$ truncated on the interval (a, b). As shown in details by Gelfand et al. [19], the recovery of the censored observations implies simulation from the corresponding truncated distribution.
3. **Posterior Step** (update parameter estimates)
   Sample $\sigma^{2^{(r+1)}}$ from $f[\sigma^2|\boldsymbol{Z_{C'}}, \boldsymbol{Z_C}^{*(r+1)}, \boldsymbol{\beta}^{(r)}]$,
   Sample $\boldsymbol{\beta}^{(r+1)}$ from $f[\boldsymbol{\beta}|\boldsymbol{Z_{C'}}, \boldsymbol{Z_C}^{*(r+1)}, \sigma^{2^{(r+1)}}]$.
   Given the complete sample data $(\boldsymbol{Z_{C'}}, \boldsymbol{Z}^{*}_{\boldsymbol{C}}{}^{(i+1)})$, simulate the posterior parameter estimates, $\sigma^{2^{(i+1)}}$ and $\boldsymbol{\beta}^{(i+1)}$. These new estimates are then used in the next imputation step.
4. Repeat steps 2 and 3 until the algorithm converges.

The resulting Markov chain from this Gibbs sampler, $(\{\boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}\}, \{\boldsymbol{\beta}^{(1)}, \sigma^{2^{(1)}}, \boldsymbol{Z}^{*}_{\boldsymbol{C}}{}^{(1)}\}, \cdots)$ converges in distribution to $f(\boldsymbol{Z}^*_{\boldsymbol{C}}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{Z})$. After discarding a burn-in of the first L iterations, the next M realizations can be used to form the multiple imputed data sets.

3.2.2 Multiple Imputation For Multiple Censored Markers

The Gibbs sampling algorithm for a single censored marker can be directly extended to the multivariate case. Suppose prior information is incorporated in the prior densities $\boldsymbol{\beta}|\boldsymbol{\Sigma} \sim \mathbb{N}(\boldsymbol{\beta}_0, \boldsymbol{B}_0^{-1}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \sim$ Inv-Wishart$(\nu_0, \nu_0\boldsymbol{\Sigma}_0)$, where $\mathbb{N}$ denotes a matrix normal distribution with a mean matrix $\boldsymbol{\beta}_{0(p\times K)}$, a $p \times p$ row covariance matrix $\boldsymbol{B}_0^{-1}$ and a $K \times K$ column covariance matrix $\boldsymbol{\Sigma}$. Let Inv-Wishart denote an inverse wishart distribution with degrees of freedom $\nu_0 > 0$ and $K \times K$ positive definite matrix $\nu_0\boldsymbol{\Sigma}_0$. Then a normal-wishart informative conjugate prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is

$$f(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathbb{N}(\boldsymbol{\beta}_0, \boldsymbol{B}_0^{-1}, \boldsymbol{\Sigma}) \times \text{Inv-Wishart}(\nu_0, \boldsymbol{\Sigma}_0). \tag{3.4}$$

The Gibbs sampling algorithm can be applied to the three blocks $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{Z}^*_{\boldsymbol{C}}$ with the respective conditional densities $f[\boldsymbol{\beta}|\boldsymbol{Z}^*, \Sigma]$, $f[\Sigma|\boldsymbol{Z}^*, \boldsymbol{\beta}]$ and $f[\boldsymbol{Z}^*_C|\boldsymbol{Z}, \boldsymbol{\beta}, \Sigma]$. The conditional distributions of $(\boldsymbol{\beta}$, $\boldsymbol{\Sigma})$ are expressed as

$$f[\boldsymbol{\beta}|\boldsymbol{Z}^*, \boldsymbol{\Sigma}] = \mathbb{N}(\boldsymbol{\beta}_n, \boldsymbol{B}_n^{-1}, \boldsymbol{\Sigma}), \tag{3.5}$$

$$f[\boldsymbol{\Sigma}|\boldsymbol{Z}^*, \boldsymbol{\beta}] = \text{Inv-Wishart}(\nu_n, \boldsymbol{\Sigma}_n), \tag{3.6}$$

where $\boldsymbol{\beta}_n = (\boldsymbol{B}_0 + \boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{B}_0\boldsymbol{\beta}_0 + \boldsymbol{X}^T\boldsymbol{Z}^*)$, ${\boldsymbol{B}_n}^{-1} = (\boldsymbol{B}_0 + \boldsymbol{X}^T\boldsymbol{X})^{-1}$, $\nu_n = \nu_0 + n$ and $\nu_n\boldsymbol{\Sigma}_n = \nu_0\boldsymbol{\Sigma}_0 + \{\boldsymbol{\beta}_0^T\boldsymbol{B}_0\boldsymbol{\beta}_0 + \boldsymbol{Z}^{*T}\boldsymbol{Z}^* - \boldsymbol{\beta}_n^T\boldsymbol{B}_n\boldsymbol{\beta}_n\}$. For a censored observation, we sample a value from the conditional distribution $f[\boldsymbol{Z}^*_C|\boldsymbol{Z}, \boldsymbol{\beta}, \Sigma]$. This requires sampling from a truncated multivariate normal distribution. As illustrated by Robert [9], we simulate each component successively based on the conditional distribution rather than generating a random vector from the truncated MVN distribution. Specifically, we impute the censored value for marker $j(j = 1, \cdots, K)$ by generating a value from the truncated normal distribution:

$$\text{TruncNormal}_{(-\infty, d_j]}(E[Z_j^*|\boldsymbol{Z}^*_{-j}], \quad \Sigma_{j|-j}),$$

where the conditional mean and variance of Z_j^* given $\boldsymbol{Z}^*_{-j} = (Z_1^*, \cdots, Z_{j-1}^*, Z_{j+1}^*, \cdots, Z_K^*)$ are

$$E[\boldsymbol{Z}_j|\boldsymbol{Z}_{-j}] = \mu_j + \boldsymbol{\Sigma}_{j,-j}^T\boldsymbol{\Sigma}_{-j,-j}^{-1}(\boldsymbol{Z}_{-j} - \boldsymbol{\mu}_{-j}), \tag{3.7}$$

$$\Sigma_{j|-j} = \sigma_j^2 - \boldsymbol{\Sigma}_{j,-j}^T\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{j,-j}, \tag{3.8}$$

with $\mathbf{\Sigma}_{-j,-j}$ being a $(K-1) \times (K-1)$ matrix, derived from $\mathbf{\Sigma}$ by eliminating its j-th row and its j-th column, and $\mathbf{\Sigma}_{j,-j}$ being a $(K-1)$ vector derived from the j-th column of $\mathbf{\Sigma}$ by removing the j-th row.

We now take $K = 2$ as an example to illustrate our imputation method based on Gibbs sampler. Assume for the i-th subject $(i = 1, \cdots, n)$,

$$\boldsymbol{Z}_i^* = \begin{bmatrix} Z_{i1}^* \\ Z_{i2}^* \end{bmatrix} \sim \text{BVN}\left[\boldsymbol{\mu}_i = \begin{pmatrix} \boldsymbol{X}_i\boldsymbol{\beta}_1 \\ \boldsymbol{X}_i\boldsymbol{\beta}_2 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right],$$

then $Z_{ij}^*|Z_{ij'}^* \sim \text{N}(\boldsymbol{X}_i\boldsymbol{\beta}_{j|j'}, \sigma_{j|j'}^2) \quad (j, j' = 1, 2; \quad j \neq j')$, where

$$\boldsymbol{X}_i\boldsymbol{\beta}_{j|j'} = \boldsymbol{X}_i\boldsymbol{\beta}_j + \rho\frac{\sigma_1}{\sigma_2}(Z_{ij'}^* - \boldsymbol{X}_i\boldsymbol{\beta}_{j'}), \quad \sigma_{j|j'}^2 = \sigma_j^2(1-\rho^2).$$

For a censored observation, we sample a value from distribution

$$f[Z_{ij}^*|Z_{ij} < d_j, \boldsymbol{X}_i\boldsymbol{\beta}_{j|j'}, \sigma_{j|j'}^2] = \text{TruncNormal}_{(-\infty,d_j]}(\boldsymbol{X}_i\boldsymbol{\beta}_{j|j'}, \sigma_{j|j'}^2).$$

Now the Gibbs sampling algorithm can be described as follows:

1. Initialize $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \rho)$ with the maximum-likelihood estimates of a Tobit model: $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \sigma_{10}^2, \sigma_{20}^2, \rho_0)$.
2. **Imputation Step**

   Given $\hat{\boldsymbol{\theta}}^r$ at r-th iteration, generate the censored observations for subject i by successively sampling from:

   - $Z_{i1}^{*(r+1)} \sim \text{TruncNormal}_{(-\infty,d_1]}(\boldsymbol{X}_i\boldsymbol{\beta}_{1|2}^{(r)}, \sigma_{1|2}^{2(r)})$
   - $Z_{i2}^{*(r+1)} \sim \text{TruncNormal}_{(-\infty,d_2]}(\boldsymbol{X}_i\boldsymbol{\beta}_{2|1}^{(r)}, \sigma_{2|1}^{2(r)})$
3. **Posterior Step**

   Sample $\mathbf{\Sigma}^{(r+1)}$ from $f[\mathbf{\Sigma}|\boldsymbol{Z_{C'}}, \boldsymbol{Z_C}^{*(r+1)}, \boldsymbol{\mu}^{(r)}]$

   Sample $\boldsymbol{\mu}^{(r+1)}$ from $f[\boldsymbol{\beta}|\boldsymbol{Z_{C'}}, \boldsymbol{Z_C}^{*(r+1)}, \mathbf{\Sigma}^{(r+1)}]$
4. Repeat steps 2 and 3 until the algorithm converges.

Once the imputed data sets are created, the multiple imputation inference described in [40] can be performed. Suppose θ is the parameter of interest. We may fill in censored data M times to generate M complete data sets and then apply the standard regression procedure to each complete data set. Let $\hat{\theta}_m$ and V_m $(m = 1, \cdots, M)$ be the estimate and associated variance for θ. The resulting combined point estimate $\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$ and the corresponding variance $Var(\hat{\theta}) = \frac{1+M}{M}S_M^2 + \frac{1}{M}\sum_{m=1}^{M}V_m$, where S_M^2 is the sample variance of estimates $\hat{\theta}_m$.

3.3 SIMULATION STUDY

We conducted various simulation studies to evaluate the performance of our MI approaches based on the Gibbs sampling method and compare these with several other imputation methods. We generated two marker measurements $\boldsymbol{Z}_1^*$ and $\boldsymbol{Z}_2^*$ from a bivariate normal distribution with means $\mu_1 = 1, \mu_2 = 2$, and variances $\sigma_1^2 = \sigma_2^2 = 1$. The correlation ρ between these two markers was set to 0 or 0.2 to represent small correlations and 0.5 or 0.8 for relatively high correlations. The association between these two markers and the binary outcome of interest was described by a logistic regression model, $\text{logit}(\pi = Pr[Y = 1]) = b_0 + b_1\boldsymbol{Z}_1^* + b_2\boldsymbol{Z}_2^*$, where $b_0 = -0.1$, $b_1 = -0.2$ and $b_2 = 0.3$. To achieve a desirable proportion of left-censored data, the detection limits d_j (j=1,2) were selected to be $F^{-1}(c; \mu_j, \sigma_j)$ $(c = 0.2, 0.4)$, implying that on average $100c$ percent of the simulated data are left-censored. The simulation was conducted for 500 data sets with a sample size of $n = 200$. For the Gibbs sampling, 700 iterations were generated for each data set and after discarding a burn-in of the first 200 realizations of the sequence, we took the imputed values from the 201st, 301st, 401st, 501st and 601st iteration to form the M=5 imputed data sets. The length of the burn-in and monitoring was sufficient to achieve convergence as assessed by trace plots and autocorrelation for each parameter. We estimated $\boldsymbol{\mu}_j$ with a linear regression model including the binary response variable Y as a predictor.

