APPLICATION OF SEMIPARAMETRIC METHODS FOR REGRESSION MODELS WITH MISSING COVARIATE INFORMATION

by

Gina D'Angelo

BS, Ohio University, 1994

BSEd, Ohio University, 1994

ScM, Johns Hopkins University School of Public Health, 1998

Submitted to the Graduate Faculty of

The Department of Biostatistics

The Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Gina D'Angelo

It was defended on

January 27, 2005

and approved by

Dissertation Advisor: Lisa A. Weissfeld, Ph.D., Professor,
Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Committee Member: H. Samuel Wieand, Ph.D., Professor,
Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Committee Member: Gong Tang, Ph.D., Assistant Professor,
Department of Biostatistics,
Graduate School of Public Health, University of Pittsburgh

Committee Member: Jong-Hyeon Jeong, Ph.D., Assistant Professor,
Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Committee Member: Bret Goodpaster, Ph.D., Associate Professor,
Department of Medicine
School of Medicine, University of Pittsburgh

Lisa A. Weissfeld, Ph.D.

APPLICATION OF SEMIPARAMETRIC METHODS FOR REGRESSION MODELS WITH MISSING COVARIATE INFORMATION

Gina D'Angelo, Ph.D.

University of Pittsburgh, 2005

This dissertation addresses regression models with missing covariate data. These methods are shown to be significant to public health research since they enable researchers to use a wider spectrum of data. Unbiased estimating equations are the focus of this dissertation, predominantly semiparametric methods utilized to solve for regression parameters in the presence of missing covariate data. The first aim of this dissertation is to evaluate the properties of an efficient score, an inverse probability weighted estimating equation approach, for logistic regression in a two-phase design. Simulation studies showed that the efficient score is more efficient than two other pseudo-likelihood methods when the correlation between the missing covariate and its surrogate is high.

The second aim of this dissertation is to develop a methodology for left truncated covariate data with a binary outcome. To address this problem, we proposed two methods, a likelihood-based approach and an estimating equation approach, to estimate the coefficients and their standard errors for a regression model with a left truncated covariate. The estimating equation technique is close to completion, and once solved should be the most efficient method. The likelihood-based method is compared to standard methods of filling in the truncated values with the lower threshold value or using only the nontruncated values. Simulation studies demonstrated that the likelihood-based method has the best variance

correction and moderate bias correction. The application of this method is illustrated in a sepsis study conducted at the University of Pittsburgh.

ACKNOWLEDGEMENT

I would like to acknowledge my advisor, Dr. Lisa Weissfeld, for her support of this dissertation. Dr. Weissfeld has been an incredible mentor showing me how to become a researcher and independent thinker. I am very grateful for her invaluable insight, expertise, patience, innovative nature, and motivation. This thesis would not have been possible without her and for that I most grateful. I know it was a long six years but definitely worth it.

I would also like to thank my committee, Dr. Gong Tang, Dr. Sam Wieand, Dr. Jong-Hyeon Jeong, and Dr. Bret Goodpaster, for their suggestions and time. The support, understanding, and insightful suggestions from Dr. Wieand made this thesis possible. I am also very grateful for the intuitive reasoning and simulation/analysis suggestions offered by Dr. Tang. These proved to be instrumental in the progress and completion of my dissertation.

Without the unending support of the faculty, staff, and students in the Biostatistics Department my tenure would have been a bumpier road. Dr. Howard Rockette has been such an inspiration. He has always supported the students academically, professionally, and emotionally. I owe a debt of gratitude to my teachers, Dr. Rockette, Dr. Sati Mazumdar, Dr. Weissfeld, and Dr. Roslyn Stone, for the knowledge I gained. I referred to their notes infinite times as they proved incredibly useful for my dissertation.

The students provided me with the support, drive and ambition needed to succeed. The study sessions with my fellow students will be memorable. I was impressed with the students genuine concern for one another. I wanted to especially thank Zorana, Zek, Wei, Kari, Iva,

and Ami for their friendship and academic support. Ami and Iva, I will always remember DSO and Sanctuary.

The UPCI Biostatistics Facility was one of the more integral part of my experience at Pitt. I am incredibly thankful for being mentored by Bill Gooding and Dr. Sam Wieand. The various data analysis techniques and clinical study experience I gained from the staff and collaborators was an invaluable learning experience. I wanted to thank Bill for providing me with knowledge, collaboration experience, and challenging projects.

Last but not least I would like to thank my parents, Debbie and Mike, and my siblings, Matt, Meri, Jen, Ren, and Stimpy. My family has always believed in me and been there for me. It is because of their love, devotion, and support that I was able to achieve all my accomplishments and complete this Ph.D. program.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The problem of missing data plagues many different studies. While the issue of missing data in the longitudinal setting has been examined in some detail, cross-sectional and case-control designs have not received as much attention. With the advent of new and more expensive technologies in medicine, the need for innovative approaches to the handling of missing data at both the design and analysis stage is necessary. The work of Robins *et al.* (1994), Bickel *et al.* (1993), and Lawless *et al.* (1999) points to the generality of the "missing data" problem. They discuss that outcome-based sampling schemes, errors-in-variable, censored data, and truncated data can be viewed as "missing data". Viewing the problem more generally has led to the development of an updated approach to estimation in the semiparametric literature. Through the clever use of estimating equations, one can provide statistical methods for the analysis of missing data from many different settings.

The goal of these estimating equation techniques is to obtain an estimator of the parameter of interest. These estimating equations sum to zero and are a function of the parameter and data. Ideally, it is of interest to find an estimating equation that is unbiased and

optimal (Godambe, 1991). A benefit of the unbiased estimating equation is a reduction in bias. The optimality property implies efficiency. The solution of these unbiased estimating equations is in fact the estimate of the parameter of interest. Once the optimal unbiased estimating equation is found, it has been shown that the estimator is consistent, asymptotically normal, and efficient, all desirable properties (Godambe, 1991; Bickel *et al.*, 1993; Robins *et al.*, 1994).

Estimating equations are a rich class of estimators flexible enough to model normal and non-normal data under various types of designs and frameworks. These techniques were introduced to the missing data literature to obtain efficient estimators of the parameter of interest in the presence of nuisance parameters. In an attempt to gain efficiency, information is drawn from both the complete and incomplete cases. Semiparametric methods have predominantly focused on estimation rather than model building, thereby limiting available inferential techniques. Thus, the only tool developed for inference with semiparametric techniques is the Wald test, while no tools comparable to the likelihood ratio test (LRT) or Akaikes Information Criterion (AIC) are available. These techniques are newer and gaining popularity due to their flexibility; however, the conceptual complexity can be a deterrent. A majority of the estimating equation techniques for missing data have been developed as a result of design issues, since the amount of incomplete data can be reduced by addressing it during the study design phase.

Sampling based approaches are applied to case-control and cohort studies to either balance data or reduce the cost of data collection; thereby, improving precision and eliminating the bias of coefficient estimates. As a motivational example for sampling by design, a sepsis study initiated by the University of Pittsburgh was designed to determine the relationship

between severe sepsis status and biomarker data. As new biomarkers became available it was not feasible to measure the biomarker for all subjects in the cohort. To design a meaningful sub-study, data were collected for a subset of subjects within categories of severe sepsis status, death status, and initial health state.

Another common example of missing data is truncated data. Truncated data arise when a variable is observed within a prespecified range of values. This is a common occurrence with laboratory data such as measuring blood samples for immunology assays. An example of truncated data in the above mentioned sepsis study is that of the inflammatory marker data. A panel of inflammatory markers was collected in a large portion of the cohort; however, assays for most of these markers are not very sensitive. This data will serve as an example for analysis.

Most consider "missing data" to be missingness by happenstance. Missing by happenstance occurs when at least one of the variables is not completely observed/reported for all subjects and the reason for the missing data is not exactly known. The inflammatory marker data in the above mentioned sepsis study provides an example of missingness by happenstance. Some of the marker data is missing with the reason unknown possibly due to administrative reasons.

A missing data mechanism is a tool which explains the cause for missing data and describes the relationship between the missing data indicator and the variables. Two types of missing data mechanisms are ignorable and nonignorable missing data (Rubin, 1976). Ignorable missing data include data that are missing completely at random (MCAR) and data that are missing at random (MAR).

For descriptive purposes, we specify R as the missing data indicator and the complete data as (Y, Z), where $Z = (X, V)$ and X is incomplete. If missingness is independent of all the variables, then $P(R = 1|Y, Z) = \alpha$, where α is a constant, and the data are MCAR. Under MCAR the observed values, X_{obs}, are a simple random sample of X; that is the distribution of the missing values, X_{mis}, is the same as X_{obs}. If missingness is dependent on the fully observed variables, then $P(R = 1|Y, Z) = P(R = 1|Y, V)$ and the data are MAR. Within each subclass of (Y, V) the observed values are a random sample of X; that is within each subclass of (Y, V) the distribution of X_{mis} is equivalent to X_{obs}.

Not missing at random (NMAR) falls under nonignorable missing data. If conditioned on the observed data, missingness is dependent on the unobserved values, then the data are NMAR. This dissertation is concerned with covariates that are MAR and NMAR. The missing by design problem will be defined as MAR and the truncated problem is defined as NMAR.

The main types of missing data methodology include imputation methods, likelihood-based approaches, estimating equation procedures, and complete and available case analysis methods. Two naive approaches are exclusion of the missing covariate and complete case analysis. Excluding the missing covariate can lead to model misspecification. Complete case is defined as a case without a missing value. Complete case (CC) analysis is characterized as performing standard statistical analysis on complete cases and is the simplest measure to address incomplete data since no methodological modifications are necessary. However, depending on the extent of incomplete cases and the cause for incomplete data there is a potential for bias and a loss of precision. A loss of precision occurs when there is a loss in information. Regardless of the cause of the missing data mechanism, as the loss of infor-

mation increases so does the variance. If the data are MCAR, then the complete cases are a simple random sample of all cases and estimates of parameters will be unbiased. Outside of the MCAR framework potential bias should be addressed. Given a fully observed data set, maximum likelihood estimates are asymptotically normal, asymptotically efficient, and consistent. Since regression models are the focus here and they often depend on likelihood theory, we are interested in developing estimators from non-likelihood based methodology that hold the same properties.

The missing data pattern and data type determine the type of method chosen to account for missing data. A common approach for addressing missing data is likelihood methods. However, when the covariate data are incomplete, specification of a full likelihood is required. In addition, the distribution of the covariates must also be specified. The likelihood-based approach proves difficult with covariates of high dimension due to the complexity of the distributions. Semiparametric methods, a class of estimating equations, are an alternative approach to specifying a full likelihood. As opposed to a likelihood-based approach, specification of the missing data mechanism is required and the distribution of the covariates is left unspecified. Unbiased estimating equations are a general technique yielding efficient estimates for regular estimates under certain conditions to be specified in Chapter 2.

Two comparable methods that address efficiency utilizing estimating equations are the efficient score function (Nan *et al.*, 2002) and the efficient influence function (Robins *et al.*, 1994). The efficient score function is considered a semiparametric approach where information bounds are obtained via scores and score operators. The efficient influence function is considered a nonparametric approach where information bounds are obtained via derivatives of functions. The efficient score and influence function are functions of each

other; thereby, construction of one can aid in construction of the other given the correct functions.

The efficient score approach has been developed for missingness by design where properties have been evaluated for a two-phase design with a time-to-event outcome. The influence function method has been developed for data missing by design or by happenstance in a general framework. Our intent is to apply the score function technique and extend it for covariates missing by design and truncated with a binary outcome.

1.1 OBJECTIVES

Estimation and prediction in a wide application of statistical models for non-normal data in the presence of missing data are the focus of this dissertation. Throughout this thesis, historical and proposed techniques for estimation of coefficient parameters and their covariance matrix will be discussed and developed. Since the Wald test is the inference tool utilized, properties such as consistency and asymptotic normality of an estimator must be met for the Wald test to be valid. Semiparametric methods are in their infancy providing a framework to address a wide class of additional problems that will be discussed in Chapter 6.

Regression models are a class of models commonly used to analyze medical studies. This dissertation specifically focuses on a binary outcome. Regression parameters are biased and inefficient when covariates are MAR. The first aim of this thesis is to evaluate the properties of the efficient score (Nan 2002), an inverse probability weighted estimating equation approach, for logistic regression of a MAR covariate under missing by design. Prior to

this publication large sample properties of this efficient score had not been evaluated and compared to other estimates under our framework.

The second aim of this thesis is to develop a methodology for truncated covariate data with a binary outcome. To date no methods exist for regression models with truncated covariates. We propose a likelihood-based approach and a semiparametric approach via score functions to obtain estimates of regression parameters and the covariance matrix.

1.2 SUMMARY

The layout of the dissertation will be as follows. Chapter 2 will review literature to handle MAR covariates in regression models and describe estimating equations in greater detail. Chapter 3 will include simulation studies for logistic regression with missing covariate data to evaluate and compare the properties of three estimators. Chapter 4 will review truncated data and develop our proposed extensions for modelling a binary outcome while adjusting for a truncated covariate. Chapter 5 will include simulation studies for logistic regression with a truncated covariate to evaluate and compare the properties of three estimators. Chapter 6 will include a discussion and describe future work.

CHAPTER 2

METHODOLOGY OF REGRESSION AND MISSING DATA

2.1 LOGISTIC REGRESSION

Logistic regression is a statistical tool that allows us to study relationships between a binary outcome variable and covariates. The outcome, $Y \in \{0, 1\}$, is binary and always observed. The covariates, $\mathbf{Z} = (\mathbf{X}, \mathbf{V})$, can be a mixture of discrete and continuous variables where $\mathbf{V}$ is always observed and $\mathbf{X}$ is possibly missing. A method of estimation for parameters of the logistic model which yields desirable properties is maximum likelihood. The likelihood function for $(Y, \mathbf{Z})$ is

$$L(\beta; y, \mathbf{z}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{Z}_i'\boldsymbol{\beta})} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{Z}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{Z}_i'\boldsymbol{\beta})} \right)^{1-y_i}.$$

The estimates of $\boldsymbol{\beta}$ are solved by differentiating the log likelihood with respect to $\boldsymbol{\beta}$ and setting these equations equal to zero

$$l_\beta = \sum_{i=1}^{n} \mathbf{z}_i \left(y_i - \frac{\exp(\mathbf{z}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{z}_i'\boldsymbol{\beta})} \right) = 0. \tag{2.1}$$

Equation 2.1 is also the score equation. The score functions are nonlinear in the parameters so an iterative procedure is employed to obtain maximum likelihood estimates of the parameters. Further description of this iterative procedure can be found in McCullagh *et al.* (1983). When the data are completely observed, the estimates of $\boldsymbol{\beta}$ are consistent, asymptotically normal, and asymptotically efficient. If the data are MCAR and complete case analysis is used, then the above mentioned properties still hold with the exception of efficiency.

Logistic regression models are a member of the class of generalized linear models (GLM) (McCullagh and Nelder, 1983). Generalized linear models use the following structure: $y = E(y|\mathbf{z}) + \varepsilon = u + \varepsilon$ and $\eta_i = g(u_i) = \mathbf{z}_i'\boldsymbol{\beta}$. The three components of a GLM are the systematic component η, the random component ε, and the link function $g(u)$. The random component, ε, measures the variability of Y after accounting for all systematic variability with inclusion of the covariates. Here, the "link" function $g(u)$ is chosen by the analyst to obtain a reasonable range for the linear function $\mathbf{Z}'\boldsymbol{\beta}$ and to describe the data adequately.

Since the outcome is binary, $Y \in \{0, 1\}$, the expectation of Y given $\mathbf{Z}$, $u = \frac{\exp(\mathbf{Z}'\boldsymbol{\beta})}{1+\exp(\mathbf{Z}'\boldsymbol{\beta})}$, is bounded by 0 and 1. The conditional mean of Y given $\mathbf{Z}$, u, is nonlinear in the parameters indicating that it is necessary to choose the logit link for the transformation of u which is defined as

$$g(u) = \ln\left[\frac{u}{1-u}\right] = \beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p. \tag{2.2}$$

The logit link function is strategically chosen since $g(u)$ is continuous, can range from -∞ to ∞, and is linear in the βs.

The form of the random part, ε, of the model will be chosen by the probability frequency function which describes the distribution of the outcome variable. Depending on the probability model chosen, the variance may be a function of the mean. That is var(y_i)=aV(u) where a=σ^2 and V(u) is a function of the mean. The errors are binomially distributed with mean zero and variance $u(1-u)$.

2.2 MISSING DATA METHODOLOGY

Numerous approaches are available for regression of missing data problems. Three types that will be discussed are likelihood-based, imputation, and estimating equation methods. Estimating equation methods will be discussed in greater detail in the next section.

Likelihood-based procedures (Rubin, 1976; Little and Rubin, 1987) are a large class of procedures commonly implemented by statisticians. The general notion is to define a model for the observed data and draw inferences from this model. The disadvantage of this approach is that one has to specify a full likelihood and a distribution for the covariates. If the covariates are of high dimension this can be difficult. An advantage of this approach is that the missing data mechanism need not be specified under MAR.

Under simple missing data patterns and specified distinct sets of parameters the likelihood can be factored leading to simple inferences for the parameters. Under most conditions the likelihood cannot be factored and the EM algorithm must be employed to solve for the parameter estimates. The EM algorithm is computationally intensive, and will i) converge

slowly if a large portion of data is missing and ii) have no solution if a closed form does not exist during the maximization stage.

Imputation is a common approach for the handling of missing data and is an option in standard statistical packages such as Stata. Imputation is characterized as filling in missing values. A predictive (or joint) distribution is defined for the missing data from which values (draws) are randomly selected from this distribution. In 1987 Rubin introduced multiple imputation, which is of Bayesian influence, to account for uncertainty of these randomly selected values. Advantages of this method are that the parameters are permitted to have high dimensions, the method is computationally and conceptually simpler than other methods, and the distribution is exact and does not rely on asymptotic approximations (Schafer 1997). The disadvantage of imputation is that it is model-based.

Estimating equations are desirable methods due to their applicability to a wide range of missing data settings. Not only can the techniques address missing by happenstance, but they can readily be extended to handle a general missing data setting including outcome-based sampling schemes, errors-in-variable, censored data, and truncated data. The beauty of these methods is that estimating equations are functions of the score equations which are well understood. In addition, the full likelihood need not be defined. Essentially these equations can be viewed as modifications, or functions, of the score equations. If one can manipulate a function of the score equation to handle the type of missingness of interest, then these estimating equations can be applied.

Weighted estimating equation techniques have been designed for missing by happenstance and outcome-based sampling schemes. These include pseudo-likelihoods and semiparametric methods, which will be reviewed more thoroughly and evaluated via simulations. The

idea is to assign the same weight to a group of subjects who have similar characteristics compensating for the subjects excluded from the analysis. These weights are inverted probabilities of selection. For example, if 30% of subjects who are over 60 years old and have disease are missing a CD8 count, then all subjects who are over 60 years old and have disease are assigned a weight of 1.4.

Weighted estimating equations borrow estimating procedures from the survey sampling literature applying the Horvitz-Thompson estimator. Lawless *et al.* (1999) cover semi-parametric methods for response-selective and missing data problems for regression analysis. The complexity of these methods vary, but the advantage is that the distribution of the covariates does not need to be specified. A complication of these techniques includes specification of the missing data mechanism and the conceptual aspect. Three methods that will be discussed in greater detail under the logistic section and compared in the following chapter are conditional maximum likelihood, weighted pseudo-likelihood, and the efficient score.

2.2.1 Survey Sampling

Survey sampling techniques have strongly influenced missing data methodology. It is crucial to understand survey sampling methodology to gain insight into weighted estimating equations. The landmark paper by Horvitz and Thompson (1952) set the stage for selective design and missing data methodology. Horvitz *et al.* developed a general technique for improving any statistic when a random sample with unequal probability within subclasses of a finite population is selected. The Horvitz-Thompson estimator, defined as a weighted

mean $\frac{\sum_j \pi_j^{-1} y_j}{\sum_j \pi_j^{-1}}$, was originally intended to address survey biased sampling. The statistic was restricted to descriptive statistics, such as the mean and variance.

Horvitz *et al.* developed an estimator under two cases. An unbiased linear estimator and unbiased estimator of the sampling variance were developed for a one and two phase sampling technique, where selection probabilities are defined a priori and used to select a subsample from a finite population. Studies that rely on these concepts were intended to increase precision in the presence of information loss. Although Horvitz *et al.* were aware that this method reduces the variance they did not address which estimator would yield a minimum or "optimal" variance. As a result, various extensions were proposed over the next 50 years.

Prior to 1974 these weighting techniques were restricted to descriptive statistics. Kish and Frankel (1974) made a major contribution by extending the Horvitz-Thompson estimator to complex statistics and designs such as confidence intervals and inference for regression models. Kish *et al.* also felt that "traditional" survey sampling methods, such as the Horvitz-Thompson estimator, could be implemented outside of the realm of survey sampling design. Survey samples tend to have large sample sizes so the asymptotic results often hold. Sample size issues and asymptotics may pose a problem and need to be addressed with other designs and types of missing data problems due to limited sample size.