Table 1: Simulation Results of Multiple Imputation for One Marker at a Time

b	$b_0 = -0.1$				$b_1 = -0.2$				$b_2 = 0.3$			
Method	**Omni**[a]	**Naive**[b]	**MI$_B$1**[c]	**MI$_G$1**[d]	**Omni**[a]	**Naive**[b]	**MI$_B$1**[c]	**MI$_G$1**[d]	**Omni**[a]	**Naive**[b]	**MI$_B$1**[c]	**MI$_G$1**[d]
					20% censored, ρ =0							
Bias	-0.009	-0.096	-0.018	-0.020	-0.009	-0.034	-0.005	-0.005	0.006	0.056	0.009	0.010
SE	0.365	0.455	0.376	0.375	0.152	0.181	0.156	0.156	0.152	0.185	0.157	0.157
MSE	0.258	0.410	0.275	0.272	0.047	0.068	0.050	0.050	0.047	0.073	0.051	0.050
CP	0.966	0.964	0.968	0.960	0.962	0.950	0.956	0.950	0.960	0.950	0.956	0.960
					40% censored, ρ =0							
Bias	-0.009	-0.204	-0.017	-0.025	-0.009	-0.070	-0.003	0.001	0.006	0.114	0.008	0.010
SE	0.365	0.615	0.401	0.395	0.152	0.225	0.166	0.164	0.152	0.235	0.169	0.167
MSE	0.258	0.782	0.306	0.303	0.047	0.109	0.056	0.055	0.047	0.126	0.057	0.056
CP	0.966	0.960	0.972	0.964	0.962	0.948	0.952	0.946	0.960	0.928	0.960	0.964
					20% censored, ρ =0.2							
Bias	-0.006	-0.090	-0.012	-0.012	-0.007	-0.028	0.001	0.001	0.005	0.050	0.004	0.004
SE	0.344	0.428	0.355	0.354	0.154	0.184	0.159	0.158	0.155	0.188	0.156	0.159
MSE	0.230	0.366	0.246	0.245	0.049	0.070	0.051	0.051	0.049	0.074	0.052	0.052
CP	0.970	0.958	0.962	0.962	0.952	0.948	0.958	0.956	0.956	0.948	0.954	0.958
					40% censored, ρ =0.2							
Bias	-0.006	-0.196	-0.007	-0.007	-0.007	-0.054	0.016	0.015	0.005	0.102	-0.005	-0.001
SE	0.344	0.579	0.380	0.375	0.154	0.228	0.169	0.167	0.155	0.237	0.171	0.168
MSE	0.230	0.697	0.279	0.277	0.049	0.111	0.057	0.057	0.049	0.125	0.058	0.057
CP	0.970	0.954	0.960	0.958	0.952	0.952	0.958	0.956	0.956	0.938	0.958	0.956

[a] **Omni**: Omniscient, [b] **Naive**: Censored observations replaced by LOD/2,
[c] **MI$_B$1**: MI-Bootstrapping, [d] **MI$_G$1**: MI-Gibbs sampling

For the case when the two markers are independent or weakly correlated, we applied our Gibbs sampling based MI method for one marker at a time and compared the results to the bootstrap based MI approach presented by Lubin et al. [32]. In the bootstrap based procedure, a bootstrap sample was generated first from the observed data with replacement and the Tobit likelihood function was used to obtain the estimates $\tilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma^2}$. Then the censored observation was imputed by a value generated from the inverse cumulative distribution function

$$\Phi^{-1}\{\text{UNIF}[0, \Phi(d; \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)]; \tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\}, \tag{3.9}$$

where UNIF$[0, a]$ is a uniform distribution on $[0, a]$. For each of these data sets we compared an omniscient estimate (**Omni**) obtained from the complete data, an naive estimate (**Naive**) based on replacing censored observation by LOD/2, a MI estimate based on bootstrapping

($\mathbf{MI}_B\mathbf{1}$) and the MI estimate based on Gibbs sampling ($\mathbf{MI}_G\mathbf{1}$). Note that the correlation between markers was essentially ignored in $\mathbf{MI}_B\mathbf{1}$ and $\mathbf{MI}_G\mathbf{1}$ as they were applied for the censored markers one at a time. Table 1 summarizes the results obtained from the simulation study where each of the two markers is treated independently in the analysis. As expected, the two MI approaches ($\mathbf{MI}_B\mathbf{1}$ and $\mathbf{MI}_G\mathbf{1}$) performed much better than the Naive method, even when the data are not heavily censored (i.e., 20%). The naive substitution with LOD/2 yielded significantly biased estimates and larger SEs and MSEs. Compared to the Omni method, which serves as the gold standard, both of the MI approaches consistently produced approximately unbiased estimates. The SEs, MSEs and CPs were also comparable. For the setting where we incorporate the correlation between two markers, we conducted a simulation study similar to that outlined above. As shown in Table 2, the estimates from methods $\mathbf{MI}_B\mathbf{1}$ and $\mathbf{MI}_G\mathbf{1}$ are considerably biased for the effects of censored markers, even though the intercept estimates remained fine when the censoring and correlations are not high (20% censoring, $\rho = 0.5$). In contrast, method $\mathbf{MI}_G\mathbf{2}$, the MI method accounting for marker correlations, still resulted in unbiased estimates for all of the coefficients when censored markers are highly correlated. Even when the correlation is moderate ($\rho = 0.5$), method $\mathbf{MI}_G\mathbf{2}$ appeared to work better for data subject to heavier censoring.

Table 2: Simulation Results of Multiple Imputation for Higher Correlated Markers

b	$b_0 = -0.1$				$b_1 = -0.2$				$b_2 = 0.3$			
Method	**Omni**[a]	**MI$_B$1**[b]	**MI$_G$1**[c]	**MI$_G$2**[d]	**Omni**[a]	**MI$_B$1**[b]	**MI$_G$1**[c]	**MI$_G$2**[d]	**Omni**[a]	**MI$_B$1**[b]	**MI$_G$1**[c]	**MI$_G$2**[d]
					20% censored, ρ =0.5							
Bias	-0.005	0.002	-0.001	-0.001	-0.005	0.015	0.015	0.002	0.003	-0.011	-0.009	0.004
SE	0.330	0.340	0.340	0.340	0.173	0.177	0.177	0.181	0.174	0.178	0.177	0.181
MSE	0.211	0.226	0.226	0.226	0.061	0.062	0.062	0.061	0.060	0.062	0.062	0.062
CP	0.966	0.968	0.962	0.962	0.962	0.970	0.968	0.962	0.952	0.954	0.948	0.954
					40% censored, ρ =0.5							
Bias	-0.005	0.018	0.011	-0.010	-0.005	0.045	0.045	0.017	0.003	-0.033	-0.029	-0.005
SE	0.330	0.363	0.357	0.356	0.173	0.184	0.182	0.183	0.174	0.185	0.182	0.181
MSE	0.211	0.257	0.252	0.252	0.061	0.070	0.070	0.070	0.060	0.068	0.066	0.065
CP	0.966	0.968	0.964	0.960	0.962	0.956	0.952	0.952	0.952	0.950	0.944	0.952
					20% censored, ρ =0.8							
Bias	-0.006	0.039	0.038	0.007	-0.005	0.053	0.053	0.027	0.004	-0.048	-0.047	-0.020
SE	0.361	0.366	0.365	0.364	0.248	0.247	0.247	0.246	0.248	0.246	0.246	0.245
MSE	0.254	0.256	0.257	0.255	0.123	0.113	0.114	0.112	0.122	0.113	0.113	0.111
CP	0.950	0.962	0.960	0.958	0.962	0.968	0.970	0.964	0.958	0.966	0.962	0.960
					40% censored, ρ =0.8							
Bias	-0.006	0.078	0.071	0.046	-0.005	0.102	0.102	0.087	0.004	-0.091	-0.088	-0.071
SE	0.361	0.375	0.372	0.370	0.248	0.240	0.239	0.238	0.248	0.240	0.238	0.237
MSE	0.254	0.272	0.267	0.265	0.123	0.111	0.112	0.110	0.122	0.108	0.107	0.105
CP	0.950	0.958	0.962	0.958	0.962	0.952	0.950	0.955	0.958	0.958	0.956	0.958

[a] **Omni**: Omniscient, [b] **MI$_B$1**: MI-Bootstrapping,
[c] **MI$_G$1**: MI-Gibbs sampling for single censored marker,
[d] **MI$_G$2**: MI-Gibbs sampling accounting for correlations between markers

3.4 APPLICATION

The Genetic and Inflammatory Markers of Sepsis (GenIMS) study was designed to identify genetic markers and biomarkers related to the development of severe sepsis as a result of community acquired pneumonia (CAP). The study enrolled 2,320 subjects with CAP from emergency department at 28 US hospitals. In addition, several secondary outcomes such as organ failure, acute kidney injury (AKI) and death were also examined. To assess the relationship between potential biomarkers and these outcomes, biomarkers in the inflammatory and coagulation pathways were measured daily during the first seven days of hospitalization and weekly thereafter. For the example presented here, we focused on the prediction of AKI using day 1 levels of cytokines and fibrinolysis markers. The analysis cohort included 1836 patients who were confirmed CAP cases admitted to the hospital and with available biomarker data. The markers analyzed for this example include tumor necrosis factor (TNF) which was censored at a lower limit of 4, interleukin-6 (IL6) which was censored at either 2 or 5 depending on the assay used, interleukin-10 (IL10) which was censored at 5, plasminogen activator inhibitor (PAI-1) which was censored at 2 and D-dimer which was censored at the lower limit of 110. The censoring proportions for these markers were 34.97%, 27.34%, 9.42%, 8.28% and 1.74%, respectively. We assumed a log-normal distribution for the biomarker concentrations and analyzed the data using the natural log scale. To apply the multiple imputation procedures, the means of each of the biomarkers were estimated. The biomarker means were modeled in the log scale using linear regression models that included baseline characteristics (age, gender and baseline creatinine as marker of kidney function) and the outcome variable AKI.

To assess the magnitude of correlations among these five markers, we used the method of Lyles et al. [41] to estimate the correlation coefficients between the two censored markers. The estimated correlation matrix is given by

$$\left(\begin{array}{ccccc} \underline{\textbf{IL6}} & \underline{\textbf{IL10}} & \underline{\textbf{TNF}} & \underline{\textbf{PAI-1}} & \underline{\textbf{D-dimer}} \\ 1.00 & \mathbf{0.47} & \mathbf{0.40} & 0.17 & 0.25 \\ & 1.00 & 0.34 & 0.21 & 0.05 \\ & & 1.00 & 0.18 & 0.30 \\ & & & 1.00 & -0.01 \\ & & & & 1.00 \end{array}\right),$$

indicating that the correlations among the markers are small to moderate. The correlations among cytokines (IL6, IL10 and TNF) are relatively stronger than those between the two fibrinolysis markers (PAI-1 and D-dimer), so we only incorporated the cytokine correlations in the Gibbs-sampling based MI method ($\mathbf{MI}_G\mathbf{2}$) and compared the results to those obtained from naive method, and two simple MI methods, $\mathbf{MI}_B\mathbf{1}$ and $\mathbf{MI}_G\mathbf{1}$. Table 3 provides the estimates, standard errors and p-values of each risk factor for the development of AKI. The results from the four methods were similar for the adjusted baseline variables and markers with small amounts of censored data which are PAI-1 and D-dimer. However, there were noticeable differences across the four methods for the coefficient estimates and significance of cytokine effects, when both of the correlations and levels of censoring were higher. In particular, the effect of TNF became significant when the correlations between markers were incorporated in the MI method. These results are consistent with the simulation study. In all cases the MI methods outperformed the naive method when the censoring proportion reaches 20%. With censoring proportions of 30% or higher, it becomes important to account for the correlations in the MI, even if the correlation is moderate (e.g, 0.4 or 0.5).

Table 3: Analysis Results for Prediction of AKI using Day 1 Cytokines and Fibrinolysis Markers

Method	**Naive (LOD/2)**[a]			**MI$_B$1**[b]			**MI$_G$1**[c]			**MI$_G$2**[d]		
Parameter	Est.	SE	p	Est.	SE	p	Est.	SE	p	Est.	SE	p
Intercept	-7.917	1.055	.000	-7.942	1.064	.000	-7.626	1.045	.000	-8.020	1.068	.000
Age	0.041	0.007	.000	0.041	0.007	.000	0.041	0.007	.000	0.040	0.007	.000
Male	-0.625	0.254	.014	-0.617	0.254	.008	-0.630	0.255	.007	-0.634	0.251	.006
Creatinine	2.418	0.819	.003	2.375	0.821	.002	2.473	0.826	.001	2.435	0.810	.001
logIL6	0.070	0.052	.177	0.070	0.053	.094	0.061	0.052	.125	0.012	0.048	.407
logIL10	0.027	0.077	.723	0.032	0.074	.336	0.036	0.075	.319	0.065	0.076	.204
logTNF	**0.157**	**0.096**	**.104**	**0.178**	**0.101**	**.047**	**0.124**	**0.085**	**.077**	**0.282**	**0.099**	**.003**
logPAI-1	0.211	0.075	.005	0.220	0.077	.002	0.219	0.076	.003	0.221	0.087	.006
logD-dimer	0.241	0.090	.008	0.240	0.092	.004	0.202	0.083	.014	0.251	0.094	.004

Censoring proportion: **IL6(9.42%)**, **IL10(34.97%)**, **TNF(27.34%)**, PAI-1(8.28%), D-dimer(1.74%)

[a] **Naive**: Censored observations replaced by LOD/2, [b] **MI$_B$1**: MI-Bootstrapping,

[c] **MI$_G$1**: MI-Gibbs sampling for single censored marker,

[d] **MI$_G$2**: MI-Gibbs sampling accounting for correlations between cytokines (IL6, IL10 and TNF)

3.5 DISCUSSION

Censoring issues due to lower or upper detection limits are not uncommon in biomarker studies, but it is often not well-documented at the data collection and can be easily neglected in the analysis stage. Motivated by the GenIMS study, we proposed MI procedures based on the Gibbs sampling method for multiple censored markers. Markov Chain Monte Carlo

(MCMC) methods such as the Gibbs sampler have been widely used for imputing data with non-monotone missing patterns. We extended these MI methods to left-censored biomarker data by incorporating the informative missing mechanism (which is known to be due to the detection limit) in the MI procedure. Although various modeling approaches were developed for analyzing censored marker data as response variables, the evaluation of diagnostic and prognostic performance of markers usually requires marker measurement to be treated as predictors/covariates. The MI approach provides a practical and flexible solution for further complex analysis. Our MI methods performed well for low to moderate levels of censoring in the data (20% ∼40%) and can easily accommodate right-censored or interval-censored data. Our simulation results showed that ignoring the correlations between censored markers may lead to biased estimates in the logistic regression when the correlation is high or moderate and the data are heavily censored. Our method requires the assumption of a multivariate normal distribution, which may not be satisfied with marker data. Appropriate transformations (e.g., Box-Cox transformation) need to be considered. When the amount of missing data is not large, there is evidence [61] that inference made from the MCMC based imputed data tend to be robust to departures from the normal distribution. Whether this is the case for censored data and how the choices of prior distributions affect the results merit further study.