Manski and Lerman (1977) developed a weighted estimating equation for complete data which used the Horvitz-Thompson approach. Manski *et al.* clearly defined a general framework and statistical model (estimating approach) for regression models under choice-based sampling. This approach can cause a loss of efficiency since it uses only complete data, but

attempts to gain efficiency by assigning larger weights to the complete pseudo-likelihood, accounting for incomplete cases.

In an effort to increase efficiency of the estimates of regression coefficients, Robins, Rotnitzky, and Zhao (1994) developed a weighted estimating approach based on semiparametric methods and influence functions. Robins *et al.* were the first to develop a semiparametrically efficient estimator for regression models with incomplete covariates. These inverse probability weighted estimating equation (IPWE) methods were shown to have desirable properties and to be flexible enough to handle MAR data under any type of regression problem and missing by design/happenstance. This class of estimators is referred to as IPWE and has prompted many other researchers to pursue extensions of Robins' IPWE method.

2.2.2 Response-Selective Designs

Response-selective designs can be considered missing data problems. In a conventional case-control study the outcome is fixed and considered a stratification variable. Within each strata, subjects are randomly chosen and covariate information is collected for all subjects. Logistic regression is the standard analysis for a case-control study. In this case, the proportion of case/control selection need not be known and specified. If the disease is rare these studies greatly reduce the amount of data collection. If balance issues or confounding arises, it is beneficial to stratify on the covariates to improve precision and eliminate bias of coefficient estimates. This is referred to as a stratified case-control study. Since the ratio of the probabilities of cases and controls varies by strata, alternative statistical methods to standard logistic regression must be utilized.

Cohort studies are prospective by design and covariate information is initially collected on all subjects, who are followed for a set time in which the outcome is measured. In many cohort studies thousands of subjects are enrolled to determine the risk factors of the outcome. Collection of data can be an arduous and expensive task possibly inducing missing data. Response-selective designs are efficient methods preventing loss of information.

Numerous designs fall under the cohort design. A common cohort design is a case-cohort design. A typical case-cohort design measures covariate information on all cases and a subset of controls. Another option is to fully measure a set of covariates V and outcome Y and randomly sample a subset of subjects within each strata defined by (Y, V). This is considered a stratified case-cohort study. A condition of these cohort studies is that the variable collected for the subsample of patients cannot be time-varying.

2.2.3 Missing Data Mechanism

For application of weighted estimating equation techniques the mechanism that generates missing data must be specified. Under the sampling based scheme, prespecified selection probabilities are often used, where $p_j = \frac{n_j}{N_j}$, n_j is the total number of subjects selected within each strata j, and N_j is the total number of subjects within each strata j. Other suggestions for estimating these probabilities after data collection are i) to use a ratio of the number of fully observed subjects to the number of subjects within each strata $\widetilde{p}_j = \frac{\widetilde{n}_j}{\widetilde{N}_j}$; or ii) to use the logit model adjusting by the fully observed data $P(R = 1|Y, V; \xi)$ where ξ is unknown. Under missing by happenstance, one could use specified probabilities suggested by a clinician familiar with the type of data being analyzed or estimate the probabilities

by either an empirical or a model based estimate. Robins *et al.* (1994) and Lawless *et al.* (1999) claim that there is a gain in the efficiency of the estimate of β if π, the probability of missingness, is estimated since more information is drawn from the data. This result also holds if π is known.

2.3 EFFICIENT SCORE AND INFORMATION BOUND FOR REGRESSION MODELS

This section will describe the efficient score in a general setting. In order to obtain a better understanding, the efficient score is described in greater detail in a general setting prior to describing the three methods specifically applied for logistic regression. Most of the concepts in this section are based on research from Robins *et al.* (1994) and Nan *et al.* (2002). They used different approaches to develop comparable methods. Robins utilized influence functions to solve the efficient influence function while Nan employed the score operator approach. For more details and proofs refer to Nan *et al.* (2002) and Robins *et al.* (1994).

The complete data are $U^o = (U_1^0, U_2^0) \sim Q$, where U_1^0 is fully observed and U_2^0 is partially observed. In this dissertation we will only deal with a parametric (logistic regression) model $\mathcal{Q}$= $\{Q_{\beta,\gamma} : \beta \in \boldsymbol{\beta} \subset \mathbb{R}^k, \gamma \in G\}$, where γ is the nuisance parameter and β is the parameter of interest. The observed data are $U = (U_1^0, U_2^0, R)^R (U_1^0, R)^{1-R} \sim P$ and the model is semiparametric where $\mathcal{P}$= $\{P_{\beta,\gamma} : \beta \in \boldsymbol{\beta} \subset \mathbb{R}^k, \gamma \in G\}$. Note that R is a missing data indicator, where $R = 1$ indicates fully observed and $R = 0$ indicates partially observed.

Since the data are MAR, the probability of missingness is modelled by $\pi(R = 1|U_1^0)$ with a restriction that this probability be greater than zero.

The density and likelihood of the observed data are

$$\begin{aligned} p_{\beta,\gamma}(u) &= (q(r,u^o))^r \left(\int q(r,u^o) d\mu(u_2^0) \right)^{1-r} && (2.3) \\ &= \left(\pi(u_1^0) q_{\beta,\gamma}(u^o)\right)^r \left(\left(1 - \pi(u_1^0)\right) \int q_{\beta,\gamma}(u^o) d\mu(u_2^0) \right)^{1-r} && (2.4) \\ L(\beta,\gamma) &= \prod_{i=1}^{n} p_{\beta,\gamma}(u_i), \end{aligned}$$

where $q_{\beta,\gamma}(u^o)$ is the density of the complete data. Equation 2.3 is general enough to be applied to any missing data setting. One can define R and the density $q(r, u^o)$ to suit their missing data problem. We will show throughout this thesis the flexibility of this model and the range of problems that can be solved using this approach. Note that Equation 2.4 includes the probability of selection and the distribution of the covariates.

Typically, the above likelihood does not require one to model $g(\mathbf{z})$, the distribution of the covariates, as $\mathbf{Z}$ is ancillary. However, if the probability of selection depends on the outcome, then the covariates are no longer ancillary. Once $\mathbf{Z}$ is no longer ancillary the distribution $g(\mathbf{z})$ must be modelled. In addition, likelihood-based methods will force the probability of selection, π, to drop out under MAR.

For estimation problems where the data are incomplete, complete case analysis is inefficient and the score operator provides an alternative for this setting. The score for the observed data for each subject i is defined as :

$$\begin{aligned} \dot{l}_{i,\beta} &= R_i \dot{l}^0_{i,\beta} + (1 - R_i) E\left(\dot{l}^0_{\beta} | U^0_{i,1} \right) \in \dot{\mathcal{P}} \\ \dot{l}_{i,\gamma} &= R_i \dot{l}^0_{i,\gamma} + (1 - R_i) E\left(\dot{l}^0_{\gamma} | U^0_{i,1} \right) \in \dot{\mathcal{P}}, \end{aligned}$$

where $\dot{l}^0_\beta$ and $\dot{l}^0_\gamma$ are scores for the complete data. However, the solutions to these score equations do not give efficient estimates since both score equations are contained in the tangent space of $\mathcal{P}$, denoted by $\dot{\mathcal{P}} = \dot{\mathcal{P}}_\beta + \dot{\mathcal{P}}_\gamma$. The tangent space of $\mathcal{P}$ is a linear span of all scores of every submodel of $\mathcal{P}$ at P, which is basically a collection of scores for that model. The tangent space for the complete data distribution, $\mathcal{Q}$, is denoted $\dot{\mathcal{Q}} = \dot{\mathcal{Q}}_\beta + \dot{\mathcal{Q}}_\gamma$.

In Equation 2.4 the distribution of the covariates, $g(\mathbf{z})$, must be estimated. However, $\gamma = g(\mathbf{z})$ is not the inferential target, permitting $g(\mathbf{z})$ to be estimated nonparametrically. Since $g(\mathbf{z})$ is nonparametric, Equation 2.4 is a semiparametric model. A semiparametric method was used to solve this problem, and is a tool used to place a semiparametric estimating equation in the appropriate space in order to estimate the parameter of interest. The ultimate goal is to obtain an estimate of β; thereby, removing the influence of the nuisance parameter γ. A solution is to restrict the score to the appropriate space. Following semiparametric methodology, an approach is to use the efficient score of the parameter of interest, β, for estimation. The efficient score of β is the orthogonal projection of the observed score $\dot{l}_\beta$ onto $\dot{\mathcal{P}}_\gamma^{\perp}$, where $\dot{\mathcal{P}}_\gamma^{\perp}$ is the orthocomplement of the linear span of scores of the nuisance parameters γ.

Score operators will be used to aid in calculating the efficient score. A score operator A can be used to map the complete scores to the observed scores ($\dot{\mathcal{Q}}$ to $\dot{\mathcal{P}}$). A mean zero square integrable space, or function, is denoted L_2^0. According to Bickel *et al.* (1993), the score operator $A : L_2^0(Q) \rightarrow L_2^0(P)$ is defined by $Aa(U) = E(a(U^0)|U) = Ra(U^0) + (1-R)E(a(U^0)|U_1^0)$ for $a \in L_2^0(Q)$. The adjoint of A, A^T: $L_2^0(P) \rightarrow L_2^0(Q)$, is defined by $A^T b(U^0) = E(b(U)|U^0)$ for $b \in L_2^0(P)$ (Bickel *et al.*, 1993). The basic idea is to rewrite Aa

as a linear combination of $f(A^TAa)$, while still being in the correct space and then restricting the space to obtain the efficient score. The efficient score l^*_β in model $\mathcal{P}$ is

$$\begin{aligned} l^*_\beta &= \frac{R}{\pi}\zeta^* - \frac{R-\pi}{\pi}E(\zeta^*|U^0_1) \in \mathcal{K} \subset \dot{\mathcal{P}}^\perp_\gamma \\ &= \prod\left(\dot{l}_\beta|\dot{\mathcal{P}}^\perp_\gamma\right) = \prod\left(\dot{l}_\beta|\dot{\mathcal{P}}\cap\dot{\mathcal{P}}^\perp_\gamma\right) = \prod\left(\dot{l}_\beta|\mathcal{M}\cap\dot{\mathcal{P}}^\perp_\gamma\right) \end{aligned}$$

since $\mathcal{M}$ is a closed subspace where $\dot{\mathcal{P}} \subset \mathcal{M} \subset L^0_2(P)$. We define $\mathcal{K} = \mathcal{M}\cap\dot{\mathcal{P}}^\perp_\gamma$ which is comprised of the closed subspace of all functions $k(U)$ defined as:

$$k(U) = \frac{R}{\pi}\zeta(U^0) - \frac{R-\pi}{\pi}E(\zeta(U^0)|U^0_1),$$

where $\zeta(U^0) \in \dot{\mathcal{Q}}^\perp_\gamma$. ζ^* is solved from

$$\Pi\left(\frac{1}{\pi}\zeta^* - \frac{1-\pi}{\pi}E(\zeta^*|U^0_1)|\dot{\mathcal{Q}}^\perp_\gamma\right) = l^{*0}_\beta,$$

where l^{*0}_β is the efficient score in model $\mathcal{Q}$.

The information bound of β, I^{*-1}_β,can be estimated with either the observed information $I^*_\beta = l^*_\beta l^{*T}_\beta$ or the expected information $I^*_\beta = E_P(l^*_\beta l^{*T}_\beta)$. If there is uncertainty in the model specification it is preferable to use the observed information, since it is robust under model misspecification. Upon calculating the efficient score and efficient information bound, an estimate of β can be found by one iteration of the Newton-Raphson estimator, also known as the one step estimator, yielding:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \widetilde{\boldsymbol{\beta}} + I^{*-1}_{\widetilde{\boldsymbol{\beta}}} l^{*s}_{\boldsymbol{\beta}} \\ l^{*s}_\beta &= \sum l^*_\beta, \end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}$ is an initial consistent starting value of $\boldsymbol{\beta}$. Nan *et al.* (2002) showed that $\widehat{\beta}$ is consistent, that $\left(\widehat{\beta}-\beta\right) \xrightarrow{d} N(0, I^{*-1}_{\widetilde{\beta}})$, and that $\widehat{\beta}$ is asymptotically semiparametrically

efficient. Nan demonstrated his method under a simulated two-stage design with lifetime data and discrete covariates.

2.4 LOGISTIC REGRESSION APPLICATION

2.4.1 Framework

The following notation is derived from Nan (2002), Lawless (1999), and Breslow (2003). Nan and Breslow developed the following notation specifically for a two-phase problem. Our contribution is made by borrowing and combining their notation; and then applying it to Nan's efficient score, the pseudo-likelihood, and the weighted pseudo-likelihood method to solve for missing by design data. Nan developed a general methodology for regression problems with missing data, but had not specifically filled in the details for a logistic regression problem. We will be the first to evaluate the properties of Nan's efficient score for logistic regression with incomplete covariate information.

The complete data are denoted by $(Y, \mathbf{Z})$, where $Y \in \{0, 1\}$ is the outcome and $\mathbf{Z} = (X, \mathbf{V})$ is a vector of covariates. The covariates consist of $\mathbf{V} = (V_1, V_2)$ where V_1 is a surrogate of X. V_1 is defined as a pure surrogate with implications that V_1 would not be included in the conditional model of Y given $\mathbf{Z}$ (Robins, 1994) since $f(Y|X, V) = f(Y|X, V_2)$. The completely observed data are $(Y, \mathbf{V})$ and X is MAR. The observed data are denoted by $W = (Y, \mathbf{Z}, R)^R(S, R)^{1-R}$, where $S = S(\mathbf{Z}, Y)$ is a function of $\mathbf{Z}$ and Y (Nan, 2002). The

missing data indicator is defined by

$$R = \begin{cases} 1 \text{ if } X \text{ is observed} \\ \\ 0 \text{ if } X \text{ is missing} \end{cases} . \tag{2.5}$$

We define S as a strata variable, where $S_j,\ j = 1, ..., J$, is specified by a combination of levels of Y and $\mathbf{V}$ where J is denoted $J = \max(Y) * \max(\mathbf{V})$ (Lawless *et al.*, 1999). If $\mathbf{V}$ is continuous, it would be necessary to categorize $\mathbf{V}$ using prespecified cutoff values to calculate S. Typically stratification variables are used for analysis of stratified designs. Post-stratification can be imposed for other types of missingness such as missing by happenstance.

Assume that the missing data mechanism only depends on the stratum defined by Y and V, where

$$\mathcal{S}_j = \mathcal{S}_{a,m} = \{(Y, \mathbf{Z}) : (Y, \mathbf{V}) \in (Y = a, \mathbf{V} = m), a = 0, 1, m = 1, .., M\}. \tag{2.6}$$

A stratum indicator is defined as $\delta_{ij} = I\{(y_i, z_i) \in \mathcal{S}_j\}\ i = 1, ..., N,\ j = 1, ..., J$ (Breslow, 2003). A stratum level variable assigns the corresponding stratum level and is defined by

$$S_i = \sum_{j=1}^{J} jI(\delta_{ij} = 1), i = 1, ..., N, \text{ or } S_i = j\ if\ I(\delta_{ij} = 1) \text{ (Lawless, 1999).} \tag{2.7}$$

The probability of being observed is modeled by

$$\pi(S) = P(R = 1|S) = \sum_{j=1}^{J} p_j \delta_{ij} = p_{s_i} \text{ (Lawless, 1999).} \tag{2.8}$$

Depending on the sampling scheme it is necessary to either use prespecified probabilities or estimate p_j, where $\widetilde{p}_j = \frac{n_j}{N_j}$, $n_j = \sum_{i=1}^{N} I(\delta_{ij} = 1, R_i = 1)$ and $N_j = \sum_{i=1}^{N} I(\delta_{ij} = 1)$ or $p_j = P(R = 1|y, v)$ where $(Y, \mathbf{Z}) \in \mathcal{S}_j$. In the case of missing by happenstance, one could use a logit model to estimate the probability of missingness.

The model for the complete data is

$$q_{\beta,g}(y,\mathbf{z}) = f(y|\mathbf{z})g(\mathbf{z}) = \left(\frac{\exp^{\boldsymbol{\beta}'\mathbf{z}}}{1+\exp^{\boldsymbol{\beta}'\mathbf{z}}}\right)^{y}\left(\frac{1}{1+\exp^{\boldsymbol{\beta}'\mathbf{z}}}\right)^{1-y} g(\mathbf{z}) \text{ (Breslow, 2003)}, \qquad (2.9)$$

where g is some density of $\mathbf{z}$. The distribution of being in the jth strata, $\mathcal{S}_j$, is $Q_j(\beta, G) = P((Y,\mathbf{Z}) \in \mathcal{S}_j)$ where $j = 1, ..., J$. The conditional distribution of being in the jth strata, $\mathcal{S}_j$, given the covariates is

$$\begin{aligned} Q_j^*(\mathbf{z},\beta) &= P((Y,\mathbf{z}) \in \mathcal{S}_j|\mathbf{Z}=\mathbf{z}) I_{\mathcal{S}_j^*}(\mathbf{z}) \text{ (Breslow, 2003)} \\ &= \sum_{Y:(Y,\mathbf{z})\in\mathcal{S}_j} f(y|\mathbf{z}) I_{\mathcal{S}_j^*}(\mathbf{z}) = f(y|\mathbf{z}) I_{\mathcal{S}_j^*}(\mathbf{z}) \\ &= \left(\frac{\exp^{\boldsymbol{\beta}'\mathbf{z}}}{1+\exp^{\boldsymbol{\beta}'\mathbf{z}}}\right)^{y}\left(\frac{1}{1+\exp^{\boldsymbol{\beta}'\mathbf{z}}}\right)^{1-y} I_{\mathcal{S}_j^*}(\mathbf{z}), \end{aligned}$$

where $j = 1, ..., J$ and $\mathcal{S}_j^* = \{\mathbf{z} \in \mathbf{Z} : \text{ for some } y, \ (y,\mathbf{z}) \in S_j\}$. The indicator function for $\mathcal{S}_j^*$ (Breslow, 2003) is

$$I_{\mathcal{S}_j^*}(\mathbf{z}) = \begin{cases} 1 \ \ if \ \ \mathbf{Z} = \mathbf{z} \text{ and } (y,\mathbf{z}) \in S_j \\ \\ 0 \ \ if \ \ \mathbf{Z} \neq \mathbf{z} \text{ or } (y,\mathbf{z}) \notin S_j \end{cases} . \qquad (2.10)$$

The distribution of $\mathcal{S}_j$ is the summation of the conditional distribution of $\mathcal{S}_j$ given the covariates over all values of $\mathbf{z}$,

$$\begin{aligned} Q_j(\beta, G) &= \sum_{\mathbf{z}} Q_j^*(\mathbf{z},\beta) g(\mathbf{z}) \text{ (Breslow, 2003)} \\ &= \sum_{\mathbf{z}} f(y|\mathbf{z}) I_{\mathcal{S}_j^*}(\mathbf{z}) g(\mathbf{z}) \\ &= \sum_{\mathbf{z}} \left(\frac{\exp^{\boldsymbol{\beta}'\mathbf{z}}}{1+\exp^{\boldsymbol{\beta}'\mathbf{z}}}\right)^{y}\left(\frac{1}{1+\exp^{\boldsymbol{\beta}'\mathbf{z}}}\right)^{1-y} I_{\mathcal{S}_j^*}(\mathbf{z}) g(\mathbf{z}). \end{aligned}$$

Since $\mathbf{z}$ contains missing observations, estimation of the distribution of $\mathbf{z}$ must take this into account to avoid an invalid distribution. Since $g(v)$ will fall out of the score equation,

we only need to estimate $g(x|v)$ by implementing the Horvitz-Thompson estimator. The empirical distribution of $x|v$ is estimated by

$$G(X|V = v_j) = \frac{1}{n_j^*} \sum_{i \in F_j^*} \frac{R_i}{\pi(y_i, v_j)} I(X_i \leq x) \text{ (Nan, 2002)}, \tag{2.11}$$

where $n_j^* = \sum_{i \in F_j} \frac{1}{\pi(y_i, v_j)}$, $F_j^* = \{i : V_i = v_j, i = 1, ..., n\}$, and $F_j = \{i : R_i = 1, V_i = v_j, i = 1, ..., n\}$.

2.4.2 Conditional Maximum Likelihood

For a stratified case-control study the default analysis is to include the stratum-specific terms. An alternative approach known as conditional maximum likelihood, also known as pseudo-likelihood, was developed by Fears and Brown (1986) based on the sampling probabilities. Fears *et al.* showed that the likelihood is properly modelled when including a ratio of the sampling probabilities of the cases to the control within each strata. In fact, this ratio is a constant within each strata and only affects the baseline coefficient. This led to the discovery of including the logarithmic transformation of this ratio as an offset term in the GLM setting. Intuitively, this makes sense since this reduces to the logit model being weighted by the ratio of selection probabilities of cases by controls within each strata. Only complete cases are included in the analysis and each case is assigned an offset term. Using these weights this method compensates for those subjects not included in the analysis.