4.0 MEDIAN REGRESSION FOR LONGITUDINAL LEFT-CENSORED RESPONSES

Biomarkers are often measured repeatedly in biomedical studies to help understand the development of the disease, identify the patients at high-risk and guide the therapeutic strategies for intervention. One common source of measurement error for biomarkers is left-censoring because the assays used may not be sensitive enough to measure the low concentrations below a detection limit. The likelihood-based approaches assuming multivariate normal distribution have been proposed to account for left-censoring problem; however the biomarker data are often highly skewed even after certain transformations. We propose a median regression model that requires minimal assumption on the distribution and leads to easier interpretation of the results in the original scale of the data. We developed the estimating procedures incorporating correlations between serial measurements for left-censored longitudinal data. We conducted simulation studies to evaluate the properties of the proposed estimators and compare median regression model with mixed models under various specifications of distributions and covariance structures. We demonstrated our method with a data set from the Genetic and Inflammatory Markers of Sepsis (GenIMS) study.

4.1 INTRODUCTION

Biomarkers are often measured repeatedly in biomedical studies for gaining insight of treatment effectiveness and establishing the potential disease pathways to guide the future treatment targets. However, the biomarker data are subject to various sources of measurement errors. Left-censoring due to the lower limit of detection (LOD) is a common source of error that may not be noticed in the analysis stage of biomarker data. In the Genetic and Inflammatory Markers of Sepsis (GenIMS) study (Kellum et al. [3]), a set of inflammatory and coagulation markers were evaluated repeatedly during the course of hospitalization for patients with community acquired pneumonia. Unfortunately, the assays used were not sensitive enough to measure low concentrations of some biomarkers, resulting in moderate to heavy left-censoring data. Figure 1 presents the censoring proportion of cytokines, TNF (tumor necrosis factor), IL6 (interleukin-6) and IL10, over the first week of hospitalization. Since left-censoring introduces informative missing data that are not ignorable, the traditional longitudinal analyses based on mixed model and generalized estimating equation (GEE) approach are no longer valid.

The ad-hoc methods using an arbitrary constant such as LOD, LOD/2, or LOD/$\sqrt{2}$ usually lead to biased estimation results. The existing statistical methods for left-censored data mainly focuses on the likelihood-based approach, where the distribution of censored variable is fully specified. The contribution of censored observations to the likelihood function is indicated by the probability of being censored. For example, a normal distribution is assumed in the Tobit linear regression model (Tobin [71]; Persson and Rootzen [49]) for independent data. Mixed models based on the idea of Tobit model have been developed for the left-censored longitudinal data with various computational algorithms presented by different researchers (Hughes [29]; Lyles et al. [42]; Jacqmin-Gadda et al. [31]; Wu [75]; Thiebaut et al. [70]). A Tobit variance-component method was demonstrated in linkage analysis of family left-censored trait data (Epstein et al. [15]). However, the normality assumption of mixed models may not be satisfied as the biomarker data are often highly skewed even after certain transformation. Additionally, misspecification of covariance structure of response variable in the mixed models may result in biased estimates. The GEE methods are usually considered as a robust alternative to mixed models, but incorporation of left-censoring data in GEE methods is not trivial without specifying the distribution of censored variable. In the literature of econometrics, quantile regression has been popularly used due to its robustness to non-normality or heteroscedasticity. The corresponding quantile regression methods for data censored at a fixed constant were also well established (Powell [50, 51]). However the computation of censored quantile regression estimators and associated variance estimators is challenging because the objective function is not convex and an unknown density function need to be estimated. Little work has been done for longitudinal censored observations until recently Wang and Fygenson [73] proposed an inference procedure for longitudinal studies with application to a HIV/AIDS study. A simple quantile rank score test was developed to test for the treatment effect, while the regression coefficient of treatment effect was not directly estimated. This approach avoided the computation of the complex variance estimator.

In observational studies such as the GenIMS study, the point estimates of all the covariate effects are of equal importance. In this paper, we focus on the estimation procedure for median regression with left-censored longitudinal data subject to fixed and known detection limits. Although the estimating procedure outlined in Wang and Fygenson [73] for nuisance parameters can be directly applied to all the regression parameters, it is based on the working assumption of independence. We will incorporate the correlations between repeated measures as done in Jung [33] for uncensored data. In Section 4.2, we present the notation and methods. In Section 4.3, we provide simulation study to evaluate the proposed methods and compare the median regression with mean regression based on mixed models under various specifications of distributions and covariance structures. In Section 4.4, we demonstrate our methods with GenIMS biomarker data.

4.2 NOTATION AND METHODS

4.2.1 Censored Median Regression

Longitudinal data are typically modeled with marginal model, random effect model and transition model. Here we focus on the marginal modeling approach for median regression. Let y^*_{it} be the continuous response on the i-th subject at time t, we consider the linear regression model

$$y^*_{it} = \boldsymbol{x}^T_{it}\boldsymbol{\beta} + e_{it}, \quad i = 1, \cdots, n; \quad t = 1, \cdots, m_i \tag{4.1}$$

where $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression parameters, $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of covariates for the i-th individual at time t, e_{it} is the random error and T denotes the transpose of a vector or matrix. The error vectors $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{i,m_i})^T$ for $i = 1, \cdots, n$ are independent, but the components of $\boldsymbol{e}_i$ are correlated to each other to reflect the serial correlation of repeated measures within an individual subject. Then a median regression model relating the median of response variable, $med(y^*_{it})$, to a set of covariates has the form

$$med(y^*_{it}) = \boldsymbol{x}^T_{it}\boldsymbol{\beta}, \tag{4.2}$$

where the median of error term is assumed to be zero. There is no other distributional assumption made on random errors. When there is a detection limit in the assay, we can not observe y^*_{it} if it has value below the detection limit, say d. In other words, y^*_{it} is a latent variable and we only observe y_{it}, where $y_{it} = y^*_{it}$, if $y^*_{it} > d$. Thus we consider the following censored regression model,

$$y_{it} = max(d, \boldsymbol{x}^T_{it}\boldsymbol{\beta} + e_{it}), \tag{4.3}$$

which is a straightforward extension of Powell's univariate censored regression model [51]. For univariate censored data, Powell first considered minimizing the objective function

$$M_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - max\{d, \boldsymbol{x}^T_i\boldsymbol{\beta}\}| \tag{4.4}$$

based on a least absolute deviations (LAD) criterion. LAD estimation method was later extended to more general 100τ-th quantiles (Powell [50, 51]) based on the objective function

$$Q_n(\boldsymbol{\beta}_\tau) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_{it} - max\{d, \boldsymbol{x}^T_{it}\boldsymbol{\beta}_\tau\}), \quad (0 < \tau < 1) \tag{4.5}$$

where $\rho_\tau(u) = u\{\tau - I(u \leq 0)\}$ and $I(\cdot)$ is an indicator function. For $\tau = 0.5$, $Q_n(\boldsymbol{\beta}_\tau)$ and $M_n(\boldsymbol{\beta})$ lead to the same estimators for censored median regression. Analogue to the idea of GEE approach, we propose the objective function under the working independence assumption for model (4.3) as

$$M_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m_i} |y_{it} - max\{d, \boldsymbol{x}^T_{it}\boldsymbol{\beta}\}|. \tag{4.6}$$

Wang and Fygenson [73] have used the same objective function to estimate the nuisance regression parameters when they made inference for a subset of quantile regression parameters. They derived the asymptotic properties for the resultant estimators and showed that under mild conditions, the estimators are strongly consistent and asymptotically normal even though the objective function treated all observations as if they were independent. They provided a close form for variance estimator under the assumption of exchangeable covariance structure. It follows along their lines of proof, the estimators that minimize $M_n(\boldsymbol{\beta})$ have the same nice asymptotic properties.

Median regression for uncensored data is often performed via linear programming algorithm because the objective function is not smooth. For censored data, the objective function (4.6) is neither smooth nor convex, which implies that multiple local optima may exist. We apply the BRCENS algorithm of Fitzenberger [16] for optimization since this algorithm was shown to perform better than the standard linear programming algorithm. The variance estimation involves the estimation of a unspecified distribution of error term, and depends on the underlying true covariance structure, we circumvent this computational problem by using the bootstrap method and evaluate the performance of bootstrap estimator in the simulation study. To retain the correlation structure of the responses, we take each subject as a sampling unit, and draw a random sample of size n with replacement from the original data. To facilitate the estimation with bootstrap sample $\{\tilde{y}_{it}, \tilde{\boldsymbol{x}}_{it}\}$, we minimize the following modified convex objective function

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m_i}\rho(\tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}^T\boldsymbol{\beta})I(\tilde{\boldsymbol{x}}_{it}^T\hat{\boldsymbol{\beta}} > d)$$

as in Wang and Fygenson [73], where the loss function $\rho(u) = u\{0.5 - I(u \leq 0)\}$ and $\hat{\boldsymbol{\beta}}$ is the minimum of function $M_n(\boldsymbol{\beta})$ in (4.6).

4.2.2 Weighted Censored Median Regression

Under the working independence assumption, intra-subject correlation structure is not incorporated in the estimating function. To improve efficiency, we may construct a weighted estimating equation as done in Jung [33] for uncensored data, with weights calculated based on the correlation structure. Let covariance matrix

$$\boldsymbol{V}_i = \text{cov}(0.5 \times \mathbf{1}_{m_i} - I(\boldsymbol{y}_i \leq \boldsymbol{X}_i\boldsymbol{\beta})), \tag{4.7}$$

where $\boldsymbol{y}_i = (Y_{i1}, \cdots, Y_{im_i})^T$, $\boldsymbol{X}_i$ is an $m_i \times p$ matrix of covariates and $\mathbf{1}_{m_i}$ is an m_i-vector of 1's. The median regression estimator for uncensored longitudinal data, $\hat{\boldsymbol{\beta}}$, can be obtained as a solution to

$$S_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{\Gamma}_i \boldsymbol{V}_i^{-1}[0.5 \times \mathbf{1}_{m_i} - I(\boldsymbol{y}_i \leq \boldsymbol{X}_i\boldsymbol{\beta})] = \mathbf{0}, \tag{4.8}$$

where $\boldsymbol{\Gamma}_i$ is an $m_i \times m_i$ diagonal matrix with t-th diagonal element being the probability density function of e_{it} evaluated at zero. Equation (4.8) is optimal in terms of asymptotic efficiency. When the random errors e_{it} are identically independently distributed, the optimal weighting matrix is simply $\boldsymbol{V}_i^{-1}$ because $\boldsymbol{\Gamma}_i$ is constant across the subjects.

Now, we apply the weighting technique to the left-censored longitudinal data by considering the estimating equation

$$S_n^w(\boldsymbol{\beta}_w) \quad = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{\Lambda}_i \boldsymbol{W}_i^{-1}[0.5 \times \mathbf{1}_{m_i} - I(\boldsymbol{y}_i \leq max\{d, \boldsymbol{X}_i\boldsymbol{\beta}_w\})] = \mathbf{0}, \tag{4.9}$$

where $\boldsymbol{W}_i = \text{cov}(0.5 \times \mathbf{1}_{m_i} - I(\boldsymbol{y}_i \leq max\{d, \boldsymbol{X}_i\boldsymbol{\beta}_w\}))$, and $\boldsymbol{\Lambda}_i$ is a diagonal matrix, denoted by $\text{diag}(\lambda_{i1}(0), \lambda_{i2}(0), \cdots, \lambda_{im_i}(0))$, and λ_{it} is the probability density function of $(y_{it} - max\{d, \boldsymbol{x}_{it}^T\boldsymbol{\beta}_w\})$. The kernel density estimator (Silverman [65]) can be used for estimation of λ_{it}.

In contrast to the unweighted approach under the working assumption of independence, we need to solve the estimating equation here rather than minimizing an objective function. We use the following iterative algorithm to obtain the solution to the equation (4.9).

Step 1: Initialize $\hat{\boldsymbol{\beta}}_w^{(0)} = (\hat{\beta}_{w1}^{(0)}, \cdots, \hat{\beta}_{wp}^{(0)})^T$ with the estimate from the unweighted approach.

Step 2: Given a current estimate $\hat{\boldsymbol{\beta}}_w^{(k)}$, compute $\hat{\boldsymbol{W}}_i^{(k)}$ and $\hat{\boldsymbol{\Lambda}}_i^{(k)}$and then substitute them into equation (4.9).

Step 3: Obtain updated $\hat{\beta}_{wj}^{(k+1)}$ of j-th parameter $(j = 1, \cdots, p)$ by solving equation (4.9) using the bisection method, fixing all other arguments. We recurrently update the parameters.

Repeat Steps 2 and 3 until the algorithm converges. We applied the classical bootstrap method to compute the variance estimator. The arguments in Jung [33] and Wang and Fygenson [73] can be extended to establish the asymptotic properties of the weighted estimators. In this paper, we will examine the performance of proposed estimators through simulation study.