Breslow and Cain (1988) and Wild (1991) improved and further developed this method for other designs. A requirement is that the probability of selection for cases and controls within each strata be known. Wild (1991) established that the estimate of β is consistent and asymptotically normal using Hsieh's (1985) method. The conditional maximum likelihood

can be used for a missing data problem, but with caution since it would be necessary for the missing data mechanism to depend on the outcome and other covariates. An advantage to this method is that standard software can be used.

The pseudo-likelihood is defined as (Breslow uses subscripts $\{i, j, k\}$ and we use subscripts $\{g, k, l\}$):

$$\begin{aligned} L_1 L_2 &= \prod_{g,k} P_{gk}^{N_{gk}} \prod_{g,k,l} p_{gkl} \text{ (Breslow 1988,1999),} \\ g &= \{0, 1\} \text{ for controls and cases} \\ k &= \{1, ..., K\} \text{ for the level of V} \\ l &= \{1, ..., n_{gk}\} \end{aligned}$$

where

$$P_{g,k} = \frac{\exp(g\delta_k)}{1 + \exp(g\delta_k)} = \Pr(Y = g | S_V = k), \ S_V \text{ is the strata variable for } V$$

and

$$p_{gkl} = \frac{n_{gk} \exp\left\{g(\beta_0 - \delta_k + x_{gkl}^T \beta)\right\}}{n_{0k} + n_{1k} \exp(\beta_0 - \delta_k + x_{gkl}^T \beta)}.$$

A pseudo-likelihood estimate is found by first maximizing $\prod_{g,k} P_{gk}^{N_{gk}}$ to obtain an estimate of δ_k, $\widehat{\delta}_k = \log(N_{1k}/N_{0k})$. Then $\prod_{g,k,l} p_{gkl}$ is maximized with the estimate $\widehat{\delta}_k$. In practice this pseudo-likelihood estimate is found by performing logistic regression with the phase two data including an offset $\log(n_{1k} N_{0k} / n_{0k} N_{1k})$. The covariance matrix must be corrected with the following formula

$$\left(X^T A X\right)^{-1} \left\{X^T A X - C^*\right\} \left(X^T A X\right)^{-1} \text{ (Breslow 1988, 1997),}$$

where A is a diagonal matrix with elements $n_{gkl} p_{0kl} p_{1kl}$ along the diagonal, $C^* = \sum_{g,k} \left(n_{gk}^{-1} - N_{gk}^{-1}\right) W_g W_g^T$, and $W_g = \sum_l n_{+kl} p_{0kl} p_{1kl} x_{kl}$.

2.4.3 Weighted Pseudo-likelihood

In 1974 Kish and Frankel suggested a weighted pseudo-likelihood approach. This approach borrows estimating procedures from the survey sampling literature applying the Horvitz-Thompson estimator. Only complete cases are included in analysis. Each case is assigned an inverted selection probability weight. In effect, the individual score equation is multiplied by this weight to compensate for the individuals excluded to implicitly draw information from the incomplete cases.

The form of the weighted log pseudo-likelihood is specified as:

$$l_{\beta}^{w} = \sum_{j=1}^{J} p_j^{-1} \sum_{i:(y_i,z_i)\in S_j} \log f(y_i|\mathbf{z}_i;\boldsymbol{\beta}) \text{ (Lawless 1999)} \tag{2.12}$$

with score function

$$S_w(\beta) = \sum_{i=1}^{N} R_i \mathbf{U}_{wi}(\mathbf{y}_i, \mathbf{z}_i, \widetilde{\mathbf{p}}; \boldsymbol{\beta}) \text{ (Lawless 1999)},$$

where $\mathbf{U}_{wi}(\mathbf{y}_i, \mathbf{z}_i, \widetilde{\mathbf{p}}; \boldsymbol{\beta}) = \sum_{j=1}^{J} p_j^{-1} \delta_{ij} \frac{\partial \log f(\mathbf{y}_i|\mathbf{z}_i;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. The covariance must be corrected by using the equation

$$Var_w(\boldsymbol{\beta}) = \left(\frac{\partial S_w}{\partial \beta^T}\right)^{-1} \widehat{\mathbf{B}}_w(\boldsymbol{\beta}) \left(\frac{\partial S_w}{\partial \beta}\right)^{-1} \text{ (Breslow 1999)},$$

$$\text{where } \widehat{\mathbf{B}}_w(\boldsymbol{\beta}) = \sum_{j=1}^{J} p_j^{-2} \left\{ \sum_{i:(y_i,z_i)\in S_j} \widetilde{\mathbf{U}}_{wi} \widetilde{\mathbf{U}}_{wi}^T - \frac{1-p_j}{p_j N_j} \left(\sum_{i:(y_i,z_i)\in S_j} \widetilde{\mathbf{U}}_{wi} \right) \left(\sum_{i:(y_i,z_i)\in S_j} \widetilde{\mathbf{U}}_{wi} \right)^T \right\}.$$

Wild (1991) proved that the estimate of β is consistent and asymptotically normal using Hsieh's (1985) method. The weighted approach has been found to perform reasonably well in terms of consistency, efficiency, and bias under MAR provided the missing data mechanism is properly defined. However, the large sample properties have not been thoroughly evaluated. This approach can easily be used for design and missing data problems. An advantage of implementing this method is that standard software can be used.

2.4.4 Efficient Score and Information Bound

An extension of the weighted pseudo-likelihood, known as IPWE estimating equations, was proposed by Nan (2002), Robins (1994), and Breslow (2003). Although, the weighted pseudo-likelihood method is generally found to be consistent and unbiased it is not always the most efficient. IPWE estimating equations attempt to gain more information from incomplete cases while placing one in the appropriate solution space. However, these methods are more complex than a simple weighting procedure.

The model for the observed data is

$$
\begin{aligned}
p_{B,g}(u) &= (\pi(s)q_{\beta,g}(y,\mathbf{z}))^r \left((1-\pi(s))\int\int_{(y,\mathbf{z}):S(\mathbf{z},y)=s} q_{\beta,g}(y,\mathbf{z})d\nu(\mathbf{z})d\mu(y)\right)^{1-r} \qquad (2.13)\\
&= \left[\left\{\left(\frac{\exp^{\mathbf{z}'\boldsymbol{\beta}}}{1+\exp^{\mathbf{z}'\boldsymbol{\beta}}}\right)^y \left(\frac{1}{1+\exp^{\mathbf{z}'\boldsymbol{\beta}}}\right)^{1-y} g(\mathbf{z})\sum_{j=1}^{J} p_j\delta_j\right\}^r * \right.\\
&\qquad \left.\left(\left(\sum_{j=1}^{J}(1-p_j)\delta_j\right) * \prod_{j=1}^{J}\left(\sum_{Z} I_{\mathcal{S}_j^*}(\mathbf{z}) \sum_{Y:(Y,z)\in\mathcal{S}_j} f(y|\mathbf{z})g(z)\right)^{\delta_j}\right)^{1-r}\right]\\
&= \left\{\left(\frac{\exp^{\mathbf{z}'\boldsymbol{\beta}}}{1+\exp^{\mathbf{z}'\boldsymbol{\beta}}}\right)^y \left(\frac{1}{1+\exp^{\mathbf{z}'\boldsymbol{\beta}}}\right)^{1-y} g(\mathbf{z})\sum_{j=1}^{J} p_j\delta_j\right\}^r \left(\left(\sum_{j=1}^{J}(1-p_j)\delta_j\right)\prod_{j=1}^{J} Q_j^{\delta_j}\right)^{1-r},
\end{aligned}
$$

where $\delta_j = 1$ if $(y,\mathbf{z}) \in S_j$. All equations in the observed density (2.13) can be found in Section 2.4.1. Specifically, $\pi(s)$ (2.8) is the probability of missingness; $q_{\beta,g}(y,\mathbf{z})$ (2.9) is the model for complete data; $\mathcal{S}$ (2.6) is a stratum variable; $I_{\mathcal{S}_j^*}(z)$ (2.10) is a strata covariate specific indicator; S (2.7) is a strata level variable; p is the probability of missingness for the specified strata; R (2.5) is the missing data indicator; and $g(z)$ (2.11) is the covariate distribution. The score function in model Q for β is $\dot{l}^0_\beta = \frac{\partial l}{\partial \beta}$. The score function in model P for β is $\dot{l}_\beta = R\overset{.0}{l}_\beta - (1-R)E(\overset{.0}{l}_\beta|S)$.

The efficient score, l^*_β, of β is the orthogonal projection of the score function of β for the observed model onto the orthocomplement of the nuisance parameter. The efficient score is (Nan *et al.* 2002):

$$l^*_\beta = \prod(\dot{l}_\beta|\mathcal{K}) = \frac{R}{\pi}c^*(\mathbf{Z},Y) - \frac{R-\pi}{\pi}E\left(c^*(\mathbf{Z},Y)|S\right),$$

where c^* is

$$\begin{aligned} c^* &= \pi(S)\left(\overset{.0}{l}_\beta - E\left(\overset{.0}{l}_\beta|\mathbf{Z},R=1\right)\right) + (1-\pi(S))\,E\left(c^*|S\right) - \\ &\quad \pi(S)E\left(\frac{1-\pi(S)}{\pi(S)}E\left(c^*(\mathbf{Z},Y)|S\right)|\mathbf{Z},R=1\right). \end{aligned}$$

One can solve $E\left(c^*|S\right) = \xi(S)$ by

$$\xi(S) = E\left(\overset{.0}{l}_\beta - E\left(\overset{.0}{l}_\beta|\mathbf{Z},R=1\right)|S\right) - E\left(E\left[\left(\frac{1-\pi(S)}{\pi(S)}\right)\xi(S)|\mathbf{Z},R=1\right]|S\right),$$

where $S = S(Z,Y)$ is a function of the fully observed data.

The information bound, I^{*-1}_β, can be estimated with either the observed information $I^*_\beta = l^*_\beta l^{*T}_\beta$ or expected information $I^*_\beta = E_P(l^*_\beta l^{*T}_\beta)$. Using one iteration of the Newton-Raphson estimator the solution for the coefficients is $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} + I^{*-1}_{\boldsymbol{\beta}} l^{*s}_{\boldsymbol{\beta}}(c(\mathbf{Z},Y))$, where $l^{*s}_{\boldsymbol{\beta}}(c(\mathbf{Z},Y)) = \sum_{i=1}^{n} l^*_{\boldsymbol{\beta}}(c(\mathbf{Z},Y))$.

2.5 SUMMARY

If one of the covariates in a regression application is missing at random, the regression coefficients must be adjusted to obtain consistent and efficient estimates. Pseudo-likelihood methods and the efficient score method for logistic regression were reviewed in detail. Asymptotic properties of these methods for logistic regression will be evaluated and compared in the next chapter.

CHAPTER 3

SIMULATION STUDY FOR COVARIATES MISSING BY DESIGN

3.1 INTRODUCTION

The goal of this chapter is to evaluate properties of estimators for missing data problems. It is of interest to determine the asymptotic properties of these estimators. Bias and precision of the estimators are evaluated by calculating the mean of the coefficient, mean of the variance, mean squared error (MSE), and asymptotic relative efficiency. The MSE is a measure that combines variance and bias, and thus, will aid in the selection of an estimator. It is preferable to have an estimator with minimal MSE. Validity of these methods under various distributional assumptions is also studied.

Summary statistics facilitate comparison of six methods. The summary statistics of the estimators of β calculated were: the mean of the coefficient of the replicates, mean of the variance of the coefficient of the replicates, MSE, ARE(full cohort, estimates) and 95% coverage. The ARE is the ratio of the empirical variance of the estimate from the full cohort by the empirical variance of the estimate from the corresponding method. The sample size and subset size varied. One thousand simulations were performed for logistic regression, where the model is $logit(Pr(Y = 1|x)) = \beta_0 + \beta_x$, $\{X, Y, V\}$ are binary, and V is a pure surrogate of X. On average 10% of the population are cases and 90% are controls. A case-cohort and stratified cohort study are generated where 200 subjects are selected with 100 cases and 100 controls. A sample of each strata j is randomly selected according to fixed probabilities. The strata j are defined by levels of the combinations of y and v where $\{j = 1\} = \{y = 0, v = 0\}$, $\{j = 2\} = \{y = 0, v = 1\}$, $\{j = 3\} = \{y = 1, v = 0\}$, and $\{j = 4\} = \{y = 1, v = 1\}$. Simulations were run for missing by design under 24 scenarios (Tables 3.1-3.24): $\beta_x = \{0, 1, 2\}$, $P(X = 1) = \{0.3, 0.5\}$, $P(V = k|X = k) = \{0.8, 0.5\}$, $k = \{0, 1\}$, $N = \{5000, 1000\}$, and $n = 200$. When $N = 5000$, the sample selected for complete data to be reported is 200 which is equivalent to 4% of the cohort. In this case, 100 of the 500 (20%) cases are selected and 100 of the 4500 (2%) controls are selected with additional stratification on V. When $N = 1000$, the sample selected for complete data to be reported is 200 which is equivalent to 10% of the cohort. In this case, 100 of the 100 (100%) cases are selected and 100 of the 900 (11%) controls are selected with additional stratification on V. Probability of selection for each strata are reported in the tables.

The first estimator (LR1) is for the full cohort, which will be used as a comparison measure to the other estimators. The full cohort includes all data for all subjects; that

is (Y, X, V) will be completely observed for all subjects. The remaining estimators will treat X as incomplete and (Y, V) as fully observed. The complete case (LR2) includes only those subjects with complete data. The complete case is known as the naive estimator, since standard logistic regression is performed with no further modification. Conditional maximum likelihood (PL) includes only those subjects with complete data but introduces an offset term in the logistic model. This offset term is a log function of the ratio of the probability of case selection within strata k by the probability of control selection within strata k, $\log\left(\frac{\pi_{1k}}{\pi_{ok}}\right)$. Weighted logistic regression (WPL) includes subjects with complete data and standard logistic regression is performed introducing a weight that is an inverse probability of selection within the strata. An initial value for the efficient score method is obtained from weighted logistic regression. The efficient score is modified as described in Chapter 2. The efficient score with the observed information is denoted as ESO and the efficient score with the expected information as ESE. If the model is incorrect, then the observed information is robust.

Although we developed code for the pseudo-likelihood and weighted pseudo-likelihood methods, we discovered after the fact that Breslow developed code for these methods. Since his software was written in a much more efficient fashion we used his code for this simulation study which is available at http://faculty.washington.edu/norm/software.html.

3.2 RESULTS

The efficient score performs as well or better than the other pseudo-likelihood methods. As anticipated, the complete case approach overestimates the variance and is biased towards

zero in all cases, where the bias is much worse for the intercept. Since the complete case performs so poorly, we will focus on comparing the efficient score to the two pseudo-likelihood methods PL and WPL. When data is completely reported for a small percentage of the populations, the efficient score (ESO and ESE) outperforms PL and WPL. In addition, ESO and ESE performed comparably.

Results for $\beta_x = 0$ can be found in Tables 3.1-3.8. We will first review results for $P(X = 1) = \{.5, .3\}$ and $P(V = k|X = k) = .8$. When $N = 5000$ (Tables 3.1, 3.3), all 4 methods produce unbiased estimates with the efficient score estimates more biased. The efficient score produces smaller variances, smaller MSEs, and more efficient estimators than the PL and WPL methods. The variances and MSE for the efficient score method are reduced by more than one half. When the sample size is reduced to $N = 1000$ (Tables 3.2, 3.4), the results hold with the exception that the the variances for PL and WPL method have been greatly reduced since a higher proportion of the data is available. As the correlation between X and V decreases, $P(V = k|X = k) = .5$, the results are the same for all four methods when $P(X = 1) = \{.5, .3\}$ and $N = \{5000, 1000\}$ (Tables 3.5-3.8). All methods have unbiased estimates. When $N = 5000$ (Tables 3.5 and 3.7), the efficient score produces slightly less biased estimates when $P(X = 1) = .5$, slightly smaller variances and MSEs when $P(X = 1) = \{.5, .3\}$, and slightly more efficient estimates when $P(X = 1) = .5$.

Results for $\beta_x = 1$ can be found in Tables 3.9-3.16. Results for $P(X = 1) = \{.5, .3\}$, $P(V = k|X = k) = .8$, and $N = 5000$ (Tables 3.9 and 3.11) are the same as $\beta_x = 0$ except the variance and MSE for the intercept are reduced by less than one-half of the pseudo-likelihood methods. When $P(X = 1) = \{.5, .3\}$, $P(V = k|X = k) = .8$, and $N = 1000$ (Tables 3.10 and 3.12) the results are the same as $\beta_x = 0$ except the efficient score produces

slightly less biased estimates. As with $\beta_x = 0$, all methods produce similar results when the correlation between X and V is reduced ($P(V = k|X = k) = .5$), $P(X = 1) = \{.5, .3\}$, and $N = \{5000, 1000\}$ (Tables 3.13-3.16). When $N = 5000$ (Tables 3.13 and 3.15), the efficient score produces slightly less biased estimates and slightly smaller variances and MSEs.

Tables 3.17-3.24 contain results for $\beta_x = 2$. An overall assessment is that as the relationship between the missing covariate and the outcome increases the variance reduction is not as large. When $P(X = 1) = \{.5, .3\}$, $P(V = k|X = k) = .8$, and $N = 5000$ (Tables 3.17 and 3.19), the results are the same as $\beta_x = 0$ except that the variance and MSE for the intercept are reduced by less than one-quarter of the pseudo-likelihood methods and, for the coefficient of X, by less than one-half. The efficient score is still more efficient than the PL and WPL but the efficiency has been reduced from when $\beta_x = \{0, 1\}$. When $P(X = 1) = \{.5, .3\}$, $P(V = k|X = k) = .8$, and $N = 1000$ (Tables 3.18 and 3.20), the results are the same as $\beta_x = 0$ except that the efficient score produces slightly less biased estimates when $P(X = 1) = .5$. Results are similar to $\beta_x = 0$ when the correlation between X and V is reduced ($P(V = k|X = k) = .5$), $P(X = 1) = \{.5, .3\}$, and $N = \{5000, 1000\}$ (Tables 3.21-3.24). All methods produce similar results. When $N = 5000$ (Tables 3.21 and 3.23), the efficient score produces slightly less biased estimates and slightly smaller variances and MSEs.

We have also summarized the results according to sample size and correlation. When $N = 5000$ and the correlation is high between X and V (Tables 3.1, 3.3, 3.9, 3.11, 3.17, 3.19), all four approaches are unbiased but the efficient score approach produces smaller variances, deflating them by at least one-half. Also the MSE is the lowest for the efficient score. The pseudo-likelihood methods have slightly less bias than the efficient score. As

N decreases, $N = 1000$, and the correlation is high between X and V (Tables 3.2, 3.4, 3.10, 3.12, 3.18, 3.20) the variance is reduced for all methods since more subjects have complete data. However, the efficient score is still the most efficient method and reduces variance the most. Once again, the four approaches are unbiased and the efficient score has the smallest MSE and slightly less bias. When the correlation between X and V is low (Tables 3.5-3.8, 3.13-3.16, 3.21-3.24), all four approaches are comparable. However, when N is larger, the efficient score is still slightly more efficient and has less biased estimates. As before, there is no bias and the MSE is about the same across all methods. The various distributions of the covariate produced similar results in all cases. It does not appear to matter whether X is equally distributed or skewed. The 95% coverage probability was accurate in all scenarios. Overall, the efficient score is an improvement if the correlation is high between X and its surrogate.

Table 3.1: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.5, P(V=k—X=k)=.8, β_0=-2.197, β_x=0, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-2.197	-0.005	-2.199	-2.199	-2.203	-2.201
Var	0.0044	0.0409	0.0161	0.0161	0.0077	0.0077
MSE	0.0049	4.8484	0.0166	0.0166	0.0084	0.0083
ARE(full cohort,est)	1		0.292	0.292	0.580	0.586
95% Cov	0.932	0	0.949	0.945	0.944	0.944
β_x						
Coef	-0.00006	0.00434	0.00510	0.00469	0.00599	0.00608
Var	0.0089	0.0816	0.0550	0.0551	0.0214	0.0216
MSE	0.0096	0.0851	0.0571	0.0572	0.0240	0.0237
ARE(full cohort,est)	1		0.169	0.169	0.403	0.408
95% Cov	0.933	0.939	0.951	0.950	0.940	0.943

Note: Probabilities of being observed are $P(R = 1|j) = \{0.0222, 0.0222, 0.2, 0.2\}$

Table 3.2: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.5, P(V=k—X=k)=.8, β_0=-2.197, β_x=0, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.205	-0.002	-2.203	-2.203	-2.206	-2.205
Var	0.0225	0.0407	0.0284	0.0284	0.0272	0.0271
MSE	0.0236	4.8593	0.0289	0.0291	0.0275	0.0274
ARE(full cohort,est)	1		0.816	0.812	0.859	0.861
95% Cov	0.946	0	0.951	0.953	0.949	0.947
β_x						
Coef	0.003	0.006	0.004	0.004	0.004	0.004
Var	0.0449	0.0814	0.0681	0.0683	0.0633	0.0632
MSE	0.0454	0.0796	0.0663	0.0665	0.0598	0.0597
ARE(full cohort,est)	1		0.685	0.683	0.759	0.761
95% Cov	0.954	0.945	0.953	0.955	0.959	0.957

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1111, 0.1111, 1, 1\}$

Table 3.3: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.3, P(V=k—X=k)=.8, β_0=-2.197, β_x=0, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-2.198	-0.005	-2.198	-2.197	-2.198	-2.197
Var	0.0032	0.0320	0.0104	0.0082	0.0047	0.0047
MSE	0.0032	4.8353	0.0103	0.0082	0.0051	0.0051
ARE(full cohort,est)	1		0.313	0.393	0.631	0.633
95% Cov	0.944	0	0.961	0.954	0.935	0.937
β_x						
Coef	0.004	0.007	0.004	0.001	-0.006	-0.002
Var	0.0106	0.0883	0.0640	0.0654	0.0274	0.0276
MSE	0.0108	0.0903	0.0643	0.0664	0.0289	0.0285
ARE(full cohort,est)	1		0.169	0.163	0.375	0.380
95% Cov	0.941	0.944	0.945	0.943	0.944	0.948

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0179, 0.0292, 0.1613, 0.2632\}$$

Table 3.4: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.3, P(V=k—X=k)=.8, β_0=-2.197, β_x=0, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=1000	n=200				
β_0						
Coef	-2.202	0.094	-2.203	-2.203	-2.205	-2.205
Var	0.0160	0.0303	0.0191	0.0185	0.0180	0.0180
MSE	0.0174	5.2807	0.0204	0.0198	0.0194	0.0193
ARE(full cohort,est)	1		0.853	0.879	0.901	0.901
95% Cov	0.934	0	0.950	0.948	0.944	0.945
β_x						
Coef	-0.006	-0.283	0.006	0.007	0.004	0.006
Var	0.0537	0.0929	0.0805	0.0812	0.0757	0.0756
MSE	0.0553	0.1705	0.0810	0.0818	0.0745	0.0744
ARE(full cohort,est)	1		0.682	0.676	0.741	0.743
95% Cov	0.961	0.857	0.953	0.952	0.951	0.954

Note: Probabilities of being observed are $P(R = 1|j) = \{0.0896, 0.1462, 1, 1\}$

Table 3.5: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.5, P(V=k—X=k)=.5, β_0=-2.197, β_x=0, expected subsample of 200

	LR1 n=5000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.197	-0.005	-2.198	-2.197	-2.197	-2.197
Var	0.0044	0.0407	0.0227	0.0228	0.0218	0.0217
MSE	0.0049	4.8481	0.0240	0.0242	0.0240	0.0240
ARE(full cohort,est)	1		0.202	0.201	0.202	0.203
95% Cov	0.932	0	0.952	0.950	0.948	0.948
β_x						
Coef	-0.000060	0.001329	0.001195	0.000002	-0.001617	-0.001622
Var	0.0089	0.0816	0.0812	0.0817	0.0778	0.0774
MSE	0.0096	0.0858	0.0869	0.0874	0.0867	0.0865
ARE(full cohort,est)	1		0.111	0.110	0.111	0.111
95% Cov	0.933	0.949	0.946	0.951	0.946	0.945

Note: Probabilities of being observed are $P(R = 1|j) = \{0.0222, 0.0222, 0.2, 0.2\}$

Table 3.6: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.5, P(V=k—X=k)=.5, β_0=-2.197, β_x=0, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.205	-0.003	-2.203	-2.203	-2.203	-2.203
Var	0.0225	0.0407	0.0317	0.0317	0.0318	0.0316
MSE	0.0236	4.8546	0.0300	0.0301	0.0300	0.0300
ARE(full cohort,est)	1		0.787	0.784	0.786	0.786
95% Cov	0.946	0	0.969	0.966	0.967	0.967
β_x						
Coef	0.003	0.007	0.007	0.007	0.006	0.006
Var	0.0449	0.0814	0.0810	0.0814	0.0816	0.0811
MSE	0.0454	0.0733	0.0736	0.0741	0.0738	0.0738
ARE(full cohort,est)	1		0.617	0.613	0.615	0.615
95% Cov	0.954	0.967	0.967	0.966	0.964	0.963

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1111, 0.1111, 1, 1\}$

Table 3.7: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.3, P(V=k—X=k)=.5, β_0=-2.197, β_x=0, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-2.198	-0.006	-2.199	-2.199	-2.199	-2.199
Var	0.0032	0.0290	0.0110	0.0110	0.0106	0.0105
MSE	0.0032	4.8323	0.0114	0.0115	0.0113	0.0113
ARE(full cohort,est)	1		0.282	0.281	0.284	0.285
95% Cov	0.944	0	0.951	0.950	0.944	0.944
β_x						
Coef	0.004	0.008	0.008	0.007	0.008	0.008
Var	0.0106	0.0981	0.0976	0.0982	0.0935	0.0930
MSE	0.0108	0.1047	0.1055	0.1065	0.1040	0.1038
ARE(full cohort,est)	1		0.103	0.102	0.104	0.104
95% Cov	0.941	0.943	0.945	0.946	0.943	0.942

Note: Probabilities of being observed are $P(R = 1|j) = \{0.0222, 0.0222, 0.2, 0.2\}$

Table 3.8: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.3, P(V=k—X=k)=.5, β_0=-2.197, β_x=0, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.202	0.001	-2.200	-2.200	-2.200	-2.200
Var	0.0160	0.0289	0.0200	0.0200	0.0200	0.0199
MSE	0.0174	4.8603	0.0212	0.0214	0.0213	0.0213
ARE(full cohort,est)	1		0.819	0.815	0.817	0.817
95% Cov	0.934	0	0.952	0.951	0.951	0.951
β_x						
Coef	-0.0061	0.0009	0.0012	0.0011	0.0021	0.0022
Var	0.0537	0.0976	0.0971	0.0975	0.0978	0.0971
MSE	0.0553	0.0970	0.0975	0.0976	0.0971	0.0971
ARE(full cohort,est)	1		0.567	0.566	0.569	0.569
95% Cov	0.961	0.956	0.956	0.954	0.956	0.956

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1111, 0.1111, 1, 1\}$

Table 3.9: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.5, P(V=k—X=k)=.8, β_0=-2.787, β_x=1, expected subsample of 200

	LR1 n=5000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.790	-0.381	-2.791	-2.791	-2.803	-2.796
Var	0.0072	0.0488	0.0234	0.0260	0.0145	0.0147
MSE	0.0084	5.8420	0.0262	0.0284	0.0165	0.0161
ARE(full cohort,est)	1		0.320	0.295	0.516	0.525
95% Cov	0.933	0	0.946	0.941	0.938	0.941
β_x						
Coef	1.005	0.664	1.005	1.006	1.021	1.015
Var	0.0104	0.0855	0.0588	0.0610	0.0283	0.0285
MSE	0.0120	0.2095	0.0666	0.0679	0.0328	0.0320
ARE(full cohort,est)	1		0.180	0.177	0.371	0.378
95% Cov	0.936	0.771	0.933	0.936	0.943	0.947

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0216, 0.0229, 0.2663, 0.1583\}$$

Table 3.10: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.5, P(V=k—X=k)=.8, β_0=-2.787, β_x=1, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.798	-0.582	-2.799	-2.798	-2.799	-2.798
Var	0.0370	0.0551	0.0429	0.0428	0.0419	0.0418
MSE	0.0386	4.9196	0.0441	0.0443	0.0429	0.0428
ARE(full cohort,est)	1		0.875	0.872	0.901	0.902
95% Cov	0.943	0	0.949	0.951	0.947	0.942
β_x						
Coef	1.008	0.982	1.015	1.014	1.012	1.011
Var	0.0534	0.0902	0.0768	0.0771	0.0725	0.0724
MSE	0.0552	0.0905	0.0764	0.0769	0.0710	0.0708
ARE(full cohort,est)	1		0.724	0.719	0.778	0.780
95% Cov	0.947	0.955	0.955	0.953	0.949	0.951

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1081, 0.1145, 1, 1\}$

Table 3.11: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.3, P(V=k—X=k)=.8, β_0=-2.577, β_x=1, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-2.580	-0.288	-2.581	-2.580	-2.588	-2.584
Var	0.0043	0.0362	0.0134	0.0123	0.0074	0.0076
MSE	0.0048	5.2789	0.0144	0.0133	0.0089	0.0088
ARE(full cohort,est)	1		0.333	0.361	0.546	0.550
95% Cov	0.938	0	0.944	0.943	0.935	0.939
β_x						
Coef	1.006	0.674	1.011	1.010	1.019	1.017
Var	0.0090	0.0861	0.0594	0.0618	0.0280	0.0282
MSE	0.0099	0.1931	0.0608	0.0638	0.0316	0.0312
ARE(full cohort,est)	1		0.163	0.155	0.316	0.319
95% Cov	0.936	0.790	0.943	0.940	0.932	0.936

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0175, 0.0304, 0.2007, 0.1962\}$$

Table 3.12: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.3, P(V=k—X=k)=.8, β_0=-2.577, β_x=1, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.580	-0.279	-2.580	-2.580	-2.581	-2.581
Var	0.0217	0.0360	0.0248	0.0242	0.0238	0.0237
MSE	0.0231	5.3181	0.0259	0.0253	0.0247	0.0247
ARE(full cohort,est)	1		0.892	0.911	0.933	0.933
95% Cov	0.947	0	0.942	0.947	0.950	0.949
β_x						
Coef	0.999	0.684	1.006	1.008	1.003	1.005
Var	0.0455	0.0854	0.0725	0.0742	0.0679	0.0678
MSE	0.0480	0.1854	0.0745	0.0763	0.0691	0.0690
ARE(full cohort,est)	1		0.645	0.630	0.695	0.696
95% Cov	0.953	0.789	0.946	0.950	0.957	0.957

Note: Probabilities of being observed are $P(R = 1|j) = \{0.0877, 0.1520, 1, 1\}$

Table 3.13: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.5, P(V=k—X=k)=.5, β_0=-2.787, β_x=1, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-2.790	-0.620	-2.806	-2.806	-2.793	-2.793
Var	0.0072	0.0556	0.0376	0.0377	0.0368	0.0367
MSE	0.0084	4.7557	0.0414	0.0416	0.0413	0.0412
ARE(full cohort,est)	1		0.204	0.203	0.203	0.204
95% Cov	0.933	0	0.953	0.952	0.945	0.944
β_x						
Coef	1.005	1.030	1.030	1.028	1.007	1.007
Var	0.0104	0.0913	0.0909	0.0914	0.0887	0.0884
MSE	0.0120	0.0965	0.0973	0.0979	0.0971	0.0970
ARE(full cohort,est)	1		0.125	0.124	0.124	0.124
95% Cov	0.936	0.950	0.947	0.948	0.941	0.939

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0222, 0.0222, 0.1986, 0.1986\}$$

Table 3.14: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.5, P(V=k—X=k)=.5, β_0=-2.787, β_x=1, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.798	-0.597	-2.796	-2.796	-2.795	-2.795
Var	0.0370	0.0548	0.0458	0.0459	0.0461	0.0459
MSE	0.0386	4.8500	0.0437	0.0440	0.0437	0.0437
ARE(full cohort,est)	1		0.882	0.878	0.882	0.882
95% Cov	0.943	0	0.957	0.957	0.957	0.957
β_x						
Coef	1.008	1.012	1.012	1.012	1.010	1.010
Var	0.0534	0.0902	0.0899	0.0903	0.0906	0.0901
MSE	0.0552	0.0821	0.0824	0.0831	0.0826	0.0826
ARE(full cohort,est)	1		0.670	0.664	0.668	0.668
95% Cov	0.947	0.970	0.969	0.969	0.970	0.969

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1112, 0.1112, 1, 1\}$

Table 3.15: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.3, P(V=k—X=k)=.5, β_0=-2.577, β_x=1, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-2.580	-0.399	-2.585	-2.585	-2.578	-2.578
Var	0.0043	0.0349	0.0168	0.0169	0.0163	0.0164
MSE	0.0048	4.7802	0.0198	0.0200	0.0199	0.0199
ARE(full cohort,est)	1		0.242	0.240	0.240	0.240
95% Cov	0.938	0	0.933	0.927	0.925	0.924
β_x						
Coef	1.006	1.026	1.026	1.026	1.007	1.006
Var	0.0090	0.0925	0.0922	0.0928	0.0888	0.0891
MSE	0.0099	0.1050	0.1057	0.1074	0.1063	0.1063
ARE(full cohort,est)	1		0.094	0.092	0.093	0.093
95% Cov	0.936	0.940	0.935	0.939	0.930	0.931

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0222, 0.0222, 0.1984, 0.1984\}$$

Table 3.16: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.3, P(V=k—X=k)=.5, β_0=-2.577, β_x=1, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-2.580	-0.380	-2.579	-2.579	-2.579	-2.579
Var	0.0217	0.0345	0.0255	0.0255	0.0256	0.0255
MSE	0.0231	4.8623	0.0262	0.0264	0.0263	.02628
ARE(full cohort,est)	1		0.879	0.876	0.878	0.878
95% Cov	0.947	0	0.944	0.946	0.950	0.950
β_x						
Coef	0.999	1.012	1.012	1.012	1.009	1.010
Var	0.0455	0.0916	0.0912	0.0917	0.0919	0.0913
MSE	0.0480	0.0898	0.0899	0.0900	0.0893	0.0893
ARE(full cohort,est)	1		0.535	0.534	0.538	0.538
95% Cov	0.953	0.962	0.959	0.958	0.959	0.958

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1112, 0.1112, 1, 1\}$

Table 3.17: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.5, P(V=k—X=k)=.8, β_0=-3.557, β_x=2, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-3.561	-0.920	-3.573	-3.578	-3.576	-3.562
Var	0.0147	0.0685	0.0432	0.0516	0.0351	0.0363
MSE	0.0156	7.0274	0.0454	0.0548	0.0350	0.0341
ARE(full cohort,est)	1		0.346	0.287	0.451	0.459
95% Cov	0.946	0	0.947	0.950	0.957	0.949
β_x						
Coef	2.005	1.431	2.020	2.025	2.032	2.018
Var	0.0175	0.1029	0.0760	0.0848	0.0505	0.0517
MSE	0.0185	0.4320	0.0825	0.0916	0.0531	0.0510
ARE(full cohort,est)	1		0.225	0.203	0.354	0.364
95% Cov	0.943	0.548	0.948	0.945	0.948	0.950

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0212, 0.0223, 0.3509, 0.1381\}$$

Table 3.18: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.5, P(V=k—X=k)=.8, β_0=-3.557, β_x=2, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-3.586	-1.359	-3.586	-3.586	-3.576	-3.572
Var	0.0777	0.0961	0.0841	0.0843	0.0838	0.0850
MSE	0.0829	4.9306	0.0895	0.0894	0.0846	0.0842
ARE(full cohort,est)	1		0.926	0.927	0.975	0.978
95% Cov	0.953	0	0.946	0.941	0.950	0.948
β_x						
Coef	2.026	1.978	2.033	2.033	2.021	2.018
Var	0.0917	0.1293	0.1158	0.1166	0.1128	0.1137
MSE	0.0958	0.1337	0.1209	0.1213	0.1131	0.1124
ARE(full cohort,est)	1		0.794	0.792	0.844	0.849
95% Cov	0.954	0.951	0.952	0.954	0.955	0.960

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1060, 0.1169, 1, 1\}$

Table 3.19: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.3, P(V=k—X=k)=.8, β_0=-3.157, β_x=2, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-3.160	-0.662	-3.163	-3.164	-3.171	-3.165
Var	0.0073	0.0442	0.0209	0.0224	0.0157	0.0160
MSE	0.0079	6.2700	0.0227	0.0247	0.0186	0.0183
ARE(full cohort,est)	1		0.347	0.319	0.428	0.433
95% Cov	0.949	0	0.943	0.946	0.937	0.942
β_x						
Coef	2.005	1.349	2.014	2.017	2.026	2.019
Var	0.0110	0.0913	0.0621	0.0695	0.0393	0.0397
MSE	0.0118	0.5189	0.0651	0.0736	0.0453	0.0445
ARE(full cohort,est)	1		0.181	0.160	0.263	0.266
95% Cov	0.947	0.408	0.957	0.953	0.932	0.939

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0172, 0.0316, 0.2688, 0.1584\}$$

Table 3.20: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.3, P(V=k—X=k)=.8, β_0=-3.157, β_x=2, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-3.162	-1.220	-3.162	-3.162	-3.164	-3.162
Var	0.0371	0.0462	0.0395	0.0396	0.0396	0.0395
MSE	0.0384	3.8006	0.0407	0.0410	0.0403	0.0402
ARE(full cohort,est)	1		0.943	0.937	0.955	0.956
95% Cov	0.947	0	0.945	0.945	0.952	0.947
β_x						
Coef	2.001	2.379	2.014	2.013	2.015	2.013
Var	0.0556	0.1057	0.0938	0.0938	0.0895	0.0891
MSE	0.0554	0.2542	0.0939	0.0927	0.0876	0.0873
ARE(full cohort,est)	1		0.591	0.598	0.634	0.635
95% Cov	0.960	0.806	0.954	0.956	0.956	0.957

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1578, 0.0858, 1, 1\}$

Table 3.21: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.5, P(V=k—X=k)=.5, β_0=-3.557, β_x=2, expected subsample of 200

	LR1	LR2	PL	WPL	ESO	ESE
	n=5000	n=200				
β_0						
Coef	-3.561	-1.409	-3.594	-3.593	-3.578	-3.577
Var	0.0147	0.0973	0.0796	0.0801	0.0795	0.0794
MSE	0.0156	4.7131	0.0869	0.0875	0.0871	0.0870
ARE(full cohort,est)	1		0.183	0.181	0.180	0.180
95% Cov	0.946	0.001	0.954	0.946	0.945	0.944
β_x						
Coef	2.005	2.047	2.047	2.046	2.021	2.020
Var	0.0175	0.1314	0.1313	0.1320	0.1306	0.1302
MSE	0.0185	0.1426	0.1432	0.1443	0.1443	0.1441
ARE(full cohort,est)	1		0.131	0.130	0.128	0.128
95% Cov	0.943	0.946	0.945	0.944	0.936	0.936

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0222, 0.0222, 0.1982, 0.1982\}$$

Table 3.22: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.5, P(V=k—X=k)=.5, β_0=-3.557, β_x=2, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-3.586	-1.383	-3.583	-3.583	-3.581	-3.581
Var	0.0777	0.0955	0.0867	0.0872	0.0875	0.0873
MSE	0.0829	4.8212	0.0886	0.0885	0.0884	0.0885
ARE(full cohort,est)	1		0.934	0.934	0.934	0.934
95% Cov	0.953	0.001	0.954	0.952	0.954	0.954
β_x						
Coef	2.026	2.027	2.027	2.028	2.023	2.023
Var	0.0917	0.1293	0.1291	0.1299	0.1303	0.1298
MSE	0.0958	0.1246	0.1253	0.1255	0.1252	0.1252
ARE(full cohort,est)	1		0.764	0.763	0.763	0.763
95% Cov	0.954	0.966	0.964	0.963	0.965	0.964

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1112, 0.1112, 1, 1\}$

Table 3.23: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 5000, P(X=1)=.3, P(V=k—X=k)=.5, β_0=-3.157, β_x=2, expected subsample of 200

	LR1 n=5000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	-3.160	-0.987	-3.177	-3.177	-3.167	-3.167
Var	0.0073	0.0502	0.0321	0.0323	0.0316	0.0319
MSE	0.0079	4.7638	0.0363	0.0366	0.0365	0.0365
ARE(full cohort,est)	1		0.220	0.218	0.216	0.216
95% Cov	0.949	0	0.950	0.949	0.944	0.945
β_x						
Coef	2.005	2.044	2.044	2.044	2.023	2.022
Var	0.0110	0.1056	0.1050	0.1057	0.1034	0.1039
MSE	0.0118	0.1193	0.1198	0.1216	0.1212	0.1212
ARE(full cohort,est)	1		0.100	0.098	0.098	0.097
95% Cov	0.947	0.939	0.937	0.937	0.930	0.930

Note: Probabilities of being observed are

$$P(R = 1|j) = \{0.0222, 0.0222, 0.1994, 0.1994\}$$

Table 3.24: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 1000, P(X=1)=.3, P(V=k—X=k)=.5, β_0=-3.157, β_x=2, expected subsample of 200

	LR1 n=1000	LR2 n=200	PL	WPL	ESO	ESE
β_0						
Coef	3.162	-0.963	-3.162	-3.163	-3.162	-3.162
Var	0.0371	0.0494	0.0403	0.0404	0.0405	0.0404
MSE	0.0384	4.8627	0.0412	0.0413	0.0412	0.0412
ARE(full cohort,est)	1		0.933	0.931	0.931	0.932
95% Cov	0.947	0	0.953	0.951	0.954	0.954
β_x						
Coef	2.001	2.020	2.020	2.020	2.016	2.016
Var	0.0556	0.1044	0.1038	0.1044	0.1047	0.1042
MSE	0.0554	0.0977	0.0978	0.0983	0.0978	0.0978
ARE(full cohort,est)	1		0.568	0.566	0.567	0.567
95% Cov	0.960	0.960	0.960	0.959	0.960	0.961

Note: Probabilities of being observed are $P(R = 1|j) = \{0.1112, 0.1112, 1, 1\}$

CHAPTER 4

TRUNCATED DATA

4.1 INTRODUCTION

Medical studies are frequently interested in determining the relationship between biological markers and other variables. Many of these biological markers are measured with assays that have a lower threshold for detection of the substance. When this is the case, the measurement is recorded to be that of the lower threshold value and the reading is assumed to be "normal". This results in left truncated data.