4.3 SIMULATION STUDY

We conduct simulation study to investigate the finite sample performance of two median regression methods (unweighted and weighted) for censored longitudinal data. We examine the relative efficiency of these two methods and compare them with the naive method where censored observations are replaced by the half of the detection limit. When normality assumption is questionable in practice, median regression analysis provides an important alternative to traditional mean regression analysis. We demonstrate the performance of median regression method for data from non-normal distributions and compare the results with those using mixed models. We also assess whether median regression model is more robust to misspecification of covariance structure as compared to the mixed model.

We generate the latent longitudinal data, y_{it}^*, from the model

$$y_{it}^* = \beta_0 + \beta_1 x_i + \beta_2 t + e_{it} - F_e^{-1}(0.5), \tag{4.10}$$

for $i = 1, \cdots, n$. The covariates include a time-invariant binary variable x_i, generated from a Bernoulli(0.3) distribution and a time factor $t = 1, 2, 3$ to index three follow-up times of measurements. The parameters are selected as $(\beta_0, \beta_1, \beta_2) = (-1, 1.5, 0.5)$. Random error vectors, $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n$, are assumed to be mutually independent with a multivariate distribution. $F_e(\cdot)$ is the cumulative distribution function of e_{it} and $F_e^{-1}(0.5)$ corresponds to the median of e_{it}. To achieve a desirable censoring proportion ($c = 0.2$ and 0.4), we choose the detection limits as $(100\times c)$-th sample percentile of the simulated data. The variance estimation is based on 1000 bootstrap samples. In particular, we consider the following configurations.

(1) For evaluation of performance of median regression methods, two hundred simulations are conducted with sample size n=200.

$\boldsymbol{e}_i \sim \text{MVN}(\mathbf{0}, \sigma^2\boldsymbol{R})$, where $\sigma^2 = 1$ and correlation matrix $\boldsymbol{R}$ is exchangeable with correlation coefficient $\rho = 0, 0.3, 0.5, 0.6$ and 0.8.

(2) For comparison between median regression and mean regression under various distributions, five hundred simulations with n=200 were conducted.

- case 1: Multivariate normal distribution, exchangeable correlation matrix

  $\boldsymbol{e}_i \sim \text{MVN}(\mathbf{0}, \sigma^2\boldsymbol{R})$, where $\sigma^2 = 1$ and $\boldsymbol{R}$ is exchangeable with $\rho = 0.8$.

- case 2: Multivariate normal distribution, unstructured correlation matrix

  $\boldsymbol{e}_i \sim \text{MVN}(\mathbf{0}, \sigma^2\boldsymbol{R})$, where $\sigma^2 = 1$ and

$$\boldsymbol{R} = \begin{pmatrix} 1.00 & 0.37 & 0.55 \\ & 1.00 & 0.77 \\ & & 1.00 \end{pmatrix}.$$

- case 3: Asymmetric distribution

  $\boldsymbol{e}_i = exp(\boldsymbol{\xi}_i) - 1$ and $\boldsymbol{\xi}_i \sim \text{MVN}(\mathbf{0}, \sigma^2\boldsymbol{R})$, where $\sigma^2 = 1$ and $\boldsymbol{R}$ is exchangeable with $\rho = 0.8$.

- case 4: Heteroscedastic model where variance depends on covariates

  $\boldsymbol{e}_i = exp(\boldsymbol{\xi}_i) - 1$ and $\boldsymbol{\xi}_i \sim \text{MVN}(\mathbf{0}, 1/(1 + x_i + t)\boldsymbol{R})$, where $\boldsymbol{R}$ is exchangeable with $\rho = 0.8$. The conditional distribution of y_i given x_i and t is asymmetric about its median and its variance varies with x_i and t.

Table 4: Simulation Results of Censored Median Regression

β	$\beta_0 = -1$				$\beta_1 = 1.5$				$\beta_2 = 0.5$			
Method	**Omni**[a]	**CMR**[b]	**wCMR**[c]	**Naive**[d]	**Omni**[a]	**CMR**[b]	**wCMR**[c]	**Naive**[d]	**Omni**[a]	**CMR**[b]	**wCMR**[c]	**Naive**[d]
					ρ **=0.5 ,**	**20% censored**						
Bias	-0.012	-0.021	-0.020	0.268	0.010	-0.013	0.009	-0.077	0.003	0.006	0.005	-0.092
SE	0.132	0.127	0.127	0.217	0.147	0.147	0.147	0.157	0.053	0.054	0.053	0.079
empSE	0.122	0.133	0.133	0.061	0.144	0.145	0.145	0.131	0.047	0.050	0.049	0.040
MSE	0.034	0.037	0.037	0.125	0.034	0.045	0.045	0.050	0.005	0.006	0.006	0.017
CP	0.935	0.915	0.960	0.900	0.935	0.925	0.950	0.935	0.950	0.945	0.950	0.895
					ρ **=0.8 ,**	**20% censored**						
Bias	-0.007	-0.012	0.011	0.270	0.005	0.007	0.006	-0.080	0.002	0.004	-0.003	-0.092
SE	0.124	0.114	0.113	0.190	0.168	0.167	0.166	0.177	0.044	0.044	0.043	0.065
empSE	0.118	0.122	0.122	0.058	0.172	0.173	0.173	0.151	0.038	0.039	0.038	0.041
MSE	0.030	0.030	0.030	0.116	0.060	0.060	0.060	0.063	0.003	0.004	0.003	0.015
CP	0.930	0.890	0.930	0.760	0.900	0.885	0.940	0.920	0.960	0.962	0.945	0.750
					ρ **=0.5 ,**	**40% censored**						
Bias	-0.012	-0.051	-0.050	0.820	0.010	0.024	0.022	-0.255	0.003	0.014	0.013	-0.278
SE	0.132	0.280	0.278	0.068	0.147	0.178	0.177	0.133	0.053	0.092	0.092	0.053
empSE	0.122	0.256	0.253	0.046	0.144	0.168	0.168	0.133	0.047	0.081	0.080	0.036
MSE	0.034	0.156	0.153	0.680	0.034	0.064	0.063	0.102	0.005	0.017	0.016	0.082
CP	0.935	0.935	0.945	0.000	0.935	0.920	0.950	0.545	0.950	0.960	0.950	0.005
					ρ **=0.8 ,**	**40% censored**						
Bias	-0.007	-0.021	-0.020	0.817	0.005	0.012	0.011	-0.260	0.002	0.006	0.005	-0.276
SE	0.124	0.244	0.238	0.070	0.168	0.189	0.188	0.147	0.044	0.077	0.073	0.055
empSE	0.118	0.216	0.205	0.049	0.172	0.178	0.178	0.152	0.038	0.065	0.064	0.035
MSE	0.030	0.108	0.107	0.677	0.060	0.076	0.075	0.114	0.003	0.011	0.010	0.081
CP	0.930	0.935	0.940	0.000	0.900	0.900	0.940	0.590	0.960	0.965	0.950	0.005

[a] **Omni**: Omniscient, [b] **CMR**: Censored Median Regression, [c] **wCMR**: Weighted Censored Median Regression, [d] **Naive**: Censored observations replaced by LOD/2

In Table 4, we display the results of median regression under various censoring proportions and correlation coefficients. We compare four estimators, Omniscient estimator based on the complete latent data without censoring, censored median regression (CMR) estimator, weighted CMR (wCMR) estimator accounting for serial correlation, and naive estimator where censored data are replaced by half of the detection limit. All the estimators except for wCMR estimators were obtained under working assumption of independence. The weighting matrix $\boldsymbol{W}_i^{-1}$ was specified in the weighted estimating equations. We report the biases along with the standard errors (SE) estimated from bootstrap method, empirical SE (empSE) calculated as sample standard deviation of estimates, mean squared errors (MSE), and empirical 95% coverage probability (CP). As expected, naive approach leads to poor estimators with large bias and its performance becomes worse as the censoring proportion is increased. In contrast, two censored median regression estimators perform well even when 40% of data are censored, as compared to the Omniscient estimators which serve as a gold standard. Bootstrap estimators of standard errors are in good agreement with the empirical standard errors for all the cases. The empirical 95% confidence intervals also have reasonable coverage rates. Comparing to CMR estimators, wCMR estimators are in general less biased and more efficient as indicated by the smaller standard errors. The wCMR estimators are also associated with better empirical coverage probability rates. The improvement resulted from the weighted approach is more significant when the correlations between measurements are higher and the censoring proportion is bigger (say, rho=0.8, c=0.4). These results are consistent with those observed for the uncensored data. As demonstrated previously (He et al., [24], Yi and He, [77]), the relative efficiency gain from the weighted methods may be mild for finite sample size when the serial correlations is not high enough. For censored data, the proportion of censoring also appears to be an important factor affecting the efficiency gain of the weighted approach.

Table 5: Simulation Results Comparing Censored Median Regression with Tobit Mixed Model

β	$\beta_0 = -1$				$\beta_1 = 1.5$				$\beta_2 = 0.5$			
Method	**Omni**[a]	**CMR**[b]	**wCMR**[c]	**TM**[d]	**Omni**[a]	**CMR**[b]	**wCMR**[c]	**TM**[d]	**Omni**[a]	**CMR**[b]	**wCMR**[c]	**TM**[d]
Case 1: MVN, Exchangeable, 40% censored												
Bias	0.009	-0.036	-0.020	0.018	0.002	0.021	0.016	-0.003	-0.003	0.009	0.002	-0.005
SE	0.120	0.321	0.310	0.101	0.116	0.171	0.165	0.058	0.044	0.100	0.097	0.030
empSE	0.114	0.300	0.293	0.097	0.109	0.158	0.152	0.056	0.042	0.095	0.092	0.029
MSE	0.028	0.204	0.194	0.020	0.026	0.056	0.053	0.006	0.004	0.020	0.017	0.002
CP	0.944	0.914	0.920	0.950	0.946	0.952	0.948	0.962	0.954	0.942	0.952	0.950
Case 2: MVN, Unstructured, 40% censored												
Bias	0.008	-0.037	-0.031	0.119	0.002	0.025	0.024	-0.022	-0.003	0.009	0.003	-0.045
SE	0.122	0.322	0.303	0.110	0.116	0.172	0.166	0.074	0.053	0.104	0.100	0.042
empSE	0.117	0.319	0.316	0.111	0.110	0.169	0.167	0.076	0.050	0.103	0.103	0.041
MSE	0.029	0.214	0.205	0.039	0.026	0.060	0.058	0.012	0.005	0.022	0.020	0.005
CP	0.934	0.888	0.900	0.780	0.950	0.936	0.942	0.938	0.954	0.942	0.945	0.828
Case 3: Asymmetric distribution, 40% censored												
Bias	0.012	-0.010	-0.009	-0.403	0.002	0.006	0.006	0.690	-0.002	0.002	0.001	0.131
SE	0.122	0.373	0.360	0.270	0.118	0.190	0.187	0.176	0.044	0.112	0.109	0.089
empSE	0.113	0.356	0.335	0.270	0.111	0.179	0.171	0.234	0.042	0.110	0.106	0.083
MSE	0.029	0.272	0.251	0.310	0.027	0.070	0.066	0.563	0.004	0.025	0.023	0.032
CP	0.952	0.954	0.940	0.740	0.950	0.956	0.954	0.012	0.950	0.945	0.945	0.752
Case 4: Heteroscedastic model, 40% censored												
Bias	0.008	0.012	0.011	0.457	0.000	-0.001	-0.001	-0.017	-0.002	-0.004	-0.004	-0.170
SE	0.080	0.161	0.160	0.085	0.048	0.060	0.059	0.057	0.024	0.050	0.049	0.029
empSE	0.074	0.160	0.160	0.105	0.045	0.060	0.058	0.062	0.022	0.050	0.049	0.031
MSE	0.012	0.053	0.052	0.228	0.004	0.007	0.006	0.007	0.001	0.005	0.005	0.031
CP	0.946	0.940	0.943	0.020	0.948	0.952	0.954	0.914	0.956	0.940	0.944	0.004

[a] **Omni**: Omniscient, [b] **CMR**: Censored Median Regression,
[c] **wCMR**: Weighted Censored Median Regression, [d] **TM**: Tobit Mixed Model

Table 5 shows the results of median regression models (CMR and wCMR) comparing to those from the Tobit-Mixed (TM) model which is a direct extension of Tobit regression to longitudinal data. We fit the mixed models using SAS procedure PROC NLMIXED (SAS Institute Inc. 2000) as illustrated in Thiebaut and Jacqmin-Gadda [69]. We specify the correlation structure in the Tobit-Mixed model as exchangeable by including only the random intercept. When data are generated from multivariate normal distribution with exchangeable correlation structure (case 1), mixed model results in better estimators with smaller bias in general and much smaller variance. It is not surprising since the mixed models are specified correctly in this case and should lead to more efficient estimators than median regression models. On the other hand, when the underlying correlation matrix is unstructured (case 2), CMR and wCMR estimators correspond to smaller bias and better coverage rates comparing to TM estimators. For non-normally distributed data or even heteroscedastic data (cases 3 and 4), median regression models still perform reasonably well as expected, while mixed models fail to give comparable results.