In this setting, regression models are the analysis tool of choice with logistic, survival, and linear models seeing the greatest use. Each of these modeling techniques places different assumptions on the outcome variables, while requiring that the covariates be fixed. These assumptions can create problems when data are truncated since the models do not explicitly handle truncated covariate data. In addition, a limited number of modeling approaches are available for the modeling of truncated outcome data. The Tobit (Tobin 1958) model is the most popular approach for regression modelling with truncated data.

In many studies, the observed level of truncated data may be quite low since most of the observed values are above the threshold level, particularly for markers of "illness". In this instance, truncation may have little impact on data analysis results. A sepsis study conducted at the University of Pittsburgh has motivated us to determine the impact of a large amount of truncation. One aim of this study was to determine the relationship between severe sepsis status and measures of inflammation such as tumor necrosis factor, interleukin-10, and interleukin-6.

Thirty-eight hospitals participated in this study from November 2001 and November 2003, enrolling 2320 patients. Patient eligibility criteria included being at least 18 years old and having both a clinical diagnosis of pneumonia and a new pulmonary infiltrate on chest x-ray. During a patient's stay in the hospital, blood was drawn for cytokine assays at enrollment, and on days 2-8, 15, 22 and 30. Baseline inflammatory marker samples were collected for 1815 subjects, of which 1809 samples were collected for IL-10 and TNF and 1811 samples for IL-6. The detectable limit for IL-10 and IL-6 was 5 and for TNF was 4, indicating that the concentration of the sample for these markers was below the detectable limit. After a majority of the IL-6 samples were assayed, a more sensitive assay was developed for it and the new detection limit was 2. Of the 1809 patients, 714 (39%) were below the detectable limit for TNF and 900 (50%) were below the detectable limit for IL-10. Of the 1811 patients, 278 (15%) were below the detectable limit for IL-6 where 25 (1%) were below the detectable limit of 2 and 253 (14%) were below the detectable limit of 5. Of the 1815 patients, 477 (26%) developed severe sepsis during their stay in the hospital. Upon discharge, 78 of the 1815 patients (4%) died.

The Tobit model can handle truncated outcome data in a regression model. A generalization of the Tobit model known as the censored normal regression model can handle truncated outcomes when multiple thresholds exist. We demonstrate how inference and prediction are affected if appropriate measures are not taken to adjust an analysis for truncated data. We also demonstrate how the data analysis results are impacted when the data are analyzed using the Tobit model, standard regression with filling in the truncated values, and standard regression with complete cases. The Tobit model is used for TNF and IL-10 and the censored normal regression (CNREG) model for IL-6. Results are reported in Table 4.1. Inference is the same regardless of the method chosen. But the predicted value of the inflammatory markers conditional on severe sepsis status increases with both standard regression methods. This is occurring because the Tobit/CNREG model assumes that truncated values are below the detectable limit and that the slope is not as steep. The other two methods produce steeper slopes due to either assuming that all truncated values are equivalent to the detectable limit, or discarding the data. These results should convince one to use appropriate modelling techniques to handle truncated data. This implies that if covariate data is also truncated the inference and prediction could potentially be impacted.

Table 4.1: Linear Regression and Tobit Model Results for Cytokines

TNF	N	E(tnf|no sssoap)	E(tnf|sssoap)	p-value
Tobit Model	1809	4.8	6.5	<.0001
LR filling in truncated value	1809	6.6	8.2	<.0001
LR with nontruncated values	1095	9.4	11.9	<.0001
IL-10	N	E(IL-10|no sssoap)	E(IL-10|sssoap)	p-value
Tobit Model	1809	4.4	7.1	<.0001
LR filling in truncated value	1809	9.1	11.9	<.0001
LR with nontruncated values	909	17.8	21.9	<.0001
IL-6	N	E(IL-6|no sssoap)	E(IL-6|sssoap)	p-value
CNREG	1811	3.4	4.2	<.0001
LR filling in truncated value	1811	3.6	4.3	<.0001
LR with nontruncated values	1533	4.0	4.6	<.0001

Note: LR=Linear Regression, SSSOAP= Severe Sepsis SOAP

Truncated and censored data methodology have been developed for the last 30 years with the focus on the outcome variable. This has lead to the development of models such as the Tobit regression model for truncated data in addition to numerous models for censored data. Another commonly encountered problem is that of truncated covariate data. This type of data is generally observed in the laboratory setting, where the lower limit of detection of an assay is often observed. Currently two methods exist to handle a truncated covariate. The first method is a complete case method. Estimates from this approach will be consistent

but have inflated variances due to deletion of cases. The second approach includes all subjects filling in the truncated values with the lower threshold value.

To address this problem, we propose two methods to estimate the coefficients and their standard errors for a regression model with a left truncated covariate. The first method is a likelihood-based approach. The second approach uses estimating equation techniques. The likelihood-based method is solved and will be compared to a standard method of filling in the truncated values with the lower threshold value in the next chapter. The estimating equation method is close to completion and once solved should be the most efficient method.

The application of the likelihood-based method is illustrated in the sepsis study conducted at the University of Pittsburgh, referenced above. One aim of this study was to determine the relationship between severe sepsis status and measures of inflammation such as tumor necrosis factor, interleukin-10, and interleukin-6.

4.2 LITERATURE REVIEW

The goal of this chapter is to develop a regression model for a binary outcome adjusting for a truncated variable. Very little literature has been devoted to regression with truncated variables. Henery (1981) is the only known author to develop a normal conditional distribution for a random variable given a truncated variable. Tobin (1958) developed the Tobit model for regression modelling with a truncated outcome for both a truncated and censored sample. The Tobit model is the most popular approach for regression modelling with truncated data. Breen (1996) reviews truncated and censored samples as well as the Tobit model in greater detail.

Dempster *et al.* (1977) developed an EM algorithm approach for truncated variables in a general framework. McLaren *et al.* (1986, 1991) has extended the EM algorithm approach to handle truncated immunology data. McLaren has focused on descriptive and regression modelling tools where the truncated variable is the outcome. We base our first approach on a likelihood-based method covered in the next section.

Bickel *et al.* (1983) developed an efficient score approach for regression with a truncated sample where the outcome is truncated. Bickel chose an estimating equation method to adjust for this biased sampling. We base our second proposed approach on estimating equations for a censored sample with a truncated covariate. Our proposed extension will be described in Section 4.4.

4.3 LIKELIHOOD-BASED EXTENSION

Truncated data are defined as data that are observed within a fixed interval. A random variable that is observed above a threshold is known as left truncation; whereas, right truncation is defined as observing a random variable below a threshold. Double truncation is defined as observing a random variable between lower and upper thresholds. Left truncation is the focus of this chapter. A left truncated variable is denoted by

$$x_c = \begin{cases} x \;\; if \; x > c \\ \\ c \;\; if \; x \leq c \end{cases} .$$

Two types of samples involving truncated variables are censored samples and truncated samples (Breen, 1996). For definition purposes, let X be truncated and $Z = (Y, V)$ be fully observed where (X, V) are covariates and Y is the outcome. In the case of a censored

sample, Z is observed for all subjects and X is observed for some subjects. In the case of a truncated sample, Z is observed only for those subjects who have an observed value for X. Censored samples will be addressed in this chapter.

The likelihood-based approach will reduce bias and variance in the coefficients but will not produce an efficient estimate partially due to bias. The full likelihood must be specified including a distribution of the covariate space. The truncated indicator is defined as $R = \begin{cases} 1 \; if \; x > c \\ 0 \; if \; x \leq c \end{cases}$. To simplify matters, a specific example will be used where $W = (Y, V, X, R)$ and X is truncated. We assume that X and V are independent. The density of the data are

$$p(Y, V, X, R) = q(Y, V, X)^R \left(\int_{x \leq c} q(Y, V, X) d\mu(x) \right)^{1-R},$$

where $q(Y, V, X) = f(y|v, x)f(v, x) = f(y|v, x)f(v)f(x)$. The first part of the likelihood contributes to the nontruncated data and the second part to the portion that is truncated. The second component of the likelihood is an average of the density over all values $X \leq c$.

The log likelihood is

$$\begin{aligned} \log p(w) &= R \log q(Y, V, X) + (1 - R) \log \left(\int_{x \leq c} q(Y, V, X) d\mu(x) \right) \\ &= R \left\{ \log f(y|v, x) + \log f(x) + \log f(v) \right\} \\ &\quad + (1 - R) \log \left(\int_{x \leq c} f(y|v, x) f(x) f(v) d\mu(x) \right). \end{aligned}$$

The derivative of the log-likelihood wrt β is defined as:

$$\begin{aligned}
\frac{\partial}{\partial\beta}\log p(w) &= \frac{\partial}{\partial\beta}R\log q(Y,V,X) + \frac{\partial}{\partial\beta}(1-R)\log\left(\int_{x\leq c} q(Y,V,X)d\mu(x)\right) \\
&= R\dot{l}(y,x,v|\beta) \\
&\quad + (1-R)\,\frac{\int_{x\leq c}\dot{l}(y,x,v|\beta)f(y|v,x)f(x)f(v)d\mu(x)}{\int_{x\leq c} f(y|v,x)f(x)f(v)d\mu(x)} \\
&= R\dot{l}(y,x,v|\beta) + (1-R)\,E(\dot{l}(y,x,v|\beta)|x\leq c,y,v). \qquad (4.1)
\end{aligned}$$

In our specific case, we have a binary outcome and the covariate data is inflammatory marker data which implies that we are concerned with only the logistic model and we can assume the covariate $log(X)$ is normally distributed. The score of the logistic model is $\dot{l}_1(y,z,v|\beta) = \begin{pmatrix} 1 \\ x \\ v \end{pmatrix}\left(y - \frac{\exp(\beta_0+\beta x+\beta_v v)}{1+\exp(\beta_0+\beta x+\beta_v v)}\right)$. In other scenarios, one may not be able to assume a distribution for the covariate space which is another reason for choosing a semi-parametric approach over the likelihood-based approach.

Solving for the parameters from 4.1 directly is intractable so an alternative method must be used. The Newton-Raphson estimator is used to solve for the coefficient parameters β. The following steps are taken:

- Step 1: Obtain a consistent estimate, $\widetilde{\beta}$, of β to use as an initial starting value. A consistent estimate can be obtained from the analysis using only nontruncated values.

- Step 2: Differentiate the log-likelihood wrt β which reduces to Equation 4.1 denoted $\overset{.o}{l}$.

- Step 3: Use the Newton-Raphson estimator to solve for an estimate of β, $\widehat{\beta}$, where

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_X \\ \widehat{\beta}_V \end{pmatrix} = \begin{pmatrix} \widetilde{\beta}_0 \\ \widetilde{\beta}_X \\ \widetilde{\beta}_V \end{pmatrix} + I_{\widetilde{\beta}}^{o-1} \begin{pmatrix} \sum_{i=1}^{n} \overset{.o}{l}_{\widetilde{\beta}_0,i} \\ \sum_{i=1}^{n} \overset{.o}{l}_{\widetilde{\beta}_X,i} \\ \sum_{i=1}^{n} \overset{.o}{l}_{\widetilde{\beta}_V,i} \end{pmatrix}$$ and the information bound $I_{\widetilde{\beta}}^{o-1}$ is

$$I_{\widetilde{\beta}}^{o} = \left(\begin{pmatrix} \overset{.o}{l}_{\widetilde{\beta}_0} \\ \overset{.o}{l}_{\widetilde{\beta}_X} \\ \overset{.o}{l}_{\widetilde{\beta}_V} \end{pmatrix} \begin{pmatrix} \overset{.o}{l}_{\widetilde{\beta}_0} \\ \overset{.o}{l}_{\widetilde{\beta}_X} \\ \overset{.o}{l}_{\widetilde{\beta}_V} \end{pmatrix}^T \right).$$

4.3.1 Extension for multiple truncated variables

The likelihood-based approach can be extended to handle more than one truncated covariate. We have discussed such a scenario in the sepsis study where cytokines TNF, IL-10, and IL-6 are truncated. We define Y as the outcome and (X, V) as the covariates where (Y, V) are fully observed and $X = (X_1, X_2)$ are truncated. The truncated indicator is defined as $R_1 = \begin{cases} 1 \; if \; x_1 > c_1 \\ 0 \; if \; x_1 \leq c_1 \end{cases}$ for the first truncated variable X_1 and $R_2 = \begin{cases} 1 \; if \; x_2 > c_2 \\ 0 \; if \; x_2 \leq c_2 \end{cases}$ for the second truncated variable X_2. The density of the data is

$$\begin{aligned} p(w) &= q(Y,V,X)^{R_1R_2} \left(\int_{x_1 \leq c_1} q(Y,V,X) d\mu(x) \right)^{(1-R_1)R_2} \times \\ &\quad \left(\int_{x_2 \leq c_2} q(Y,V,X) d\mu(x) \right)^{R_1(1-R_2)} \times \\ &\quad \left(\int_{x_1 \leq c_1, x_2 \leq c_2} q(Y,V,X) d\mu(x) \right)^{(1-R_1)(1-R_2)}, \\ \text{where } q(Y,V,X) &= f(y|v,x) f(v,x) = f(y|v,x) f(v) f(x) = f(y|v,x_1,x_2) f(v) f(x_1) f(x_2). \end{aligned}$$

The log likelihood is

$$\begin{aligned}
\log p(w) &= R_1 R_2 \log q(Y,V,X) + (1-R_1) R_2 \log\left(\int_{x_1\le c_1} q(Y,V,X)d\mu(x)\right) \\
&\quad + R_1(1-R_2)\log\left(\int_{x_2\le c_2} q(Y,V,X)d\mu(x)\right) \\
&\quad + (1-R_1)(1-R_2)\log\left(\int_{x_1\le c_1, x_2\le c_2} q(Y,V,X)d\mu(x)\right) \\
&= R_1 R_2 \{\log f(y|v,x) + \log f(x) + \log f(v)\} \\
&\quad + (1-R_1) R_2 \log\left(\int_{x_1\le c_1} f(y|v,x)f(x)f(v)d\mu(x)\right) \\
&\quad + R_1(1-R_2)\log\left(\int_{x_2\le c_2} f(y|v,x)f(x)f(v)d\mu(x)\right) \\
&\quad + (1-R_1)(1-R_2)\log\left(\int_{x_1\le c_1, x_2\le c_2} f(y|v,x)f(x)f(v)d\mu(x)\right).
\end{aligned}$$

The derivative of the log-likelihood wrt β is defined as: $\frac{\partial}{\partial\beta}\log p_{\beta,\gamma}(w) =$

$$\begin{aligned}
&R_1 R_2 \dot{l}(y,z,v|\beta) \\
&+ (1-R_1) R_2 \frac{\int_{x_1\le c_1} \dot{l}(y,x_1,x_2,v|\beta) f(y|v,x_1,x_2) f(x_1)f(x_2)f(v)d\mu(x_1)}{\int_{x_1\le c_1} f(y|v,x_1,x_2)f(x_1)f(x_2)f(v)d\mu(x_1)} \\
&+ R_1(1-R_2)\frac{\int_{x_2\le c_2} \dot{l}(y,x_1,x_2,v|\beta) f(y|v,x_1,x_2) f(x_1)f(x_2)f(v)d\mu(x_2)}{\int_{x_2\le c_2} f(y|v,x_1,x_2)f(x_1)f(x_2)f(v)d\mu(x_2)} \\
&+ (1-R_1)(1-R_2)\frac{\int_{x_1\le c_1, x_2\le c_2} \dot{l}(y,x_1,x_2,v|\beta) f(y|v,x_1,x_2) f(x_1)f(x_2)f(v)d\mu(x_1)d\mu(x_2)}{\int_{x_1\le c_1, x_2\le c_2} f(y|v,x_1,x_2)f(x_1)f(x_2)f(v)d\mu(x_1)d\mu(x_2)}
\end{aligned}$$

$$\begin{aligned}
&= R_1 R_2 \dot{l}(y,z,v|\beta) \\
&\quad + (1-R_1) R_2 E\left(\dot{l}(y,x_1,x_2,v|\beta)|y, x_1\le c_1, x_2, v\right) \\
&\quad + R_1(1-R_2)E\left(\dot{l}(y,x_1,x_2,v|\beta)|y, x_1, x_2\le c_2, v\right) \\
&\quad + (1-R_1)(1-R_2)E\left(\dot{l}(y,x_1,x_2,v|\beta)|y, x_1\le c_1, x_2\le c_2, v\right).
\end{aligned}$$

4.4 ESTIMATING EQUATION EXTENSION

For the estimating equation approach, an efficient score approach will be used via the inverse operator technique. The type of data is as before with an outcome Y which is fully observed and a cytokine variable X that is truncated at c. The remaining covariates V are fully observed. The complete data are (Y, X, V, C) where c is a constant value. Since X is truncated we claim X is either x, where $X > c$, or c, where $X \leq c$. We are treating the problem as a Type I fixed censoring problem. The complete data are $W^o = (Y, X, V, C) \sim Q$. The following assumptions are needed: 1) X and C are conditionally independent, given (V, Y) and 2) $X \perp V$ and $C \perp V$. We observe $W = (Y, \max(X, C), V, I(C \leq X)) = (Y, D, V, \Delta) \sim P$.

The density $f(Y, X, V, C)$ can be expressed as

$$\begin{aligned} q(Y, X, V, C) &= e(X, C|Y, V)h(Y, V) \\ &= \frac{f(Y|X, V)g(X)e(V)}{e(Y|V)} e(Y|C, V)e(C). \end{aligned}$$

The log-likelihood for model Q is

$$\begin{aligned} log f(Y, X, V, C) &= \log \frac{f(Y|X, V)f(X)f(V)}{f(Y|V)} f(Y|C, V)f(C) \\ &= \log f(Y|X, V) + \log f(X) + \log f(V) - \log f(Y|V) + \log f(Y|C, V) + \\ &\quad \log f(C). \end{aligned}$$

The score for β in model Q is

$$\begin{aligned}\dot{Q}_1 &= \dot{l}_1(y,z,v|\beta,\mathbf{Q}) \\ &= \frac{\partial}{\partial\beta}\log f(Y|X,V) \\ &= \begin{pmatrix}1\\ x\\ v\end{pmatrix}\left(y-\frac{\exp\left(\beta_0+\beta x+\beta_v v\right)}{1+\exp\left(\beta_0+\beta x+\beta_v v\right)}\right).\end{aligned}$$

Correctly identifying the nuisance parameters and determining the nuisance parameters that add information to the observed score is crucial to construction of the efficient score. The following parameters are treated as nuisance parameters: $\eta = (\log f(X),\ \log f(V),\ \log f(C),\ \log f(Y|C,V),\ \log f(Y|V))$. The scores for the nuisance parameters $f(X)$, $f(V)$, $f(C)$, $f(y|c,v)$, and $f(Y|V)$ in model Q are $\dot{Q}_2 = \dot{l}_2 a_2(x|\mathbf{Q}) = a_2(X) = \frac{\partial}{\partial k}\log g_k(X)$, $\dot{Q}_3 = \dot{l}_3 a_3(v|\mathbf{Q}) = a_3(V) = \frac{\partial}{\partial k}\log f_k(V)$, $\dot{Q}_4 = \dot{l}_4 a_4(c|\mathbf{Q}) = a_4(c) = \frac{\partial}{\partial k}\log g_k(C)$, $\dot{Q}_5 = \dot{l}_5 a_5(Y,V,C|\mathbf{Q}) = a_5(Y,V,C) = \frac{\partial}{\partial k}\log g_k(Y,V,C)$, and $\dot{Q}_6 = \dot{l}_6 a_6(Y,V|\mathbf{Q}) = a_6(Y,V) = \frac{\partial}{\partial k}\log g_k(Y,V)$, respectively. We assume that $\dot{Q}_2$, $\dot{Q}_3$, $\dot{Q}_4$, $\dot{Q}_5$, $\dot{Q}_6$ are mutually orthogonal and $\dot{Q}^* = \left\{\dot{Q}_1+\dot{Q}_2+\dot{Q}_3+\dot{Q}_4+\dot{Q}_5+\dot{Q}_6\right\} = \{h_1(y,x,c,v) + h_2(x) + h_3(v) + h_4(c) + h_5(y,v,c) + h_6(y,v)\}$. The only nuisance parameter that is not orthogonal to $\dot{Q}_1$ is $\dot{Q}_2$. To solve the efficient score we must find the function $a(X) \in L_2^0(G)$.

Our problem is a left censoring problem which leads us to discuss counting process concepts. Counting processes will be used to aid in solving the efficient score. The following counting process notation is from Keiding (1992) which describes the left censoring process as a left filtering process by use of the Aalen filter. According to Andersen *et al.* (1993) we

can analyze left censored data by handling the counting process as being left-filtered or as being observed with delayed entry.

The counting process for left filtering is (Keiding uses notation $\{t, C, Y\}$ and we use $\{d, H, E\}$ respectively)

$$N_i^H(d) = \int_0^d H_i(u) dN_i(u) = I\{X_i \leq d, H_i(X_{i-}) = 1\},$$

where the superscript H represents left censoring, $H_i(u) = I\{C_i < u \leq X_i\}$, and $N(d) = \sum_{i=1}^{n} N_i(d) = \sum_{i=1}^{n} I(X_i \leq d)$. In this case, $H_i(u)$ is the filtering process for left filtering and right censoring. $N(d)$ is the counting process for right censoring studied by Aalen. The intensity process of $N_i^H(d)$ is

$$\alpha(u) H_i(u) E_i(u) = \alpha(u) E_i^H(u),$$

where $E_i(u) = I(X_i \geq u)$ and $E_i^H(u) = I\{H_i(u) = 1, X_i > u\}$. Note that $E_i(u)$ represents the subjects at risk at time u under the filtering process $H_i(u)$. The sigma field is defined as

$$\mathcal{G}_d = \sigma\{I(C \leq d), CI(C \leq d), I(X \leq d), XI(X \leq d), Y, V\}.$$

Keiding has proven that left filtering is independent, that is, $f(C, X | C < X) = g(c) f(x)$ (Andersen *et al.*,1993, p. 49 and 166).

The $\alpha(u)$ from the intensity process of $N_i^H(d)$ is defined as

$$\begin{aligned} \alpha &= \frac{f(D|y,v)}{S(D|y,v)} \\ &= \frac{f(y|D,v) f(D)}{f(y|v)} \frac{F(y|v)}{F(y|v) - F(y|D,v) F(D)} \end{aligned}$$

where $S(D|y,v) = 1 - F(D|y,v)$. We define the martingale for left censored data as

$$\begin{aligned} M &= N^H(d) - \Lambda^H(d,\beta) \\ &= \sum_{i=1}^{n} I\{X_i \leq d, H_i(X_{i-}) = 1\} - \int_0^d \alpha(s,\beta) I\{X_i \geq u\} ds. \end{aligned}$$

More specifically we define the martingale and the adjusted martingale for the observed data as:

$$M(d) = I(c < X \leq d) - \int_0^d \alpha(u,\beta) I\{X \geq d\} ds$$

and

$$M_{uc}(d) = I(X \vee C \leq d, \Delta = 1) - \int_0^d \alpha(u,\beta) I\{X \vee C \geq u\} ds.$$

Let the score be defined as $\Psi(X,y,v) = \frac{\partial}{\partial \beta} \log f(Y|X,V)$. Let the operator R be defined as $R\Psi(u,y,v) = \Psi(u,y,v) - E(\Psi(X,y,v)|y,v,X > u)$ (Bickel *et al.*, 1993). The L operator is defined as $Lb(U) = \int_{-\infty}^{\infty} b dM$. The conditional expectation $E(\Psi(X,y,v)|y,v,X > u)$ is defined as

$$\begin{aligned} E(\Psi(X,y,v)|y,v,X > u) &= \frac{\int_u^\infty \Psi(X,y,v) f(y,v,x) dx}{\int_u^\infty f(y,v,x) dx} \\ &= \frac{\int_u^\infty \Psi(X,y,v) f(y|v,x) f(x) f(v) dx}{\int_u^\infty f(y|v,x) f(x) f(v) dx}. \end{aligned}$$

Conditioning on the filtration process, the conditional expectation of the observed score for the parameter of interest and nuisance parameter are:

$$\begin{aligned} E(\dot{l}_1(W^0|\beta,\mathbf{Q})|\mathcal{G}_d) &= E(\Psi(X,y,v)|\mathcal{G}_d) = \int_0^d R\Psi(X,y,v) dM \\ E(\dot{l}_2 a(W^0|\beta,\mathbf{Q})|\mathcal{G}_d) &= E(a(X,y,v)|\mathcal{G}_d) = \int_0^d Ra(X,y,v) dM. \end{aligned}$$

For model P the score for the estimation of β is also the observed score:

$$\begin{aligned}
\dot{l}_1(W|\beta,\mathbf{P}) &= E(\dot{l}_1(W^0|\beta,\mathbf{Q})|W) \\
&= \int_c^d R\Psi(X,y,v)dM \\
&= \int_0^\infty R\Psi(X,y,v)dM_{uc} \\
&= \Delta\Psi(d,y,v) + (1-\Delta)\,E(\Psi(X,y,v)|X>d,y,v).
\end{aligned}$$

The observed score for the estimation of a is:

$$\begin{aligned}
\dot{l}_2 a(W|\beta,\mathbf{P}) &= E(\dot{l}_2 a(W^0|\beta,\mathbf{Q})|W) \\
&= \int_c^d Ra(X,y,v)dM \\
&= \int_0^\infty Ra(X,y,v)dM_{uc} \\
&= \Delta a(X,y,v) + (1-\Delta)\,E(a(X,y,v)|X>d,y,v).
\end{aligned}$$

The above are true since the following holds:

$$\begin{aligned}
&\int_0^\infty R\Psi(X,y,v)dM_{uc} \\
=\ &\int_0^\infty R\Psi(X,y,v)d\left(I(X\vee C\le d,\Delta=1) - \int_0^d \alpha(u,\beta)I\{X\vee C\ge u\}du\right) \\
=\ &\int_0^\infty [(R\Psi(X,y,v)dI(X\vee C\le d,\Delta=1)dt - R\Psi(X,y,v)d\Lambda(d|y,v))] \\
=\ &\Delta R\Psi(d,y,v) + E(\Psi(X,y,v)|X>d,y,v) \\
&\left(\text{since } \int_0^\infty R\Psi(X,y,v)d\Lambda(d|y,v) = E(\Psi(X,y,v)|X>d,y,v) \text{ due to } L\circ R = identity\right) \\
=\ &\Delta(\Psi(d,y,v) - E(\Psi(X,y,v)|X>d,y,v)) + E(\Psi(X,y,v)|X>d,y,v) \\
=\ &\Delta\Psi(d,y,v) + (1-\Delta)E(\Psi(X,y,v)|X>d,y,v).
\end{aligned}$$

The efficient score for estimation of β is $\dot{l}_1^*(W,P|\beta,\mathbf{P}) = \dot{l}_1 - \Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$. The following theorems are from Bickel *et al.* (1993). We will use the method via inversion (Bickel

et al. p.79) of A^TA and projection $\Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$ by applying theorem A.2.2 to solve for $\dot{l}_1^*(W,P|\beta,\mathbf{P})$. According to theorem A.2.2, $\Pi_0(\cdot|H_0) = A\left(A^TA\right)^{-1}A^T\cdot$, implying that $\Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right) = A\left(A^TA\right)^{-1}A^T\dot{l}_1$. According to Theorem 3.4.1, the efficient score can be solved by $\dot{l}_1^*(W,P|\beta,\mathbf{P}) = \dot{l}_1 - \Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$.

The first step is to determine the projection of the score of the parameter of interest, $\dot{l}_1$, onto the score of the nuisance parameter $\dot{\mathbf{P}}_\eta$. This will enable us to calculate the efficient score of the parameter of interest, $\dot{l}_1^*(W,P|\beta,\mathbf{P}) = \dot{l}_1 - \Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$. A few items are needed to calculate $\Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$. We first need to define the operator $A : \dot{\mathbf{Q}}_\eta \to \dot{\mathbf{P}}$ by

$$\begin{aligned} Aa &= E(a_1(X)|W) = \int_0^\infty a(D)dM_{uc} \\ &= \Delta a(d) + (1-\Delta)\int_0^\infty a(D)I\{d \geq u\}\lambda(u|y,v)ds. \end{aligned}$$

Now we define the adjoint of A as $A^T : \dot{\mathbf{P}} \to \dot{\mathbf{Q}}_\eta$ by

$$\begin{aligned} A^Tb &= E(b(W)|X) - E(b(W)) = E(b(W)|X) \\ &= E(b(Y,D,V,\Delta)|X). \end{aligned}$$

The adjoint of the observed score is defined as $A^T\dot{l}_1$ where

$$\begin{aligned} A^T\dot{l}_1 &= A^T\int_0^\infty R\Psi(X,y,v)dM_{uc} \\ &= A^T\left[\Delta\Psi(d,y,v) + (1-\Delta)E(\Psi(X,y,v)|X>d)\right] \\ &= E\left[\Delta\Psi(d,y,v) + (1-\Delta)E(\Psi(X,y,v)|X>d)|X\right] \\ &= \int_v\int_y\int_\Delta \left[\Delta\Psi(d,y,v) + (1-\Delta)E(\Psi(X,y,v)|y,v,X>d)\right]q(y,v,\Delta,d|x)dydvd\Delta \\ &= \int_v\int_y \begin{bmatrix} \Delta\Psi(d,y,v)q(y,v,\Delta=1,d|x) \\ +(1-\Delta)E(\Psi(X,y,v)|y,v,X>d)q(y,v,\Delta=0,d|x) \end{bmatrix} dydv. \end{aligned}$$

For the adjoint there are two scenarios for our case: 1) $X > c$ and 2)$X \leq c$. For the first scenario, when $X > c$, we have $\Delta = 1$ and

$$\begin{aligned} A^T\dot{l}_1 &= \int_v \int_y \Delta\Psi(d,y,v)q(y,v,\Delta=1,d|x)dydv \\ &= \int_v \int_y \Psi(d,y,v)q(y,v,\Delta=1,d|x)dydv \text{ where } d=x \text{ is fixed.} \end{aligned}$$

For the second scenario, when $X \leq c$, we have $\Delta = 0$ and

$$\begin{aligned} A^T\dot{l}_1 &= \int_v \int_y [(1-\Delta)E(\Psi(X,y,v)|y,v,X>d)q(y,v,\Delta=0,d|x)]\, dydv \\ &= \int_v \int_y [E(\Psi(X,y,v)|y,v,X>d)q(y,v,\Delta=0,d|x)]\, dydv \end{aligned}$$

with

$$\begin{aligned} q(y,v,\Delta &= 0,d|x) = q(y,v,\Delta=0,d)/F(x\leq c) \\ &= q(y,v,\Delta=0,d)/F(c). \end{aligned}$$

To obtain the adjoint of the observed score, we need to calculate $q(y,v,\Delta,d)$, where the notation is borrowed from Andersen *et al.* (1993). According to Andersen (p.142 and p.166), the full likelihood is

$$\begin{aligned} L_t^*(\theta,\phi) &= L_t^{c'}(\theta)L_t^{''}(\theta,\phi) \\ &= \left(\frac{S_x(X,\theta|y,v)}{S_x(C,\theta|y,v)}a_x(X,\theta|y,v)\right)^{\Delta} \times \\ &\quad S_c(C,\phi|y,v)a_c(C,\phi|y,v)S_x(C,\theta|y,v)^{1-\Delta}F_x(C,\theta|y,v)^{\Delta}f(y,v) \\ &= \left(\frac{f_x(X,\theta|y,v)}{S_x(C,\theta|y,v)}F_x(C,\theta|y,v)\right)^{\Delta} f_c(C,\phi|y,v)S_x(C,\theta|y,v)^{1-\Delta}f(y,v). \end{aligned}$$

The density $q(y,v,\Delta,d)$ should be equivalent to the likelihood $L_t^*(\theta,\phi)$:

$$q(y,v,\Delta,d) = \left(\frac{f_x(X,\theta|y,v)}{S_x(C,\theta|y,v)}F_x(C,\theta|y,v)\right)^{\Delta} f_c(C,\phi|y,v)S_x(C,\theta|y,v)^{1-\Delta}f(y,v).$$

The pieces of $q(y, v, \Delta, d)$ can be calculated the following way. First to derive the density $f_x(X, \theta|y, v)$:

$$\begin{aligned} f_x(X, \theta|y, v) &= \frac{f_x(X, y, v)}{f(y, v)} \\ &= \frac{f_x(y|X, v)f(X)}{f(y|v)}. \end{aligned}$$

The detailed derivation of the density $q(y, v, \Delta, d)$ is:

$$\begin{aligned} q(y, v, \Delta, d) &= \left(\frac{f_x(X, \theta|y, v)}{S_x(C, \theta|y, v)} F_x(C, \theta|y, v)\right)^{\Delta} f_c(C, \phi|y, v) S_x(C, \theta|y, v)^{1-\Delta} f(y, v) \\ &= \left(\frac{\frac{f_x(y|X,v)f_x(X)}{f(y|v)}}{1 - \frac{F_x(y|C,v)F_x(C)}{F(y|v)}} \frac{F_x(y|C, v)F_x(C)}{F(y|v)}\right)^{\Delta} \frac{f_c(y|C, v)f_c(C)f(v)}{f(y, v)} \times \\ &\quad \left(1 - \frac{F_x(y|C, v)F_x(C)}{F(y|v)}\right)^{1-\Delta} f(y, v) \\ &\propto \left(\frac{f_x(y|X, v)f_x(X)}{f(y|v)\left[F(y|v) - F_x(y|C, v)F_x(C)\right]} F_x(y|C, v)F_x(C)\right)^{\Delta} f(y|v)f(v) \times \\ &\quad \left(\frac{F(y|v) - F_x(y|C, v)F_x(C)}{F(y|v)}\right)^{1-\Delta} \end{aligned}$$

(since C is constant $f_c(y|C, v) = f(y|v)$ and $f_c(C) = k$) .

Since C is a constant, $f_c(C)=\frac{\sum I(X \leq c)}{n}$ and $f_c(y|C, v) = f_c(y|v)$. A consistent estimate of θ is based on a model with $f(y|x, v, \theta)$. For $f(y|x, v, \theta)$, we would fit a logistic model including x and v to estimate θ. We define the density of X as $f_x(c) = F_x(c) = \frac{\sum I(X \leq c)}{n}$ and $f_x(x) = F_x(x) - F_x(x-)$, where $F_x(x) = \frac{\sum I(X \leq x)}{n}$ and n is the total sample size. More precisely recalling that $y = \{0, 1\}$:

$$F(y|v) - F_x(y|C, v)F_x(C) = f(y = 0|v) - f_x(y = 0|C, v)\frac{\sum I(X \leq c)}{n} \quad \text{if } y = 0$$

and

$$F(y|v) - F_x(y|C, v)F_x(C) = 1 - 1\frac{\sum I(X \leq c)}{n} = 1 - \frac{\sum I(X \leq c)}{n} \quad \text{if } y = 1.$$

Also, since C is a constant, $f_x(c)$ and $F_x(C)$ are always the same value. Since C is the same value for all subjects, $F_x(y|C, v)$ varies for each combination of (y, v).

We need to calculate the adjoint of A onto A which is defined as $A^T Aa : \dot{\mathbf{Q}}_2 \rightarrow \dot{\mathbf{Q}}_2$:

$$\begin{aligned} A^T Aa &= E(Aa|X) \\ &= E(E(a|W)|X) \\ &= E\left(\left[\Delta a(d) + (1-\Delta)\int_0^\infty a(D)I\{d \geq s\}\lambda(s|y,v)ds.\right]|X\right) \\ &= \Delta a(d) + E\left((1-\Delta)\int_0^\infty a(D)I\{X \vee C \geq D\}\lambda(D|y,v)dD|X\right). \end{aligned}$$

The next step is to calculate the inverse operator of $A^T A$ s.t. $\left(A^T A\right)^{-1} A^T Aa = a$. It can be very difficult to solve an inverse operator of an integral operator. If we can assume it is acceptable to sum rather than integrate over v, y, D then we can solve a linear operator.

$$\begin{aligned} A^T Aa &= \Delta a(d) + (1-\Delta)\int_v\int_y\int_0^\infty a(D)I\{X \vee C \geq D\}\lambda(D|y,v)dDq(y,v,d,\Delta=0|x)dydv \\ &= \Delta a(d) + (1-\Delta)\sum_V\sum_Y\sum_D a(D)I\{X \vee C \geq D\}\lambda(D|y,v)q(y,v,d,\Delta=0|x) \end{aligned}$$

$$\rightarrow \left(A^T A\right)^{-1} = \begin{pmatrix} \sum_V\sum_Y q(d_1,y,v,\Delta=0)\lambda(d_1|y,v) & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1}.$$

To obtain $\left(A^T A\right)^{-1}$ one calculates the conditional expectation of $b(x)$ conditioning on X and not truncated $(\Delta = 1)$

$$\left(A^T A\right)^{-1} b(x) = E(b(x)|\Delta = 1, X).$$

Using the above solution, the inverse operator of the adjoint of A onto A of the adjoint of A onto the observed score gives:

$$\begin{aligned}\left(A^TA\right)^{-1}A^T\dot{l}_1 &= E\left[\int_v\int_y\int_\Delta[\Delta\Psi(d,y,v)+(1-\Delta)E(\Psi(X,y,v)|y,v,X>d)]\times\right.\\ &\quad \left. q(y,v,\Delta,d|x)dydvd\Delta|\Delta=1,X\right]\\ &= \int_v\int_y\Psi(d,y,v)q(y,v,\Delta=1,d|x)dydv.\end{aligned}$$

The last step needed to calculate $\Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$ is to calculate the A operator onto $\left(A^TA\right)^{-1}A^T\dot{l}_1$ which is defined as $A\left(A^TA\right)^{-1}A^T\dot{l}_1 = E(\left(A^TA\right)^{-1}A^T\dot{l}_1|W)$. We can now reach our ultimate goal which is the efficient score. To do this, we use the projection of the score of the parameter of interest onto the score of the nuisance parameter space, $\Pi_0\left(\dot{l}_1|\dot{\mathbf{P}}_\eta\right)$, to calculate the efficient score of β. The efficient score is:

$$\begin{aligned}l^* &= \dot{l}_1 - A\left(A^TA\right)^{-1}A^T\dot{l}_1 =\\ &\quad \dot{l}_1 - E(\left(A^TA\right)^{-1}A^T\dot{l}_1|W)\\ &= \dot{l}_1 - \int_0^\infty\left(A^TA\right)^{-1}A^T\dot{l}_1dM_{uc}.\end{aligned}$$

We solve for the new estimates of β by the Newton-Raphson estimator $\widehat{\beta} = \widetilde{\beta} + I^{-1}\begin{pmatrix}\sum_i l^*_{\beta_0,i}\\ .\\ .\\ \sum_i l^*_{\beta_k,i}\end{pmatrix}$

where $I = \begin{pmatrix}l^*_{\beta_0,i}\\ .\\ .\\ l^*_{\beta_k,i}\end{pmatrix}\begin{pmatrix}l^*_{\beta_0,i}\\ .\\ .\\ l^*_{\beta_k,i}\end{pmatrix}^T$.

CHAPTER 5

SIMULATION STUDIES AND EXAMPLE FOR TRUNCATED COVARIATE DATA

This chapter demonstrates the utility of the likelihood-based method compared to standard methods. The first section covers simulation studies. The following section uses the sepsis study as an example.

5.1 SIMULATION STUDIES

Simulation studies have been performed to compare four methods in the presence of truncated covariate data. The outcome, Y, is binary and the one covariate, $\log(X)$, in the model is normally distributed with mean μ and variance σ^2, i.e. $log(x) \sim N(\mu, \sigma^2)$.

Logistic regression is the model of choice for this setting, where we model $logit(\Pr(Y = 1|x)) = \beta_0 + \beta_x \log(x)$. Summary statistics facilitate comparison of four different modeling approaches for this problem. The summary statistics of the estimators of β calculated were the mean of the coefficient of the replicates, mean of the variance of the coefficient of the replicates, MSE, and 95% coverage.

The four approaches are presented in the following tables. In each case, the first column is standard logistic regression (LR) with all true values of X; here nothing is truncated. The second column is our likelihood-based approach. The starting values used for the likelihood-based approach are from standard logistic regression with $log(X)$ in the model, including only nontruncated values. The third column is standard logistic regression with $log(X)$ in the model, including only nontruncated values. In this case, the sample size will be reduced to including subjects who have a value of X above the threshold value. The fourth column is a standard logistic regression with $log(D)$ in the model where $D = max(X, C)$; i.e. D is the maximum of X and C, the threshold value. In this case, all subjects are included in the analysis with either their observed value or a truncated value.