4.4 APPLICATION

In this section, we illustrate the proposed methods with the cytokine data of GenIMS study. GenIMS is a multi-center cohort study of 2320 subjects with Community-Acquired Pneumonia (CAP) presenting to the emergency departments of 28 US academic and community hospitals between 2001 and 2003. One of the primary goals is to investigate the inflammation pathways of severe sepsis defined as CAP complicated by new-onset organ dysfunction. Cytokines including tumor necrosis factor, IL6 (interleukin-6) and IL10 were measured for patients admitted to the hospital daily during the first week and weekly thereafter. As mentioned previously, unexpected left-censoring data (Figure 1) were seen due to low sensitivity of assays. Now we take IL6 as an example to demonstrate our median regression methods. IL6 concentrations were measured using an Immulite assay (Diagnostic Products, Los Angeles, CA). The minimum detectable limit for IL6 was 5 pg/ml per the manufacturer's specifications and the overall proportions of left-censoring was 27.34%. We assume a log-normal distribution for IL6 and analyzed data in a natural log scale. Excluding those whose diagnosis of CAP were ruled out after hospital admission, we include 1886 inpatients in the analysis. Among these patients, 583 (31%) subjects developed severe sepsis.

Table 6: Longitudinal Analysis of GenIMS IL6 vs Severe Sepsis

Method	**Tobit Mixed**[a]			**CMR**[b]			**Weighted CMR**[c]		
Parameter	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Intercept	3.766	0.182	0.000	3.060	0.156	0.000	3.084	0.160	0.000
Sepsis	0.423	0.091	0.000	0.341	0.134	0.011	0.322	0.143	0.024
Day	-0.538	0.016	0.000	-0.329	0.017	0.000	-0.345	0.019	0.000
Sepsis∗Day	0.162	0.024	0.000	0.108	0.027	0.000	0.123	0.029	0.000
Age	0.001	0.003	0.843	0.007	0.002	0.002	0.008	0.002	0.002
White	0.058	0.099	0.557	-0.032	0.080	0.694	0.030	0.081	0.715
Male	0.257	0.080	0.001	0.260	0.069	0.000	0.267	0.068	0.000
Charlson>0	-0.296	0.091	0.001	-0.298	0.074	0.000	-0.335	0.074	0.000

[a] **Tobit Mixed**: Tobit Mixed Model, [b] **CMR**: Censored Median Regression
[c] **Weighted CMR**: Weighted Censored Median Regression

We examine the relationship between severe sepsis and IL6 trajectory during the first week of hospitalization using Tobit-Mixed model, censored median regression (CMR) model and weighted CMR model. The response variables is log transformed IL6 and the covariates adjusted in the models are age, gender, race (whites vs. non-whites) and charlson comorbidity index (>0 vs. 0). Table 6 summarizes the estimate (Est.), standard error (SE), and p-value for each variable of interest. The p-values were calculated based on the Wald test using the bootstrap estimator of standard error. The results from all the models are comparable except for the effect of age. Both mean and median regression analysis suggest that IL6 is higher in patients who developed severe sepsis compared to those who did not. The effect of age is not significant in mean regression model, but highly significant in median regression model. We observe very similar results from weighted and unweighted median regression analysis. It is possible that the overall intra-subject correlation is not strong enough to see the difference between these two methods. Figure 2 displays the estimated trajectories of IL6 by severe sepsis group from the three methods (TM, CMR and wCMR) for a white male at age 72 (median) with Charlson index > 0. There is a decreasing trend in IL6 concentration over time and IL6 decreases faster in patients without severe sepsis. The difference in mean and median trajectories indicates that the normality assumption may

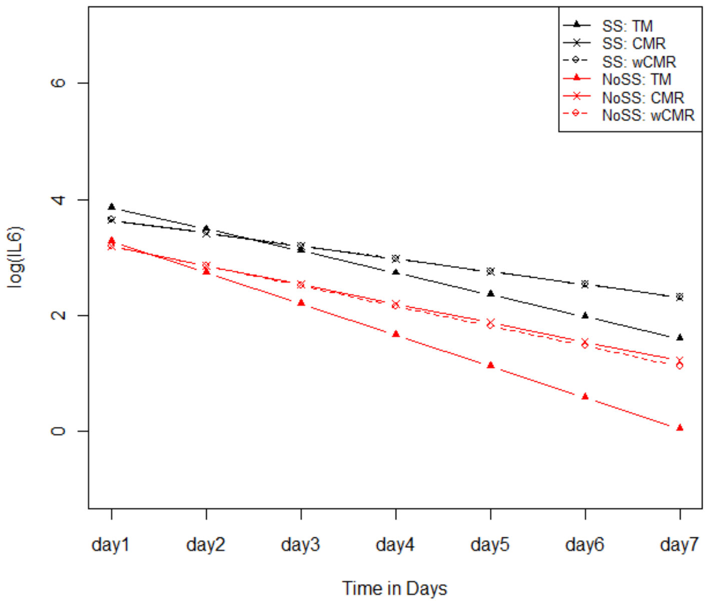

Figure 2: Mean and Median Regression Results for GenIMS Cytokine IL6

not be appropriate. Comparing to the median regression model, the mean regression model demonstrates a bigger difference between patients who developed severe sepsis and those who did not.

4.5 DISCUSSION

Since biomarker data are often highly skewed, median regression has been increasingly used for analyzing longitudinal data in biomedical studies. As a more flexible and robust method, median regression not only provides a valid approach for data that are not normally distributed, but also may provide additional insights on biological mechanisms that are not revealed by mean regression models. We considered a censored median regression model to accommodate the data censored at a fixed detection limit. We proposed the weighted estimating equations to incorporate the serial correlations between the repeated measurements. Simulation study showed that our estimators performed well under various distributional assumptions. The improvement upon the unweighted estimator is much noticeable when the censoring is heavy or the marker measurements are highly correlated. Our estimating procedure can be directly extended to the general quantile regression models. Other resampling methods such as bootstrapping estimating equations (Parzen [47]; Wei and Ying [47]) and Markov chain marginal bootstrap method (He and Hu [24]) are more efficient for variance estimation in the quantile regression analysis. Application of these methods to the censored data and weighted estimating equations merits further study.

5.0 QUANTILE REGRESSION FOR LONGITUDINAL BIOMARKER DATA SUBJECT TO LEFT CENSORING AND DROPOUTS

Quantile regression is increasingly used in longitudinal analysis of biomarker data due to its robustness to non-normality and heteroscedasticity. However, in some biomedical studies, the biomarker data can be censored by detection limits of the bioassay used or missing when the subjects drop out from the study. Inappropriate handling of these two issues leads to biased estimation results. We consider the censored quantile regression approach to account for the censoring data and apply the inverse weighting technique to adjust for dropouts. In particular, we develop a weighted estimating equation for censored quantile regression, where an individual's contribution is weighted by the inverse probability of dropout at the given occasion. We conduct simulation studies to evaluate the properties of the proposed estimators and demonstrate our method with a real data set from Genetic and Inflammatory Marker of Sepsis (GenIMS) study.

5.1 INTRODUCTION

With the advance of biotechnology, more and more biomedical studies are attempting to find out the informative biomarkers to better understand the natural history and development of a complex disease, identify the patients at high-risk and guide the therapeutic strategies for intervention. The biomarker data are often measured over a period of time to determine if the temporal changes differ between the patients who develop disease and those who do not. The Genetic and Inflammatory Markers of Sepsis (GenIMS) study (Kellum et al. [70]) is such a cohort study of 2320 patients with community-acquired pneumonia (CAP). CAP is the leading cause of sepsis. Multiple biomarkers on different pathways were measured daily in the first week and weekly thereafter for the CAP patients admitted to the hospitals. The investigators hope to improve the understanding of the biological mechanisms of sepsis, and identify the biomarkers indicating the risk for subsequent outcomes such as severe sepsis, multiple organ failure, and death. Unfortunately, the assays used were not sensitive enough. There are moderate to heavy censoring in pro-inflammatory markers interleukin (IL6) and tumor necrosis factor (TNF), and anti-inflammatory marker IL10. The censoring percentage can be as high as 50%-70% on later days of hospitalization. Furthermore, biomarker measurements are missing at certain days due to administrative errors, death or discharge early. The dominant missing pattern is monotone missing because clinically too ill and too well individuals were dead or discharged prior to one week. Current methods to deal with both censoring and missing data due to dropout are mainly likelihood based approaches. The common strategy is to apply a joint analysis of longitudinal data and dropout process. Mixed models have been extended to accommodate the censoring data due to detection limit (Hughes [29]; Lyles et al. [42]; Jacqmin-Gadda et al. [31]; Wu [75]). To account for the informative dropout, Lyles et al. [42] assumed a joint multivariate normal distribution for the random effects of the mixed models and time (in natural log scale) to dropout. The impact of the dropout process on the biomarker trajectory was explained by the association between time to dropout and the individual random effects. Thiebaut et al. [70] considered a similar joint model including a bivariate linear mixed model for the two markers and a log normal survival model for time to dropout. Gao and Thiebaut [66] took the joint modeling approach for longitudinal and binary outcomes. Under the framework of the shared random effect model, they used mixed model for the longitudinal outcomes and a binary survival model for the incidence of dropping out.

Generalized estimating equation (GEE) approach is another popular method for longitudinal analysis of continuous or categorical data. Comparing to mixed models which assume normality for the outcome variable, GEE method only requires correct specification of the mean structure and is robust to misspecification of the covariance structure. However, development of appropriate methods using GEE approach for censored data is challenging because GEE approach is not likelihood based method. In the field of econometrics, quantile regression has been used due to its robustness to non-normality or heteroscedasticity. The recent improvements in computational methods for quantile regression make it an appealing approach for biomedical studies. Quantile Regression imposes minimal assumption on the quantiles of the response variable, and allows one to relate various quantile levels (e.g., median, 25th, 75th percentiles) to the covariates differently. As an important alternative to the mean regression models, quantile regression models may provide a global assessment of covariate effects. Quantile regression methods for data censored at a fixed constant were well established for independent data (Powell [50, 51]) and extended to the longitudinal data (Wang and Fygenson [73]). If the biomarker measurements are missing due to dropout, standard estimating functions of quantile regression models leads to biased estimates when the missing mechanism is related to the observed responses, namely missing at random (MAR). Under the assumption of monotone missing and MAR, Lipsitz et al. [36] and Yi and He [77] adopted the inverse probability weighted GEE approach for quantile regression models. The basic idea of this approach is that an individual's contribution to the estimating equations is weighted by the inverse probability of dropout at the given occasion. In this study, we will apply such a weighting technique for censored quantile regression model to address both censoring and dropout issues in the biomarker analysis of GenIMS study. We introduce the notation and methods in section 5.2. The simulation results are presented in section 5.3, followed by a numerical example given in section 5.4.

5.2 NOTATION AND METHODS

5.2.1 Censored Quantile Regression

Let y_{it}^* be the biomarker measurement on the i-th subject at time t. We can consider the following linear regression model

$$y_{it}^* = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + e_{it}, \quad i = 1, \cdots, n; \quad t = 1, \cdots, m_i, \tag{5.1}$$

where $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression parameters, e_{it} is the random error and T denotes the transpose of a vector or matrix. The random errors are correlated within the subject to reflect the serial correlations of repeated measurements within each individual. If the τ-th quantile of error term is assumed to be zero, a quantile regression model relating the τ-th quantile of response variable, $q_\tau(y^*_{it})$, to a set of covariates has the form

$$q_\tau(y^*_{it}) = \boldsymbol{x}^T_{it}\boldsymbol{\beta}_\tau, \quad 0 < \tau < 1, \tag{5.2}$$

where $\boldsymbol{\beta}_\tau$ is a vector of quantile specific regression parameters. When there exists a lower detection limit, say c, for biomarker measurements, we can not observe y^*_{it} if it has a value below c. In other words, y^*_{it} is a latent variable and we only observe $y_{it} = y^*_{it}$, if $y^*_{it} > c$. The quantile regression model for censored longitudinal data can be defined as

$$y_{it} = max(c, \boldsymbol{x}^T_{it}\boldsymbol{\beta}_\tau + e_{it}), \tag{5.3}$$

which is a straightforward extension of Powell's [50] censored regression model (CQR) for the univariate case. To obtain the parameter estimator of CQR, Powell [51] proposed to minimize an objective function

$$Q_n(\boldsymbol{\beta}, \tau) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - max\{c, \boldsymbol{X}_i\boldsymbol{\beta}_\tau\}), \tag{5.4}$$

where the loss function $\rho_\tau(u) = u\{\tau - I(u \leq 0)\}$ and $I(\cdot)$ is an indicator function. The function ρ_τ reflects the contribution of residuals; The absolute values of residuals are weighted by τ if the original residual is positive, and weighted by $1-\tau$ if it is negative. When $\tau = 0.5$, $Q_n(\boldsymbol{\beta}, \tau)$ is equivalent to the objective function for median regression based on the least absolute deviations criterion. For longitudinal censored data, we can mimic the idea of GEE approach to define the objective function as

$$Q_n(\boldsymbol{\beta}, \tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m} \rho_\tau(y_{it} - max\{c, \boldsymbol{x}^T_{it}\boldsymbol{\beta}_\tau\}). \tag{5.5}$$

under the working independence assumption. The resulting estimates are equivalent to the solution of estimating equation

$$S_n(\boldsymbol{\beta}, \tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m} \boldsymbol{x}_{it}[\tau - I(y_{it} \leq max\{c, \boldsymbol{x}^T_{it}\boldsymbol{\beta}_\tau\})] = 0 \tag{5.6}$$

Wang and Fygenson [73] have used the same objective function to estimate the nuisance regression parameters when they made inference for a subset of quantile regression parameters. They derived the asymptotic properties for the resultant estimators and showed that under mild conditions, the estimators are strongly consistent and asymptotically normal even though the objective function treated all observations as if they were independent. For standard quantile regression, the objective function is not smooth, linear programming algorithm or iterative bisection methods have been used for parameter estimation. For censored quantile regression, the objective function is neither smooth nor convex, which implies that multiple local optima may exist. The BRECNS algorithm of Fitzenberger [16] has been shown to perform better than the standard linear programming algorithm.