Twelve scenarios have been simulated. The first three tables use data generated from $log(X) \sim N(2, 1)$ and the last three tables use data generated from $log(X) \sim N(3, .7)$. In Table 5.1, $\beta_0 = 1$ and $\beta_x = -0.5$ with 50% and 35% truncation. In Table 5.2, $\beta_0 = -2$ and $\beta_x = 1$ with 50% and 35% truncation. In table 5.3, $\beta_0 = -4$ and $\beta_x = 2$ with 50% and 35% truncation. In Table 5.4, $\beta_0 = 0$ and $\beta_x = 0$ with 50% and 35% truncation. In Table 5.5, $\beta_0 = -3$ and $\beta_x = 1$ with 50% and 35% truncation. In table 5.6, $\beta_0 = -6$ and $\beta_x = 2$ with 50% and 35% truncation.

5.1.1 Results

Our simulation studies demonstrate that our likelihood-based method has better variance correction than the two competitors. The likelihood-based method is less biased than the method using standard logistic regression with $d = max(x, c)$, but is more biased than the complete case method. The complete case method has no bias, but results in an inflated variance. The standard logistic regression with d has better variance correction than the complete case. Our method has the smallest MSE due to some bias and variance reduction. As the amount of truncation decreases, the likelihood method becomes less biased. When the relationship between the outcome and covariate is small, $\beta_x = 0$, there is no bias for any of the methods. The likelihood-based method does not perform as well when the relationship between the outcome and the covariate is large, $\beta_x = 2$. When $\beta_x = 2$, the likelihood method performs better when there is less truncation. In this scenario, the likelihood method has a larger bias, but the smallest variance causing the MSE to be larger than the complete case method. Overall, the 95% coverage probabilities are accurate for the likelihood-based method, except when $\beta_x = \{1, 2\}$ and the amount of truncation increases (50%).

The likelihood-based method is an improvement but still needs a bias correction. The coefficients are overestimated in the likelihood-based method and the standard method using the truncation value. The method of filling in the truncation values assumes all truncated values equal the truncation value, forcing a steeper slope and larger intercept term (in absolute terms). The likelihood-based method assumes that the truncated values are between the smallest value possible ($X = 0$) and the truncated point. The likelihood method is averaging the scores over this range which essentially yields a coefficient value somewhere

between the complete case coefficient and the fill-in truncated method. The slopes and intercept are being weighted down for LB but not as much as with filling in the truncated value. This issue is known as the bias-variance trade-off. We have reduced the variance in exchange for bias. We are still in the process of solving the efficient score approach and suspect the efficient score will correct for both the bias and variance.

Table 5.1: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 200, log(X)∼N(2,1), 50% and 35% truncation, β_0=1, β_x=-.5

		$P(X < c) = .50$			$P(X < c) = .35$		
	LR1	LB	LR2	LR3	LB	LR2	LR3
	$n = 200$	$n = 200$	$n = 100$	$n = 200$	$n = 200$	$n = 130$	$n = 200$
β_0							
Coef	1.017	1.504	1.031	1.817	1.239	1.011	1.460
Var	0.120	0.290	1.108	0.444	0.184	0.554	0.255
MSE	0.124	0.548	1.105	1.122	0.243	0.543	0.471
95% Cov	0.949	0.857	0.951	0.782	0.924	0.963	0.860
β_x							
Coef	-0.505	-0.666	-0.513	-0.761	-0.584	-0.505	-0.656
Var	0.024	0.053	0.141	0.076	0.037	0.082	0.049
MSE	0.024	0.082	0.141	0.147	0.044	0.081	0.074
95% Cov	0.957	0.907	0.956	0.861	0.948	0.959	0.912

Table 5.2: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 200, log(X)∼N(2,1), 50% and 35% truncation, β_0=-2, β_x=1

		$P(X < c) = .50$			$P(X < c) = .35$		
	LR1	LB	LR2	LR3	LB	LR2	LR3
	$n = 200$	$n = 200$	$n = 100$	$n = 200$	$n = 200$	$n = 130$	$n = 200$
β_0							
Coef	-2.040	-3.058	-2.141	-3.759	-2.476	-2.073	-2.934
Var	0.169	0.422	1.515	0.665	0.258	0.706	0.344
MSE	0.172	1.586	1.566	3.872	0.485	0.710	1.239
95% Cov	0.961	0.643	0.960	0.433	0.885	0.957	0.649
β_x							
Coef	1.019	1.371	1.060	1.598	1.180	1.034	1.339
Var	0.036	0.082	0.209	0.123	0.054	0.113	0.071
MSE	0.036	0.233	0.221	0.507	0.088	0.117	0.193
95% Cov	0.954	0.780	0.959	0.631	0.911	0.950	0.783

Table 5.3: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 200, log(X)∼N(2,1), 50% and 35% truncation, β_0=-4, β_x=2

		$P(X < c) = .50$			$P(X < c) = .35$		
	LR1	LB	LR2	LR3	LB	LR2	LR3
	$n = 200$	$n = 200$	$n = 100$	$n = 200$	$n = 200$	$n = 130$	$n = 200$
β_0							
Coef	-4.111	-6.276	-4.343	-7.965	-4.894	-4.170	-5.796
Var	0.390	1.067	3.242	1.731	0.578	1.253	0.697
MSE	0.457	6.583	3.474	18.409	1.416	1.310	4.077
95% Cov	0.942	0.404	0.963	0.085	0.836	0.959	0.446
β_x							
Coef	2.055	2.871	2.154	3.485	2.369	2.082	2.720
Var	0.089	0.223	0.519	0.359	0.130	0.235	0.162
MSE	0.107	1.073	0.577	2.787	0.285	0.254	0.728
95% Cov	0.941	0.571	0.962	0.282	0.879	0.959	0.611

Table 5.4: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 200, log(X)∼N(3,.7), 50% and 35% truncation, β_0=0, β_x=0

		$P(X < c) = .50$			$P(X < c) = .35$		
	LR1	LB	LR2	LR3	LB	LR2	LR3
	$n = 200$	$n = 200$	$n = 100$	$n = 200$	$n = 200$	$n = 130$	$n = 200$
β_0							
Coef	-0.005	-0.031	-0.055	-0.040	-0.020	-0.020	-0.024
Var	0.400	1.016	3.144	1.375	0.659	1.728	0.860
MSE	0.395	1.021	3.235	1.394	0.660	1.792	0.865
95% Cov	0.956	0.959	0.949	0.957	0.948	0.957	0.952
β_x							
Coef	0.005	0.013	0.019	0.015	0.009	0.010	0.011
Var	0.042	0.097	0.246	0.126	0.067	0.147	0.084
MSE	0.042	0.098	0.254	0.128	0.067	0.152	0.085
95% Cov	0.948	0.958	0.951	0.960	0.954	0.957	0.955

Table 5.5: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 200, log(X)∼N(3,.7), 50% and 35% truncation, β_0=-3, β_x=1

		$P(X < c) = .50$			$P(X < c) = .35$		
	LR1	LB	LR2	LR3	LB	LR2	LR3
	$n = 200$	$n = 200$	$n = 100$	$n = 200$	$n = 200$	$n = 130$	$n = 200$
β_0							
Coef	-3.063	-4.439	-3.179	-5.085	-3.724	-3.095	-4.193
Var	0.541	1.396	4.104	1.957	0.883	2.151	1.134
MSE	0.520	3.467	4.058	6.329	1.361	2.052	2.526
95% Cov	0.961	0.800	0.958	0.694	0.927	0.964	0.829
β_x							
Coef	1.019	1.392	1.054	1.559	1.205	1.030	1.331
Var	0.057	0.136	0.332	0.185	0.091	0.189	0.114
MSE	0.054	0.291	0.328	0.502	0.128	0.181	0.222
95% Cov	0.956	0.855	0.956	0.780	0.942	0.964	0.884

Table 5.6: Summary statistics for logistic regression coefficients based on 1000 replications with a sample size of 200, log(X)∼N(3,.7), 50% and 35% truncation, β_0=-6, β_x=2

		$P(X < c) = .50$			$P(X < c) = .35$		
	LR1	LB	LR2	LR3	LB	LR2	LR3
	$n = 200$	$n = 200$	$n = 100$	$n = 200$	$n = 200$	$n = 130$	$n = 200$
β_0							
Coef	-6.120	-9.125	-6.375	-10.635	-7.441	-6.202	-8.407
Var	0.969	2.703	7.054	3.854	1.582	3.327	1.916
MSE	0.951	13.053	7.531	26.494	3.676	3.417	7.919
95% Cov	0.955	0.532	0.952	0.331	0.839	0.959	0.594
β_x							
Coef	2.039	2.883	2.118	3.295	2.423	2.067	2.697
Var	0.104	0.270	0.604	0.382	0.166	0.311	0.202
MSE	0.103	1.120	0.654	2.185	0.353	0.324	0.716
95% Cov	0.955	0.621	0.947	0.448	0.871	0.954	0.672

5.2 SEPSIS STUDY

The sepsis study described earlier serves as an example. It is of interest to assess the relationship between inflammatory markers and severe sepsis status. The literature suggests using the inflammatory markers for prediction. For this data, IL-6 has the lowest amount of truncation at 15% and parameter estimates from regression models including IL-6 should not be greatly impacted by the truncation. Overall descriptive statistics of the

three inflammatory markers, TNF, IL-10, and IL-6, are reported in Table 5.7. Descriptive statistics are reported for both the nontruncated samples (NT) and all samples using the fill-in truncated samples with detection limit D, where $D = max(X, C)$. As expected, all of the inflammatory markers are negatively skewed, also known as skewed to the left. If a predictor in a regression model is highly skewed, the linear relationship between it and the outcome can be violated and particular points may exert influence on the coefficients and/or may be poorly fit in the model. If the fit of the model is affected, then transformation of the covariate or removal of points is suggested. We do not suggest removal of any points so transformation is the optimal choice. Traditionally, the transformation of cytokine data chosen is the natural logarithm. As can be seen from Table 5.7, the raw cytokine data are so highly skewed to the left that the mean is twice the median for both the nontruncated and all other samples. The log transformation draws the values in the upper range closer to a feasible range. Descriptive statistics by sepsis status for inflammatory markers are reported in Table 5.8.

Since we are using a likelihood-based method, we need to determine the distribution of the covariate. In this section, we will only deal with a simple logit model, i.e. only include one covariate. Since we know that the log transformation of the cytokine data is normally distributed, the cytokine data has a lognormal distribution. This indicates that we should analyze the data two ways: 1) include the log transformation of the data in the model and assume it is normally distributed (Table 5.9) and 2) include the raw data in the model and assume it is lognormally distributed (Table 5.10). Based on the regression results there is a greater improvement for the likelihood-based method for Interleukin 10 and 6 when the log transformation of the data is included rather than the raw data. For IL-10 and IL-6,

the likelihood-based method corrects for the coefficients and the variance more than the fill-in method when the data is log transformed. It should be noted that if the raw data is used the likelihood-based method and fill-in method are comparable for the interleukin data. For TNF it appears to matter whether the data are transformed or not. When the data are log transformed the likelihood-based method produces more biased estimates but smaller variances than the fill-in method. When the raw data are used in this model, the likelihood-based method corrects for bias and the variance more so than the fill-in method. Diagnostics were reviewed. A few raw data points exerted influence and possible poor fit. Even with this occurring, it may be that the correct distribution is the lognormal and not the normal distribution of the transformed data. This suggests the best approach for TNF is to analyze the raw data and assume the data are lognormally distributed.

Inference is the same for all of the methods when the data are log transformed (Table 5.9); that is, all baseline inflammatory markers are found to be (statistically) significantly associated with the development of severe sepsis. When the raw data are used in the models inference varies by method for TNF but does not for IL-6 and IL-10. According to LB (p=.0006) and LR3 (p=.0513), TNF is (statistically) significantly associated with the development of severe sepsis. Although, according to LR2, TNF is not statistically associated (p=.1053) with the development of severe sepsis. Regardless of the transformation or method, the risk of developing severe sepsis increases with the concentration of the inflammatory markers.

In conclusion, the nontruncated method is the most conservative method of all due to it having the largest variance. The nontruncated method is more likely to not reject due to the larger variance; i.e. the power is probably smaller with the nontruncated method.

The likelihood-based method still appears to have the best variance and moderate bias correction.

Table 5.7: Descriptive statistics for cytokines

	Raw		Log Transformed	
	NT	D	NT	D
TNF				
N	1095	1809	1095	1809
Mean (SD)	15.0 (38.2)	10.7 (30.2)	2.31 (0.69)	1.95 (0.70)
Median	8.8	5.4	2.17	1.69
(Min, Max)	(4.1, 944)	(4, 944)	(1.41, 6.85)	(1.39, 6.85)
IL-10				
N	909	1809	909	1809
Mean (SD)	40.8 (96.5)	23.0 (70.7)	2.94 (1.05)	2.28 (1.00)
Median	14.8	5.1	2.69	1.63
(Min, Max)	(5.1, 1519)	(5, 1519)	(1.63, 7.33)	(1.61, 7.33)
IL-6				
N	1533	1811	1533	1811
Mean (SD)	506.7 (3604.6)	429.6 (3321.2)	4.15 (1.71)	3.75 (1.84)
Median	51.0	35.4	3.93	3.57
(Min, Max)	(2.1, 126000.0)	(2.0, 126000.0)	(0.74, 11.74)	(0.69, 11.74)

Note: NT=Not truncated values, $D = max(X, C)$

Table 5.8: Descriptive statistics for cytokines by severe sepsis status

	NT		D	
	No SS	SS	No SS	SS
TNF*				
N	782	313	1334	475
Mean (SD)	2.24 (0.65)	2.48 (0.78)	1.89 (0.65)	2.10 (0.81)
Median	2.13	2.31	1.63	2.60
(Min, Max)	(1.41, 6.85)	(1.41, 6.34)	(1.39, 6.85)	(1.39, 6.34)
IL-10*				
N	631	278	1334	475
Mean (SD)	2.88 (1.02)	3.09 (1.11)	2.21 (0.95)	2.47 (1.12)
Median	2.63	2.83	1.61	1.93
(Min, Max)	(1.63, 7.33)	(1.63, 6.84)	(1.61, 7.33)	(1.61, 6.84)
IL-6*				
N	1103	430	1336	475
Mean (SD)	3.99 (1.63)	4.57 (1.84)	3.56 (1.75)	4.29 (1.96)
Median	3.77	4.37	3.35	4.17
(Min, Max)	(0.74, 9.56)	(0.88, 11.74)	(0.69, 9.56)	(0.69, 11.74)

* The log transformation was taken of each cytokine value.

Note: SS=severe sepsis, NT=Not truncated values, $D = max(X, C)$

Table 5.9: Logistic regression results for cytokines with transformed data

	LB	LR2	LR3
ln(TNF)	$n = 1809$	$n = 1095$	$n = 1809$
β_0 (SE)	-1.69 (0.129)	-2.00 (0.236)	-1.84 (0.156)
β_x (SE)	0.35 (0.061)	0.46 (0.095)	0.40 (0.072)
$p - value\ for\ \beta_x$	<0.0001	<0.0001	<0.0001
ln(IL-10)	$n = 1809$	$n = 909$	$n = 1809$
β_0 (SE)	-1.48 (0.107)	-1.35 (0.210)	-1.60 (0.131)
β_x (SE)	0.21 (0.043)	0.18 (0.067)	0.24 (0.050)
$p - value\ for\ \beta_x$	<0.0001	0.0077	<0.0001
ln(IL-6)	$n = 1811$	$n = 1533$	$n = 1811$
β_0 (SE)	-1.78 (0.121)	-1.77 (0.155)	-1.85 (0.128)
β_x (SE)	0.20 (0.027)	0.19 (0.033)	0.21 (0.029)
$p - value\ for\ \beta_x$	<0.0001	<0.0001	<0.0001

Note: $log(X) \sim N(\mu, \sigma^2)$

Table 5.10: Logistic regression results for cytokines with raw data

	LB	LR2	LR3
TNF	$n = 1809$	$n = 1095$	$n = 1809$
β_0 (SE)	-1.073 (0.055)	-0.966 (0.073)	-1.080 (0.059)
β_x (SE)	0.0037 (0.0011)	0.0033 (0.0020)	0.0043 (0.0022)
$p - value\ for\ \beta_x$	0.0006	0.1055	0.0513
IL-10	$n = 1809$	$n = 909$	$n = 1809$
β_0 (SE)	-1.086 (0.056)	-0.880 (0.079)	-1.084 (0.057)
β_x (SE)	0.0020 (0.0006)	0.0014 (0.0007)	0.0021 (0.0008)
$p - value\ for\ \beta_x$	0.0009	0.0552	0.0057
IL-6	$n = 1811$	$n = 1533$	$n = 1811$
β_0 (SE)	-1.104 (0.056)	-1.013 (0.060)	-1.102 (0.056)
β_x (SE)	0.00017 (0.00004)	0.00015 (0.00004)	0.00017 (0.00004)
$p - value\ for\ \beta_x$	<0.0001	<0.0001	<0.0001

Note: $X \sim LogN(\mu, \sigma^2)$

CHAPTER 6

DISCUSSION

Unbiased estimating equations are the focus of this dissertation: predominantly semiparametric methods utilized to solve for regression parameters in the presence of missing covariate data. The scope of our research ranges from covariates missing by design and missing by happenstance to truncated covariates. In the event of expensive covariate data or difficulty of covariate data collection, sampling based approaches are applied to case-control and cohort studies to either balance data or reduce the cost of data collection. A semiparametric approach is used to solve missing data problems since these estimating equations produce an efficient estimator. A benefit to this approach is the avoidance of specifying a full likelihood. Misspecification of the likelihood can lead to incorrect estimates. However, difficulties with the efficient score approach arise in specifying a density, identifying the nuisance parameters, and construction of the score equations and score operators.

The first half of this dissertation discussed in detail semiparametric methodology and methodology developed to handle logistic regression with missing covariate information. The first aim of my dissertation was to evaluate the properties of an efficient score, an inverse

probability weighted estimating equation approach, for logistic regression with a covariate missing by design. In our simulation study, data was generated to mimic a case-cohort design where the covariate was missing by design. The efficient score was compared to two other pseudo-likelihood methods, and as anticipated the efficient score improved bias and efficiency of the estimators. Analysis of the results from our simulation study demonstrates that the efficient score approach yields the most improved estimates when a correlated surrogate of the missing covariate is available.

A future extension to the efficient score methodology is to address a missing continuous covariate in a regression model. Due to the high dimensions of the covariate data, smoothing techniques would be used to aid in solving a model with this type of data. An additional extension is a regression model with covariate data missing by happenstance. A third proposed extension is addressing multiple missing covariates in a regression model. Semiparametric methods will be used to address these three extensions.

The second aim of my dissertation was to develop a methodology for truncated covariate data with a binary outcome. To address this problem, we have developed two methods, a likelihood-based method and a semiparametric method, to handle a truncated covariate in a logistic regression model. The likelihood-based approach is solved, but the efficient score approach is still in the process of being solved. In our case, the truncated covariate is continuous. Simulation studies and a sepsis study from the University of Pittsburgh demonstrated and proved the properties of the estimator for the likelihood-based method compared to methods of using only the nontruncated samples and the fill-in method. Our simulation studies for the likelihood-based approach provide confirmation of our expectations. This method improves precision but does not eliminate bias of the coefficient estimates, increas-

ing the importance of continuing to solve the semiparametric component. Possible bias corrections include the bootstrap and jackknife methods. At this time, the approach yielding the most accurate inferences and the best bias-variance correction is the likelihood-based method. Future extensions of interest that we may pursue for the truncated piece are: multiple truncated covariates, (randomly) censored covariates, and a continuous outcome.

The original focus of my dissertation was the development of survival methods for missing data. This research was set aside temporarily to focus on truncated data. Often times the hypothesis of a study is to determine the predictors of a time-to-event outcome. The efficient score was developed for case-cohort designs. However, Nan demonstrated the properties for a discrete covariate in the presence of a surrogate of the covariate. In many studies the covariate of interest is continuous. We have recently solved the case-cohort problem with a continuous covariate to be discussed in a future paper. Smoothing techniques were implemented to solve for the case-cohort problem with a continuous covariate. Another possible extension is solving the efficient score for the Cox model when a covariate is missing by happenstance. In the event of data missing by happenstance, surrogates of the covariate may be unavailable.

Two other areas of interest are two survival endpoints and repeated measures data with missing covariate information (see Appendix for details). Modelling two survival endpoints is gaining popularity. Modelling two endpoints is often of interest from a design standpoint, for example, determining the sequence of events, or for predictive purposes. The main interest is to draw inferences on the relationship between the survival endpoints, but the endpoints may be influenced by covariates. We propose a two-stage approach via copulas (Shih and Louis, 1995), incorporating weighted pseudo-likelihoods, to solve for the marginal estimates.

A proposed extension is to model two endpoints in the presence of missing covariates. The goal is to measure the dependence between these two endpoints.

Since the advent of repeated measures methods, mainly generalized estimating equations, studies are often designed to collect repeated measures to determine changes over time, increasing the chances of data not being reported. It is conceivable that data is missing with the reason unknown, also known as missing by happenstance. We propose an extension of a method developed by Wei and Stram (1988) to handle data missing at random. The proposed method is a two-stage quasi-likelihood approach that employs estimating the parameters from the marginal estimating equations at each time point and the variance-covariance from the sandwich estimator. Our focus will be on two types of outcomes, continuous and binary, narrowing the applications to linear regression and logistic regression.