When there are dropouts and the missingness depends on the previous responses(i.e. MAR), the estimators based on equation (5.5) or (5.6) is no longer consistent because the estimating equation is not consistently unbiased as found out by Lipsitz et al. [36] for the uncensored data. The weighting technique of Robins et al. [53] has been used widely for semiparametric regression modeling of incomplete longitudinal data. An individual's contribution to the traditional estimating equations is weighted by the inverse probability of being observed. This approach was taken by Lipsitz et al. [36] for quantile regression for longitudinal data with dropouts under the MAR mechanism. However, since the weighted estimating equations for quantile regression models are not continuous, the asymptotic results presented by Robins et al. [53] for mean regression models are not directly applicable. Yi and He [77] recently established the asymptotic properties of the median regression estimators. In the next section, we will show how to apply this weighting technique to the censored quantile regression model.

5.2.2 Censored Quantile Regression accounting for dropouts

Let D_i be a random variable indicating when the i-th subject was dropped out from the study. Suppose the measurements for the first time point are observed for all the individuals, D_i can take values between 2 and $m + 1$, with $m + 1$ corresponding to a complete measurement sequence. Then the dropout probability at the d_i occasion for the i-th subject is $\pi_{id_i} =$

$Pr\{D_i = d_i\}$ $(d_i = 2, \cdots, m+1)$. Now we consider the weighted estimating equations for censored quantile regression model as

$$\begin{aligned} S_n^d(\boldsymbol{\beta}, \tau) \quad &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\pi_{id_i}} \sum_{t=1}^{d_i} \boldsymbol{x}_{it}[\tau - I(y_{it} \leq max\{c, \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau\})] \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{d_i=2}^{m+1} \frac{I(D_i = d_i)}{\pi_{id_i}} \sum_{t=1}^{d_i - 1} \boldsymbol{x}_{it}[\tau - I(y_{it} \leq max\{c, \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau\})] = 0. \qquad (5.7) \end{aligned}$$

The basic idea of weighted estimating equations is to weight each individual's contribution by the inverse probability of dropout at the given occasion. Let $\boldsymbol{x}_{it}^\star = \frac{1}{\pi_{id_i}} \boldsymbol{x}_{it}$ and $y_{it}^\star = \frac{1}{\pi_{id_i}} y_{it}$, equation (5.7) can be written in the same form as the unweighted estimating equation (5.6) as follows.

$$\begin{aligned} S_n^d(\boldsymbol{\beta}, \tau) \quad &= \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{d_i} \frac{1}{\pi_{id_i}} \boldsymbol{x}_{it}[\tau - I(\pi_{id_i}^{-1} y_{it} \leq max\{c, \pi_{id_i}^{-1} \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau\})] \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{d_i} \boldsymbol{x}_{it}^\star[\tau - I(y_{it}^\star \leq max\{c, \boldsymbol{x}_{it}^{\star T} \boldsymbol{\beta}_\tau\})] = 0, \qquad (5.8) \end{aligned}$$

Thus, the corresponding objective function is in the form of

$$Q_n^d(\beta, \tau) = \sum_{i=1}^{n} \sum_{t=1}^{d_i} \rho_\tau(y_{it}^\star - max\{c, \boldsymbol{x}_{it}^{\star\,T} \boldsymbol{\beta}\}). \qquad (5.9)$$

Now the BRECNS algorithm of Fitzenberger [16] can be straightly applied to minimize this objective function. If π_{id_i} is correctly specified, i.e., the dropout process is correctly modeled, the weighted estimating equations in (5.7) are unbiased for 0 at the true value of $\boldsymbol{\beta}_\tau$ even if the dropout depends on the previous responses. Because following the derivation of equation (8) in Lipsitz et el (1997), we can easily show that

$$\begin{aligned} &E\left[\frac{I(D_i = d_i)}{\pi_{id_i}} \sum_{t=1}^{d_i - 1} \boldsymbol{x}_{it}[\tau - I(y_{it} \leq max\{c, \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau\})]\right] \\ &= E_{\boldsymbol{X}_i}\left(E_{y_i|\boldsymbol{X}_i}\left[\sum_{t=1}^{d_i - 1} \boldsymbol{x}_{it}[\tau - I(y_{it} \leq max\{c, \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau\})]\ E_{D_i|y_i, X_i}\left(\frac{I(D_i = d_i)}{\pi_{id_i}}\right)\right]\right) \\ &= E_{\boldsymbol{X}_i}\left(E_{y_i|\boldsymbol{X}_i}\left[\sum_{t=1}^{d_i - 1} \boldsymbol{x}_{it}[\tau - I(y_{it} \leq max\{c, \boldsymbol{x}_{it}^T \boldsymbol{\beta}_\tau\})]\right]\right) \\ &= E_{\boldsymbol{X}_i}(0) = 0. \qquad (5.10) \end{aligned}$$

Since the variance estimation for quantile regression estimators involves the estimation of a unspecified distribution of error term, and depends on the underlying true covariance structure, we circumvent this computational problem by using the bootstrap method and evaluate the performance of bootstrap estimator in the simulation study. To retain the correlation structure of the responses, we take each subject as the sampling unit, and draw a random sample of size n with replacement from the original data. To facilitate the estimation with bootstrap sample $\{\tilde{y}_{it}^{\star}, \tilde{\boldsymbol{x}}_{it}^{\star}\}$, we minimize the following modified convex objective function

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m_i}\rho(\tilde{y}_{it}^{\star} - \tilde{\boldsymbol{x}}_{it}^{\star T}\boldsymbol{\beta})I(\tilde{\boldsymbol{x}}_{it}^{\star T}\hat{\boldsymbol{\beta}} > c), \tag{5.11}$$

as in Wang and Fygenson [73], where the loss function $\rho(u) = u\{\tau - I(u \leq 0)\}$ and $\hat{\boldsymbol{\beta}}$ is the estimator obtained from equation (5.8). Heuristically, it follows from the arguments in Wang and Fygenson [73] and Yi and He [77] that the estimator of β is consistent and asymptotically normal if the dropout probability π_{id_i} is either known or can be consistently estimated. If the missing data due to dropout arise from the MAR mechanism, estimation of dropout probability is straightforward. Let R_{it} represent the missing status of response variable $y_{it}(i = 1, \cdots, n; t = 1, \cdots, m)$ and $R_{it} = 1$ if y_{it} is observed and 0 otherwise. Then $R_{ij} = 0$ implies that $R_{ij'} = 0$ for all $j' > j$. As described in Liptisz et al. [36] and Yi and He [77] for standard quantile regression with dropout data, we can write dropout probability π_{id_i} at occasion d_i as

$$\begin{aligned}
\pi_{id_i} &= pr(D_i = d_i) = pr(D_i = d_i|\boldsymbol{y}_i^o, \boldsymbol{X}_i) & (5.12)\\
&= pr(R_{i2}, \cdots, R_{i,d_i-1} = 1, R_{i,d_i} = 0|y_{i1}, \cdots, y_{i,d_i-1}, \boldsymbol{X}_i) & (5.13)
\end{aligned}$$

where $\boldsymbol{y}_i^o$ is the observed response history prior to dropout, and $\boldsymbol{X}_i = \{X_{i1}, \cdots, X_{im}\}$ is a set of covariates observed in the complete study period. If we define $\eta_{it} = pr(R_{it} = 1|R_{i1} = \cdots = R_{i,t-1} = 1, y_{i1}, \cdots, y_{i,t-1}, \boldsymbol{X}_i)$, we can write the dropout probability

$$\begin{aligned}
\pi_{id_i} &= \{\prod_{t=2}^{d_i-1} pr(R_{it} = 1|R_{i1} = \cdots = R_{i,t-1} = 1, y_{i1}, \cdots, y_{i,t-1}, \boldsymbol{X}_i, \boldsymbol{\alpha})\} \times \\
&\quad \{1 - pr(R_{id_i} = 1|R_{i1} = \cdots = R_{i,d_i-1} = 1, y_{i1}, \cdots, y_{i,d_i-1}, \boldsymbol{X}_i, \boldsymbol{\alpha})\}^{I\{d_i \leq m\}} \\
&= \left(\prod_{t=2}^{d_i-1} \eta_{it}\right)(1 - \eta_{id_i})^{I\{d_i \leq m\}}, & (5.14)
\end{aligned}$$

where $I\{\cdot\}$ is an indicator function. Now appropriate regression models such as logistic regression model can be used to model η_{it}, and then we can obtain the estimate of π_{id_i} based on the equation above. We will illustrate the estimation procedure in details in section 5.4 using GenIMS data set as an example.

5.3 SIMULATION STUDY

We simulated longitudinal response variable from the model

$$y_{it}^{*} = \beta_0 + \beta_1 x_i + \beta_2 t + e_{it} - F_e^{-1}(\tau), \quad i = 1, \cdots, n; \quad t = 1, \cdots, m, \tag{5.15}$$

where covariates include the indicator variable x_i, simulated from Bernoulli(0.5) and t is the follow-up time. Error term $\boldsymbol{e}_i \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \boldsymbol{R})$, where $\sigma^2 = 1$ and $\boldsymbol{R}$ is an $m \times m$ correlation matrix with exchangeable structure (correlation coefficient $\rho = 0.3$). $F_e(\cdot)$ is the CDF of e_{it} and $F_e^{-1}(\tau)$ is the τ-th quantile of e_{it}. We set $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T = (-2, 2, 3)^T$, $m = 4$ and overall censoring percentage was 20% or 30%. We conducted two hundred simulations with sample size equal to 200. One thousand bootstrap samples were generated for variance estimation. For the dropout process we employed logistic regression model

$$logit(\eta_{it}) = \alpha_0 + \alpha_1 y_{i,t-1} + \alpha_2 x_i, \tag{5.16}$$

where the parameter vector $\boldsymbol{\alpha}$=$(\alpha_0, \alpha_1, \alpha_2)^T = (2.5, -0.3, 0)^T$.

Table 7 shows the comparison between CQR and CQR adjusting for dropout (D-CQR) using true π_{id_i} and estimated $\hat{\pi}_{id_i}$. To estimate $\hat{\pi}_{id_i}$, we used logistic regression model by replacing the censored observations of $y_{i,t-1}$ with half of detection limits. We present the results from median regression when 20% of data are left-censored and 75th percentile regression when censoring is increased to 30%. In general, D-CQR approach provides much better estimates than CQR for different quantile levels as indicated by much less bias in the estimates. The bootstrap estimators of SE are in good agreement with empirical estimators, and empirical 95% coverage rates are reasonable. The D-CQR estimates are very close to omniscient estimates when the dropout probability is known. If the estimates of dropout probability are used, we still obtained reasonable results, although the bias is getting bigger, especially in the estimate of intercept. It is consistent with the literature that good estimates of dropout process is critical when the inverse weighting approach is applied. We also found that when 30% of data are censored, along with the missing data due to dropout, median regression is no longer stable.

Table 7: Simulation Results of Censored Quantile Regression accounting for Dropout

β	$\beta_0 = -2$			$\beta_1 = 2$			$\beta_2 = 3$		
Method	**Omni**[a]	**D-CQR**[b]	**CQR**[c]	**Omni**[a]	**D-CQR**[b]	**CQR**[c]	**Omni**[a]	**D-CQR**[b]	**CQR**[c]
π_{id_i}, τ **=0.5 , 20% censored**									
Bias	-0.004	-0.008	0.037	0.002	0.001	-0.014	0.001	0.004	-0.022
SE	0.123	0.216	0.202	0.114	0.157	0.149	0.037	0.068	0.063
empSE	0.119	0.200	0.187	0.117	0.154	0.140	0.036	0.065	0.062
MSE	0.030	0.088	0.078	0.027	0.049	0.042	0.003	0.009	0.008
CP	0.940	0.945	0.925	0.910	0.935	0.975	0.945	0.950	0.930
π_{id_i}, τ **=0.75 , 30% censored**									
Bias	0.000	0.001	0.037	-0.004	-0.002	-0.012	-0.000	0.001	-0.022
SE	0.134	0.227	0.212	0.120	0.161	0.153	0.040	0.071	0.067
empSE	0.132	0.213	0.198	0.118	0.161	0.147	0.039	0.065	0.061
MSE	0.036	0.098	0.087	0.029	0.052	0.046	0.003	0.009	0.009
CP	0.950	0.965	0.965	0.955	0.945	0.955	0.970	0.960	0.955
$\hat{\pi}_{id_i}$, τ **=0.5 , 20% censored**									
Bias	-0.004	-0.022	0.037	0.002	0.009	-0.014	0.001	0.008	-0.022
SE	0.123	0.223	0.202	0.114	0.160	0.149	0.037	0.070	0.063
empSE	0.119	0.197	0.187	0.117	0.156	0.140	0.036	0.064	0.062
MSE	0.030	0.090	0.078	0.027	0.051	0.042	0.003	0.009	0.008
CP	0.940	0.960	0.925	0.910	0.950	0.975	0.945	0.940	0.930
$\hat{\pi}_{id_i}$, τ **=0.75 , 30% censored**									
Bias	0.000	-0.026	0.037	-0.004	0.005	-0.012	-0.000	0.008	-0.022
SE	0.134	0.232	0.212	0.120	0.163	0.153	0.040	0.073	0.067
empSE	0.132	0.214	0.198	0.118	0.156	0.147	0.039	0.067	0.061
MSE	0.036	0.101	0.087	0.029	0.052	0.046	0.003	0.010	0.009
CP	0.950	0.960	0.965	0.955	0.945	0.955	0.970	0.940	0.955

[a] **Omni**: Omniscient, [b] **D-CQR**: CQR adjusting for dropout,
[c] **CQR**: Censored QR

5.4 APPLICATION

The Genetic and Inflammatory Markers of Sepsis (GenIMS) study is a multi-center inception cohort study of 2320 subjects with community-acquired pneumonia presenting at the emergency departments between November 2001 and November 2003. One primary goal of the study is to identify important inflammatory markers that indicate the risk of severe sepsis and subsequent adverse outcomes. The markers of inflammatory and coagulation pathways were measured daily during the first week of hospitalization and weekly thereafter. We illustrate the proposed method with the pro-inflammatory cytokine, interleukin-6 (IL6), which is known to be elevated for patients with infection. Among the 1895 patients who were confirmed CAP cases admitted to hospitals, biomarker IL6 data were collected from 1188 patients. The objective of our analysis is to investigate whether IL6 changed over time in the first week of hospitalization and how IL6 trajectory was associated with the development of severe sepsis. Since our method assumes monotone missing pattern due to dropout, we exclude those patients who had intermittent missing data and end up with 1182 patients in the analysis cohort. The percentage of dropout was increased over time, half of the patients dropped out by day 5 and only 19% of patients had measurements up to day 7. The main reason for dropouts was discharged alive. The mortality rate is only 2.4% in the first week. It appeared that patients who had lower level of IL6 concentration (or in other words healthier), were more likely to drop out at later occasions. As we mentioned earlier, the low sensitivity assays used in GenIMS study introduced moderate to heavy censoring in the biomarker measurements. Cytokine IL6 was censored at either 2 or 5 depending on the assay used. The censoring proportion for IL6 was 26.2% overall and increased from 13.3% on day 1 to 37.3% on day 7. Figure 3 presents distribution of IL6 is still asymmetric after log transformation. We fitted the quantile regression model for natural log transformed IL6 with adjusted covariates such as age, gender, race and charlson comorbidity index. As described in section 5.2, the dropout probability can be calculated through the probability of being observed at each occasion. We applied logistic regression models to estimate the probability of being observed at the t-th occasion for the i-th subject, η_{it}. The initial models include covariates (age, gender, charlson comorbidity index) that are significantly associated with

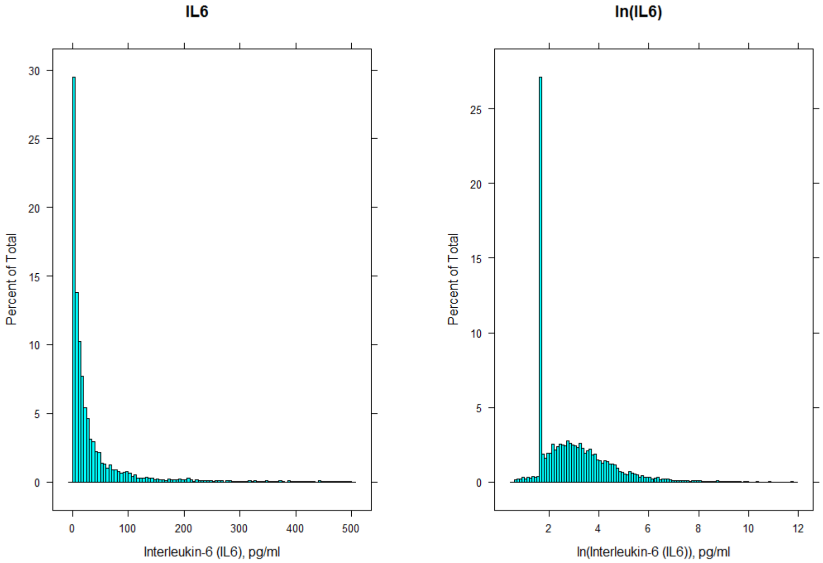

Figure 3: Histogram of IL6 and ln(IL6)

IL6; and the observed IL6 values up to occasion $t-1$. We chose the final models by stepwise selection method with cut point of p-value at 0.2. Table 8 summaries the models we used to estimate probability of being observed for each time point. It appears that older patients with higher IL6 level are less likely to have missing data. In Table 9, we present median regression and 75th percentile regression for IL6 using our method (D-CQR), and the standard CQR without accounting for dropouts. In median regression, both methods yielded similar significance and direction of covariate effects except for the main effect associated with severe sepsis (SS). Although the interaction between day and SS is significant in both analyses, sepsis did not show significant in the analysis using our method, but still significant in the CQR analyses. On the other hand, we can find the change in significance from the 75th percentile regression, especially in age and charlson comorbidity. Figure 4 presents the estimated median and 75th percentile for white males with median age 72 and charlson index>0. Two groups SS vs. NoSS are indicated by different colors and also the lines with triangles and circles represent D-CQR and CQR method, respectively. D-CQR method overestimates the quantile and underestimates the decreasing trend over time, especially for median of IL6 in NoSS group.

Table 8: Estimation Results for Missing Data Model

Pr(observed)[a]	**Parameter**	**Estimate**	**SE**	**P-value**
η_{i2}	Intercept	1.86	0.52	0.0002
	Age	0.01	0.01	0.1311
	Male	0.52	0.26	0.0449
η_{i3}	Intercept	-0.34	0.45	0.4518
	y_{i2}	0.43	0.09	< 0.00001
	Age	0.02	0.01	0.0091
	Charlson>0	0.46	0.22	0.0343
η_{i4}	Intercept	-1.27	0.33	0.0001
	y_{i2}	0.38	0.06	<0.00001
	Age	0.02	0.00	0.00001
η_{i5}	Intercept	-1.83	0.32	<0.00001
	y_{i3}	0.36	0.06	<0.00001
	Age	0.02	0.00	<0.00001
η_{i6}	Intercept	-1.85	0.40	<0.00001
	y_{i1}	-0.12	0.05	0.0080
	y_{i3}	0.27	0.12	0.0244
	y_{i4}	0.20	0.12	0.1036
	Age	0.02	0.00	0.0003
	Male	0.21	0.17	0.1999
η_{i7}	Intercept	-2.41	0.41	<0.00001
	y_{i1}	-0.08	0.04	0.0820
	y_{i4}	0.48	0.08	<0.00001
	Age	0.02	0.00	0.0009
	Male	0.21	0.15	0.1570

[a] **Pr(observed)**: Probability of being observed

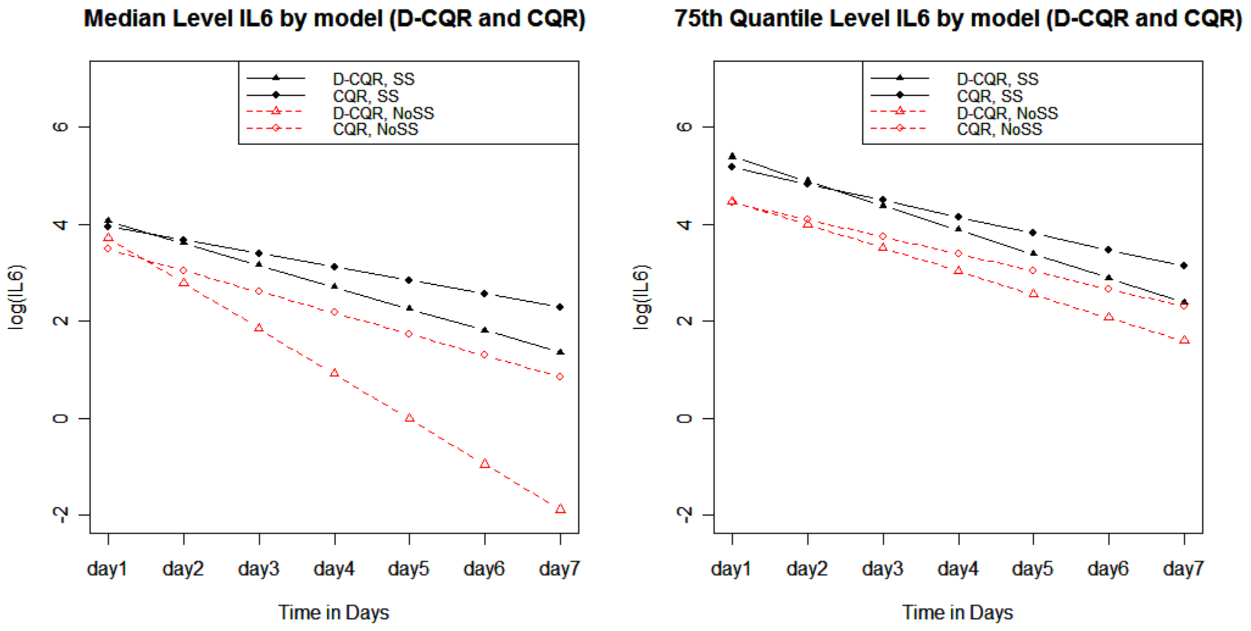

Figure 4: Quantile Regression for IL6 vs. Severe Sepsis (SS)

Table 9: Quantile Regression for IL6 vs. Severe Sepsis

Method	**CQR**[a]			**D-CQR**[b]		
Parameter	Est	SE	p-value	Est	SE	p-value
			Median Regression			
Intercept	3.45	0.21	<0.0001	3.99	0.39	<0.0001
SS	0.30	0.15	0.0530	-0.14	0.29	0.6273
Day	-0.44	0.02	<0.0001	-0.93	0.09	<0.0001
Day*SS	0.16	0.04	<0.0001	0.48	0.10	<0.0001
Age	0.00	0.00	0.1104	0.01	0.00	0.0828
White	0.11	0.09	0.2360	0.27	0.15	0.0694
Male	0.35	0.09	<0.0001	0.52	0.12	<0.0001
Charlson>0	-0.31	0.10	0.0011	-0.58	0.13	<0.0001
			75th **percentile regression**			
Intercept	4.50	0.27	<0.0001	4.40	0.32	<0.0001
SS	0.72	0.18	<0.0001	0.93	0.36	0.0092
Day	-0.35	0.02	<0.0001	-0.48	0.03	<0.0001
Day*SS	0.01	0.04	0.7585	-0.02	0.06	0.7240
Age	0.00	0.00	0.8770	0.01	0.00	0.0655
White	0.10	0.13	0.4731	0.06	0.12	0.5829
Male	0.25	0.10	0.0109	0.30	0.10	0.0040
Charlson>0	-0.09	0.11	0.3884	-0.25	0.13	0.0478

[a] **CQR**: Censored QR,
[b] **D-CQR**: CQR adjusting for dropout

5.5 DISCUSSION

Since biomarker data collected in biomedical studies are often highly skewed even after transformations, quantile regression models are increasingly used to complement the mean regression models. By selecting a set of quantile levels of interest, one can obtain a global assessment of treatment or covariate effect on the biomarker profiles. However, left censoring due to lower detection limit and missing data due to dropout hamper the use of standard quantile regression models. The estimation procedure for marginal quantile regression model is based on estimating equation approach. Thus, like the mean regression model based on generalized estimating equation method, quantile estimating equations are biased when the longitudinal responses are not missing at random. We applied inverse probability weighting technique to incorporate the dropouts in censored quantile regression. This method leads to consistent estimates of the quantile regression parameters provided that the model for dropouts is correctly specified. As shown in the simulation study, the proposed estimators have nice finite sample properties. The presented Bootstrap method provided reasonable variance estimator.

Although our estimators are not fully efficient since we used the working assumption of independence, our method is easy to implement with standard software packages that fit quantile regression. As noted in various contexts, incorporating correlations in the estimating equation may not appreciably improve the efficiency unless the repeated measurements are highly correlated. Censored quantile regression has been extended to data censored at both lower and upper thresholds, so our method can be directly extended to doubly censored biomarker data. Since we considered a MAR scenario for dropout process, the dropout probability may depend on the previous responses that can also be left censored. We replaced the censored observations by the half of the detection limit in the logistic regression model. The simulation results showed that such a naive method performed reasonably well if the censoring percentage was not high. In practice, it is common to have non-monotone pattern of missing data. Whether the results of Robins et al. [53] for this case is applicable for quantile regression merits further study.

6.0 CONCLUSION AND DISCUSSION

Motivated by the GenIMS study, we proposed several analysis methods to handle the left-censoring data in the biomarker measurements due to the sensitivity of given assay. we considered MI procedures based on Gibbs sampling method for multiple censored covariates. We extended such MI method to the left-censored marker data by accounting for the informative missing mechanism in the MI procedure. MI approach provides a practical and flexible solution for further complex analysis. Our MI methods performed well for low to moderate censoring data and can easily accommodate right-censored or interval-censored data. The simulation presented that if we ignore the correlations between censored markers then it may lead to biased estimates in the model when the correlation is high alone or moderate combined with high proportion of censoring. Since our method requires assumption of multivariate normal distribution, which may not be satisfied with marker data, appropriate transformation need to be considered. When the amount of missingness is not large, there is evidence [61] that inference made from MCMC based imputed data tend to be robust to departures from normal distribution. Whether this is the case for censored data and how the choices of prior distributions affect the results merit further study.

We considered a flexible and robust quantile regression for left-censoring longitudinal data and extend the censored QR approach to the approach accounting for the missingness due to dropouts. We first considered a censored median regression (CMR) model to accommodate the data censored at a fixed detection limit. We proposed an improved CMR estimator by incorporating the serial correlations in the estimating equation as done in Jung [33] for uncensored data. From the simulation study we found that our estimators performed well under various distributional assumptions. The improvement upon the estimator from the approach ignoring the intra subject correlation was much noticeable when censoring was heavy or the correlation was strong. Our estimating procedure is easy to fit with standard software and can be extended to general censored quantile regression models. In addition to the left-censoring problem, we also encountered missing data due to dropouts in GenIMS study. We applied inverse weighting technique accounting for the dropouts in censored quantile regression and our proposed estimators have nice asymptotic and finite sample properties. Bootstrap methods provided reasonable variance estimator and also our method is easy to implement with standard software package that fit QR. Usually people think the efficiency gain is minimal unless correlation is very high, but as we have seen from our simulation study, correlation matters if censoring is moderate or heavy. We need to have a better handling of censored response variable when using them to estimate the dropout probability. Since our approach only handles monotone missing pattern due to MAR dropouts, it is worth considering NMAR mechanism for dropout process. If we consider random effect model for QR, then we can apply a random effect model approach for both longitudinal process and dropout process. In this dissertation, the modeling with multiple longitudinal markers was not considered. Multiple biomarkers are often measured in biomedical studies to explore the mechanism of the disease development and progression. In GenIMS study, multiple markers were measured over time from the same or different potential pathways to better understand the biological mechanisms of sepsis. Because biomarkers from the same, or different, pathways are intrinsically correlated and likely to play roles interactively in the development of an adverse outcome, jointly modeling these biomarkers can greatly increase the efficiency and power of the analysis and provide more insight into their relationship with the treatment and clinical outcomes. The correlations between markers and the serial correlations within each individual maker should be taken into account in the joint analysis of multiple longitudinal markers. In the future, we will examine how the proposed censored quantile regression can be extended to simultaneously modeling the multiple markers in the same pathway or across the pathways.

BIBLIOGRAPHY

[1] Allison, P.D. (2002) Missing Data. Thousand Oaks, CA: Sage.

[2] Amemiya, T. (1984) Tobit models: A survey. *Journal of Econometrics*, **24**: 3-61.

[3] Angus, D. C., Yang, L., Kong, L., Kellum, J. A., Delude, R. L., Tracey, K. J., Weissfeld, L. A.; GenIMS Investigators (2007). Circulating high-mobility group box 1 (HMGB1) concentrations are elevated in both uncomplicated pneumonia and pneumonia with severe sepsis. *Crit Care Med* **35**(4):1061-7.

[4] Anneclaire J. De Roos et al. (2005) Persistent Organochlorine Chemicals in Plasma and Risk of Non-Hodgkin's Lymphoma *Cancer Research* **65(23)**: 11214-26.

[5] Austin, P.C., and Hoch, J. J. (2004). Estimating Linear Regression Models in the Presence of a Censored Independent Variable. *Statistics in Medicine* **23**, 411-429

[6] Bilias, Y., Chen, S., and Ying, Z. (2000). Simple resampling methods for censored regression quantiles. *Journal of Econometrics* **99** 373-386.

[7] Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest* 1992; **101**:1644-55.

[8] Bradley Efron (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics 7 (1)*: 1-26.

[9] Christian P. Robert (1995) Simulation of truncated normal variables. *Statistics and Computing* **5**: 121-125.

[10] D'Angelo G, Weissfeld L (2008) An index approach for the Cox model with left censored covariates *Stat Med*. Sep 30;**27(22)**:4502-14.

[11] Demirtas, H. (2004). Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods*, Volume **3, No 2**, 305-321.

[12] Dempster, A. P., Laird, N. M. Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **B, 39**, 1-38.

[13] Dufouil C, et al. (2004) Analysis of longitudinal studies with death and drop-out: a case study. *Stat. Med.* **23**:2215-2226

[14] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, **38**.

[15] Epstein, J.R.; Leung, A.P.; Lee, K.H. and Walt, D.R. High-density, microsphere-based fiber optic DNA microarrays. *Biosensors and Bioelectronics*, May 2003, vol. **18**, no. 5-6, p. 501-546.

[16] Fitzenberger, B. (1997). A guide to censored quantile regressions. In Handbook of Statistics, Volume 15: Robust Inference (ed. Maddala, G. S. and Rao, C. R.) 405-437. Amsterdam, North-Holland.

[17] Fitzmaurice GM, Laird NM and Ware JH. (2004). Applied Longitudinal Analysis. New York: John Wiley and Sons.

[18] Garland, T., Jr, A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.* **42**:265-292.

[19] Gelfand A.E. and A.F.M. Smith (1990) Sampling based approaches to calculating marginal densities *Journal of the American Statistical Association* **85**: 398-409.

[20] Gilbert RO. (1987). Statistical Methods for Environmental Pollution Monitoring. New York:Van Nostrand Reinhold.

[21] Gleit A. Estimation for small normal data sets with detection limits. *Environ Sci Technol.* 1985;**19**:1201-1206

[22] Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society,***Series B**, 30, 67-82.

[23] Han and Kronmal, (2004). Box-Cox transformation of left-censored data with application to the analysis of coronary artery calcification and pharmacokinetic data. *Statist. Med.* **v23**. 3671-3679.

[24] He, X. and Hu, F. (2003). Markov Chain Marginal Bootstrap. *Journal of the American Statistical Association* , Vol. 97, no. **459**, 783-795.

[25] Helsel DR. Less than obvious statistical treatment of data below the detection limit. *Environ Sci Technol.* 1990;**24**:1766-1774.

[26] Herring, A.H., Ibrahim, J.G., and Lipsitz, S.R.(2004) Nonignorable Missing Covariate Data in Survival Analysis: A Case Study of an International Breast Cancer Study Group Trial, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **53(2)**:293-310.

[27] Hogan JW, Lin X, Herman B (2004) Mixtures of varying coefficient models for longitudinal data with discrete or continuous nonignorable dropout. *Biometrics* **60(4)**:854-864

[28] Hornung R, Reed L. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg* 1990;**5**:46.51.

[29] Hughes, J. P. (1999) Mixed effects models with censored data with application to HIV RNA Levels. *Biometrics* **55**,625-629.

[30] Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005), Missing Data Methods for Generalized Linear Models:A Comparative Review, *Journal of the American Statistical Association*, **100**, 332.346.

[31] Jacqmin-Gadda, H., Thiebaut, R., Chene, G. and Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics* **1** 355-368.

[32] Jay H. Lubin et al. (2004) Epidemiologic Evaluation of Measurement Data in the Presence of Detection Llimits, *Environmental Health Perspectives*, vol. **112** No.17 pp. 1691-1696.

[33] Jung, S. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association* **91** 251-257.

[34] Keet, I. P., M. Janssen, P. J. Veugelers, F. Miedema, M. R. Klein, J. Goudsmit, R. A. Coutinho, and F. de Wolf. 1997. 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. *J. Infect. Dis.* **176**:665-671

[35] Kellum, J. A., Kong, L., Fink, M. P., Weissfeld, L. A., Yealy, D. M., Pinsky, M. R.; GenIMS Investigators (2007). Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study. *Arch. Intern. Med.* **167**:1655-63.

[36] Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., and Zhao, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: Application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Applied Statistics* **46**, 463-476.

[37] Little RJA (1988) Missing data in large surveys. *Journal of Business and Economic Statistics* **6**:287-301.

[38] Little RJA (1992) Regression with missing X's: a review. *Journal of the American Statistical Association* **87**:1227-1237.

[39] Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.

[40] Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd edition, New York: John Wiley.

[41] Lyles, R. H., Fan, D., Chuachoowong, R. (2001). Correlation coefficient estimation involving a left censored laboratory assay variable. *Statistics in Medicine* **20**, 2921-2933.

[42] Lyles, R. H., Lyles, C. M. and Taylor, D. J. (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. *J. Roy. Statist. Soc. Ser. C* **49** 485-497.

[43] Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. Statistics in Medicine 2001; 20:33-45

[44] Moschandreas DJ, Karuchit S, Kim Y, Ari H, Lebowitz MD, ORourke MK, et al. On predicting multi-route and multimedia residential exposure to chlorpyrifos and diazinon. *J Expo Anal Environ Epidemiol.* 2001a;**11**:56-65.

[45] Moschandreas DJ, Kim Y, Karuchit S, Ari H, Lebowitz MD, ORourke MK, et al. In-residence, multiple route exposures to chlorpyrifos and diazinon estimated by indirect method models. *Atmos Environ.* 2001b;**35**:2201-2213

[46] Moulton, L.H. and Halsey, N.A. (1995), A mixture model with detection limits for regression analysis of antibody response to vaccine, *Biometrics*, **51**, 1570-1578.

[47] Parzen MI, Wei LJ, Ying Z. (1994) A resampling method based on pivotal estimating functions. *Biometrika.* ;**81**:341-350

[48] Paxton WB, Coombs RW, McElrath MJ, et al. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with ¿ or = 400 CD4 lymphocytes: implications for applying measurements to individual patients. National Institute of Allergy and Infectious Diseases AIDS Vaccine Evaluation Group. *J Infect Dis.* 1997;**175**:247-254.

[49] Persson T, Rootzen H. (1977) Simple and highly efficient estimators for a type I censored normal sample. *Biometrika.* **64**:123-128.

[50] Powell, J. L. (1984). Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics* **25** 303-325.

[51] Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics* **32** 143-155.

[52] Rigobon, Roberto and Stoker, Thomas M. (2004), Censored Regressors and Expansion Bias, *Massachusetts Institute of Technology (MIT), Sloan School of Management Working papers* 4451-03.

[53] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data. emphJournal of the American Statistical Association **90**, 106.121.

[54] Ronco C., Bellomo R., Lonneman G., Agarwal P. K., Kumari R., Netea M. G., Van der Meer J. W., Kullberg B. J., Hotchkiss R. S., Karl I. E. *N Engl J Med* 2003; **348**:1600-1602.

[55] Rotnitzky and Robins, (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine.* **v16**. 81-102.

[56] Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. emphJournal of the American Statistical Association **93**, 1311.1339.

[57] Rubin, D. B. (1976), Inference and Missing Data *Biometrika*, **63**.

[58] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

[59] Russell, J. A. (2006). Management of Sepsis. *NEJM* **355**: 1699-1713.

[60] Schafer, A. (1997) Prosodic Parsing: The Role of Prosody in Sentence Comprehension Ph.D.dissertation, University of Massachusetts, Amherst.

[61] Schafer, J.L. (1997). Analysis of incomplete multivariate data, New York: Chapman & Hall.

[62] Scharfstein, D., Rotnitzky, A. and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). emphJournal of the American Statistical Association **94**, 1096-1120.

[63] Siddhartha Chib (1992), Bayes Inference in the Tobit Censored Regression Model, *Journal of Econometrics*, **51**: 79-99.

[64] Siddhartha Chib and Edward Greenberg (1996), Markov Chain Monte Carlo Simulation Methods in Econometrics *Econometric Theory*, **12**: 409-431.

[65] Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. *Chapman and Hall: London.*

[66] Sujuan Gao and Rodolphe Thiebaut, Mixed-effect Models for Truncated Longitudinal Outcomes with Nonignorable Missing Data, *Journal of Data Science*, **v.7, no.1**, p.27-42

[67] Tanner, M. A. (1993) Tools for statistical inference (2nd ed.), New York: Springer-Verlag

[68] Tanner, M. A.and Wong, W.H. (1987) The calculation of posterior distribution by data augmentation *Journal of the American Statistical Association* **82**: 528-540.

[69] Thiebaut R, Jacqmin-Gadda H. (2004) Mixed models for longitudinal left-censored repeated measures. *Comput Methods Programs Biomed.* **74(3)**: 255-260.

[70] Thiebaut RH, Jacqmin-Gadda H, Babiker A, et al. Joint modeling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Stat Med*. 2005;**24**:65-82.

[71] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Ecomometrica* **26**, 24-36.

[72] Tobin, James (1958). Liquidity Preference as Behavior Towards Risk. *Review of Economic Studies* **25.1**:65-86

[73] Wang, J. H., and Fygenson, M. (2009). Inference for Censored Quantile Regression Models in Longitudinal studies. *Annals of Statistics*, to appear.

[74] Wolfinger, R., and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233-243.

[75] Wu, L. (2002). A Joint Model for Nonlinear Mixed-Effects Models with Censoring and Covariates Measured with Error, with Application to AIDS Studies. *Journal of the American Statistical Association* **97** 955-964.

[76] Yende, S., D'Angelo, G., Kellum, J. A., Weissfeld, L. A., Fine, J., Welch, R. D., Kong, L., Carter, M., Angus, D. C. (2008). Inflammatory markers at hospital discharge predict subsequent mortality after pneumonia and sepsis. *Am J Respir Crit Care Med* **177**(11):1242-7.

[77] Yi, G. Y. and He, W. (2008). Median Regression Models for Longitudinal Data with Dropouts. *Biometrics*, to appear.