Based on our findings we suggest using a semiparametric approach for missing data. Our likelihood approach for truncated covariate data is an adequate starting point. Our results for the truncated problem suggest the need for modifications to regression models in the presence of truncated covariate data. As we have discussed there are numerous problems that need to be solved.

APPENDIX A

FUTURE WORK

METHODOLOGY FOR TWO ENDPOINTS AND MISSING COVARIATE INFORMATION

Modelling two survival endpoints is gaining popularity. Modelling two endpoints is often of interest from a design standpoint, for example determining the sequence of events, or for predictive purposes. The main interest is to draw inferences on the relationship between the survival endpoints, but the endpoints may be influenced by covariates. We propose a two-stage approach via copulas (Shih and Louis, 1995), incorporating weighted pseudo-likelihoods, to solve for the marginal estimates. A proposed extension is to model two endpoints in the presence of missing covariates. The goal is to measure the dependence between these two endpoints. Methods exist for modelling two endpoints adjusting for covariates and for modelling one endpoint adjusting for missing covariate data. However, at this juncture no method exists for modelling two endpoints with missing covariate data. This problem is of a practical nature since missing data is an every day nuisance that cannot be ignored. The main interest is to draw inferences on the relationship between two survival

endpoints. However, the endpoints may be influenced by covariate information. If covariates are missing at random, it has been shown in other literature (Chen and Little, 1999; Herring and Ibrahim, 2001) that the marginal model will be inefficient, unbiased, and inconsistent, which has implications for the bivariate model.

Modelling two endpoints is also of interest from a design standpoint, such as determining the sequence of events as well as for predictive purposes. Types of multiple endpoints include competing risks, recurring events, and different events. Methodology for different events will be addressed. A practical and straightforward method to model multiple endpoints is through the use of copulas. Copulas were first developed in the 1950s (Sklar, 1959) and their usefulness has gained attention in recent years. Multivariate distributions can be calculated with ease via copulas. However, as with any other method, drawbacks exist. One drawback which will need to be addressed is model selection in copulas.

Two-Stage Method

Event times are denoted as (T_{1i}, T_{2i}) and censored times are denoted as (C_{1i}, C_{2i}), where survival times $(Y_{1i}, Y_{2i}) = (\min(T_{1i}, C_{1i}), \min(T_{2i}, C_{2i}))$ and censoring indicators $(\delta_{1i}, \delta_{2i}) = (I\{Y_{1i} = T_{1i}\}, I\{Y_{2i} = T_{2i}\})$ are observed for each subject, $i = 1, ...n$. It is also assumed that (C_{1i}, C_{2i}) and (T_{1i}, T_{2i}) are independent. Due to the simplistic nature and statistical properties of the Shih and Louis two-stage method, it is the method of choice for analyzing bivariate survival data and will be extended to account for missing covariate information.

Two-stage parametric estimation is performed if one chooses a parametric form for the marginal distributions and bivariate distribution. For two-stage parametric estimation the

likelihood of θ is

$$\prod_{i=1}^{n} f(Y_{1i}, Y_{2i}; \theta)^{\Delta_{1i}\Delta_{2i}} \frac{\partial S(Y_{1i}, Y_{2i}; \theta)^{\Delta_{1i}(1-\Delta_{2i})}}{\partial Y_{1i}} \frac{\partial S(Y_{1i}, Y_{2i}; \theta)^{(1-\Delta_{1i})\Delta_{2i}}}{\partial Y_{2i}} \times \quad (A.1)$$
$$S(Y_{1i}, Y_{2i}; \theta)^{(1-\Delta_{1i})(1-\Delta_{2i})}.$$

Two-stage semiparametric estimation is performed if one chooses a nonparametric form for the marginal distributions and a parametric form for the bivariate distribution. For the two-stage semiparametric estimation the likelihood of α is

$$\begin{aligned} L(\alpha, u_j, v_j) &= \Pi c_\alpha(u_i, v_i)^{\Delta_{1i}\Delta_{2i}} \frac{\partial C(u_i, v_i)^{\Delta_{1i}(1-\Delta_{2i})}}{\partial u_i} \frac{\partial C(u_i, v_i)^{(1-\Delta_{1i})\Delta_{2i}}}{\partial v_i} \times \quad (A.2) \\ & \quad C(u_i, v_i)^{(1-\Delta_{1i})(1-\Delta_{2i})}, \end{aligned}$$

where $u_i = S_1(Y_{1i})$ and $v_i = S_2(Y_{2i})$ are the Kaplan-Meier estimates. The score function is the derivative of the log likelihood (A.2) wrt α:

$$U_\alpha(\alpha, \widehat{S}_1(y_{1i}), \widehat{S}_2(y_{2i})) = \sum_i \frac{\partial l(\alpha, \widehat{S}_1(y_{1i}), \widehat{S}_2(y_{2i}))}{\partial \alpha}. \quad (A.3)$$

The solution of this estimating equation obtained by setting A.3 equal to 0 is an estimate of α. The Newton-Raphson estimator can be used to solve for α:

$$\widetilde{\alpha} = \widehat{\alpha} + \left(\sum_i U_{j\alpha}^T U_{j\alpha} \right)^{-1} U_\alpha(\alpha, \widehat{S}_1(y_{1i}), \widehat{S}_2(y_{2i})).$$

According to Shih *et al.*, if $\widehat{S}_1$ and $\widehat{S}_2$ are consistent estimates of S_1 and S_2 then $\sqrt{n}(\widetilde{\alpha} - \alpha) \xrightarrow{d} N(0, \tau^2)$, implying that $\widetilde{\alpha}$ is asymptotically normal. This two-stage approach is a pseudo-likelihood approach. A loss of efficiency occurs when nonparametric survival functions are chosen for the margins.

Bivariate Extension For Missing Covariate Information

Since the marginal estimates have to be consistent in Shih's two-stage method, the non-parametric survival functions can be estimated from the weighted Cox model. The weighted Cox model is

$$l_\beta^w = \sum_{i=1}^{n} \int_0^\infty \frac{R_i}{\pi_i} \left\{ \mathbf{Z}_i - \frac{\sum_{h=1}^{n} \frac{R_h}{\pi_h} I\left(Y_h \geq t\right) \mathbf{Z}_h \exp\left\{\boldsymbol{\beta}' \mathbf{Z}_h\right\}}{\sum_{h=1}^{n} \frac{R_h}{\pi_h} I\left(Y_h \geq t\right) \exp\left\{\boldsymbol{\beta}' \mathbf{Z}_h\right\}} \right\} dN_i(t) \text{ (Pugh } et \text{ } al., \text{ 1993).} \quad \text{(A.4)}$$

The modified marginal distributions to incorporate missing covariate information are $\widehat{S}_k(y) = \exp\left(-\widehat{\Lambda}_k(y)\right)$ where the cumulative hazards are defined by:

$$\widehat{\Lambda}_k(y) = \begin{cases} \sum_{i=1}^{n} I\left\{Y_{ik} \leq y\right\} \frac{\Delta_{ik}}{\sum_{j=1}^{n} I\left\{Y_{jk} \geq Y_{ik}\right\} \frac{R_{jk}}{\pi_{jk}} e^{\beta_k Z_{jk}}}, \text{ if } \max(Y_{ik|R=0}) < \max(Y_{ik}) \\ \sum_{i=1}^{n^*} I\left\{Y_{ik} \leq y\right\} \frac{\Delta_{ik}}{\sum_{j=1}^{n} I\left\{Y_{jk} \geq Y_{ik}\right\} \frac{R_{jk}}{\pi_{jk}} e^{\beta_k Z_{jk}}}, \text{ if } \max(Y_{ik|R=0}) \geq \max(Y_{ik}) \end{cases}, \quad \text{(A.5)}$$

where $n^* = n - I\left\{(Y_{ik|R=0}) \geq \max(Y_{ik|R=1})\right\}$.

The simulations for the logistic model indicated that the weighted score may perform as well as the efficient score approach. In this case, the interest is in solving for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ from the two models $f(y_1|x)$ and $f(y_2|x)$. If the efficient score approach were used instead, it would be necessary to define the complete and observed data. Conditioning on the observed data would be necessary, and this is extremely complex to calculate since the complete data is (Y_1, Y_2, X, V) where (Y_1, Y_2, V) is observed. We propose obtaining marginal parameter estimates from the weighted pseudo-likelihood (A.4) and solving Equation A.5 for each k. The survival margins are then treated as fixed while obtaining an estimate for the association parameter.

LONGITUDINAL DATA

Since the advent of repeated measures methods, mainly generalized estimating equations, studies are often designed to collect repeated measures to determine changes over time, increasing the chances of data not being reported. It is conceivable that data is missing with the reason unknown, also known as missing by happenstance. We propose an extension of a method developed by Wei and Stram (1988) to handle data missing at random. The proposed method is a two-stage quasi-likelihood approach that employs estimating the parameters from the marginal estimating equations at each time point and the variance-covariance from the sandwich estimator. Our focus will be on two types of outcomes, continuous and binary, narrowing the applications to linear regression and logistic regression. This section will focus on the latter.

Data collected over a period of time are defined as repeated measures and longitudinal data. Standard regression procedures assume observations are independent. However, since measurements are obtained at multiple time points, the repeated measurements within each patient are no longer independent. This within-patient correlation must be taken into account during analysis. The assumption of independence between patients is still valid. Data of interest is repeated binary outcome data, baseline covariates, and time-varying covariates with ignorable nonresponse. This means that subjects can miss a visit at any time with the exception that baseline covariates must be fully observed for the nonresponse to be ignorable. We propose a quasi-likelihood approach that employs estimating the parameters from marginal estimating equations at each time point and the variance-covariance from the sandwich estimator.

Methodology For Repeated Measures And Missing Data

Various methods have spun off of quasi-likelihood theory that are in the class of estimating functions. McCullagh and Nelder (1983) introduced quasi-likelihood theory as an alternative to GLM. Since the GLM is a likelihood-based specification of the full likelihood, quasi-likelihood is an estimation tool of choice when one is uncertain about the mechanism for generating data or when there is insufficient data to specify an accurate likelihood. In a missing data framework, one is rarely 100% certain of the data generation mechanism, and once data is incomplete, model specification becomes unstable. This approach reduces the assumptions needed for model specification since only the first two moments are required to be specified for the quasi-likelihood. McCullagh *et al.* naturally developed a quasi-likelihood methodology for independent observations. Recognizing the need for methods that can accommodate dependent observations, McCullagh *et al.* extended the quasi-likelihood to handle such data.

Zeger and Liang developed a generalized estimating approach (GEE) for longitudinal data in 1986. The GEE is an extension of the GLM implementing the multivariate quasi-score approach. Dependent upon the hypothesis, there are three main types of models to choose from: marginal, random effects, and transitional models. Marginal models are concerned with the population average and model the regression of the outcome on the covariates and dependence structure separately. Random effects models allow the coefficient to vary by subject taking into account heterogeneity for latent variables. Transitional models include prior outcome measurements in the model, assuming that prior outcomes will predict the current outcome. Only the first two moments, the mean and variance, are required to be specified for GEE. Zeger and Liang have shown that the GEE is robust to specification of

the within subject correlation matrix. The GEE is not unbiased given MAR data. For complete data, $\boldsymbol{\beta}$ is estimated from the following equation

$$S_\beta(\beta, \alpha) = \sum_{i=1}^{n} \left(\frac{\partial \mathbf{u}_i}{\partial \boldsymbol{\beta}} \right)' Var(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mathbf{u}_i) = 0.$$

Lang (2000) proposed a parametric approach based on copulas for modelling repeated measures outcome data. He developed the copula methodology for normally distributed and binary outcome data. Based on simulation results, Lang concluded that the copula based model produced marginal parameter estimates which are robust to marginal distributional misspecification and the copula model chosen. Lang also extended this method to handle nonignorable missing outcomes since the copula approach is a more powerful inferential tool, due to the fact that one can apply likelihood-based methods.

Lipsitz *et al.* (1999) developed a likelihood-based method incorporating the EM algorithm for a nonignorable response problem. This likelihood approach forces one to specify the full likelihood and a distribution for covariates. As has been previously mentioned, specification of the distribution for covariates can be complex due to the high dimension of the covariates. An advantage of this approach is the flexibility of the missing data patterns. Any of the time-varying variables and baseline covariates are permitted to be missing.

Wei and Stram (1988) chose a two-stage approach to estimate the coefficient parameters and their covariance matrix. At the first stage, the coefficients are estimated from the quasi-score equation at each time point independently. The variance is estimated at the second stage. An advantage of the quasi-score is that no parametric model is specified for the data and only the first two moments are specified. A parametric model is not specified for the covariance matrix. The joint estimation of $\left(\widehat{\beta} - \beta\right)$ is shown to be multivariately normal,

and the covariance is estimated by the Information-sandwich (Huber-White) estimator. Wei *et al.* extended this method to accommodate MCAR data and time-dependent covariates. Stram, Wei, and Ware (1988) specifically evaluated properties of their estimating approach for ordinal data under MCAR. Since each coefficient is estimated at every time point, a linear combination of these estimates can give one a population average of β.

Carlin (1999) reviews weighted estimating equation approaches, known as weighted GEE (Flanders and Greenland, 1991; Clayton *et al.*, 1998). The goal of weighted estimating equations is to weight each observation by an assigned/estimated inverted probability of selection. Only complete cases are included in analysis. Although these unbiased equations lead to consistent estimates, they are not the most efficient. Robins *et al.* (1995, 1997) further modified and developed these inverse probability weighted estimating equation methods by drawing more information from the incomplete cases. Robins *et al.* developed these methods to yield more efficient estimates for ignorable and nonignorable missing data.

Two-Stage Approach

The assumptions needed to specify a quasi-likelihood are the mean and variance of the response. The form of the quasi-likelihood is

$$\begin{aligned} \mathbf{Q}(\mathbf{u};\mathbf{y}) &= \sum_{i=1}^{n} \int_{y}^{u} \frac{y_i - t}{\sigma^2 V(t)} dt \\ E(Y) &= u \\ Var(Y) &= \sigma^2 V(u). \end{aligned}$$

For the binomial case:

- $u_{it} = \frac{\exp^{Z\beta}}{1+\exp^{Z\beta}}$ is the mean of Y and $logit(u_{it}) = \beta_0 + \beta_1 z_{1it} + .. + \beta_p z_{pit}$ is the link function

- $Var(Y_{it}) = u_{it}(1 - u_{it})$ is the variance of Y.

The quasi-score function is defined as

$$\mathbf{U}(\boldsymbol{\beta})=\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{u}) = 0,$$

where $D = \partial u/\partial\beta$ and $V = cov(y)$. Wei solves the score equation at each time point independently, indicating at each time point $u = \frac{\exp^{xB}}{1+\exp^{xB}}$, $\partial u/\partial\beta = \mathbf{Zu}(\mathbf{1} - \mathbf{u})$, $V(u) = \mathbf{u}(\mathbf{1} - \mathbf{u})$, and $\mathbf{D}^T\mathbf{V}^{-1} = \mathbf{Z}$. The quasi-score equation at each time point reduces to $R\mathbf{Z}(\mathbf{Y} - \mathbf{u}) = 0$, where R is the missing indicator defined in Section 2.4. The estimates of β are used to solve the variance-covariance. The variance-covariance chosen is the robust estimator constructed from the information-sandwich (Huber-White) estimate of the variance.

Two-Stage Extension

We propose to extend the Wei *et al.* (1988) two-stage approach to accommodate MAR data. It is a well known fact that under MAR the complete quasi-score is not unbiased. In an effort to find an unbiased estimating equation as was indicated in Chapter 2 the weighted pseudo-likelihood (2.12) can fulfill this criteria. Although these equations are not the optimal unbiased estimating equation, they can serve as a simple estimating equation and building block for estimators with desirable properties. At each time point, k, Equation 2.12 can be expressed as a quasi-score function letting $\mathbf{D}_i^T\mathbf{V}_i^{-1} = \mathbf{Z}_i W_i R_i$, where $W_i = \pi_i^{-1}$. Once the quasi-score function is specified, one can solve for a consistent and asymptotically normal estimator of $\boldsymbol{\beta}_k$. The variance-covariance matrix of $(\beta_1, ..., \beta_k)$ is constructed in the

following fashion. The variance-covariance matrix of $(\beta_1, ..., \beta_k)$ is

$$G = n^{-1} \begin{bmatrix} D_{11}(\beta_1, \beta_1) & D_{12}(\beta_1, \beta_2) & D_{1k}(\beta_1, \beta_k) \\ & & \\ & & \\ D_{k1}(\beta_k, \beta_{1l}) & D_{k2}(\beta_k, \beta_2) & D_{kk}(\beta_k, \beta_k) \end{bmatrix},$$

where

$$D_{kl}(\beta_k, \beta_l) = A_k^{-1}(\beta_k) C_{kl}(\beta_k, \beta_l) A_l^{-1}(\beta_k).$$

The components of $D_{kl}(\beta_k, \beta_l)$ are defined as

$$A_k(\beta_k) = n^{-1} \sum_{i=1}^{n} l^{wT}_{\boldsymbol{\beta}_k, i} l^{w}_{\boldsymbol{\beta}_k, i}$$

and

$$C_{kl}(\beta_k, \beta_l) = n^{-1} \sum_{i=1}^{n} l^{wT}_{\boldsymbol{\beta}_k, i} l^{w}_{\boldsymbol{\beta}_l, i}.$$

According to Wei and Stram (1988) and Fahrmeir and Kaufman (1985), G is a consistent estimator of the true covariance matrix. Since $A_k \to c$ as $n \to \infty$ and by the multivariate central limit theorem, $(\beta_1, ..., \beta_k)$ is asymptotically normal, $\sqrt{n}\left(\widehat{\beta}_1 - \beta_1, ..., \widehat{\beta}_k - \beta_k\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{G})$ (Wei and Stram, 1988). An average coefficient can be estimated by $\boldsymbol{\beta} = \sum_{k=1}^{K} c_k \boldsymbol{\beta}_k$, where $c = (c_1, .., c_k) = (e' G^{-1} e)^{-1} G^{-1} e$, $e = (1, .., 1)'$, and var(β)=$(e' G^{-1} e)^{-1}$ (Wei and Johnson, 1985).

The MAR assumption defined by Mark and Gail (1994) will be applied to our data. The complexity of defining and modelling the missing data mechanism can be a barrier. The outcomes and other time-varying covariates are permitted to be missing. According to our definition of MAR, the missingness is dependent on the baseline data, outcome, and time-varying covariate data from prior time points. The data up to time point $t-1$ for (Y, V)

and R is defined as $\overline{H}_{ti} = (Y_1, ..., Y_{t-1}, V_1, .., V_{t-1})$ and $\overline{R}_{ti} = (R_1, ..., R_{t-1})$. The missing data mechanism is defined as

$$\begin{aligned} P(R_{1i}|X_i, V_{1i}, Y_{1i}) &= P(R_{1i}|X_i) \\ P\left(R_{ti}|X_i, (\overline{R}_{ti} \times \overline{H}_{ti}, \overline{R}_{ti} = 0), (\overline{R}_{ti} \times \overline{H}_{ti}, \overline{R}_{ti} = 1)\right) &= P(R_{ti}|X_i, (\overline{R}_{ti} \times \overline{H}_{ti}, \overline{R}_{ti} = 1), \overline{R}_{ti} = 0) \\ \text{for } t &> 1. \end{aligned}$$

BIBLIOGRAPHY

[1] Bickel PJ, Klaasen CA, Ritoz Y, Wellner JA (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press.

[2] Binder DA (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* 79(1): 139-147.

[3] Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* 6:39-58.

[4] Breen R (1996). *Regression Models Censored, Sample Selected, or Truncated Data*. California: Sage Publications.

[5] Breslow NE, Cain KC (1988). Logistic regression for two-stage case-control data. *Biometrika* 75(1): 11-20.

[6] Breslow NE, Chatterjee N (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Applied Statistics* 48(4): 457-468.

[7] Breslow NE, McNeney B, Wellner JA (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. The Annals of Statistics 31:1110-1139.

[8] Carlin JB, Wolfe R, Coffey C, Patton GC (1999). Tutorial in biostatistics analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: prevalence and incidence of smoking in an adolescent cohort. *Statistics in Medicine* 18:2655-2679.

[9] Chen HY, Little R (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 94(447):896-908.

[10] Clayton DG (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65:141-151.

[11] Clayton D, Cuzick J (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)* 48:82-117.

[12] Clayton D, Spiegelhalter D, Dunn G, Pickles A (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B* 60:71-87.

[13] Cox DR (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2):187-220.

[14] Cox DR (1975). Partial likelihood. *Biometrika* 62:269-276.

[15] Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39:1-38.

[16] Fahrmeir L, Kaufmann H (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* 13:342-368.

[17] Fears TR, Brown CC (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* 42(4):955-960.

[18] Flanders WD, Greenland S (1991). Analytical methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* 10:739-747.

[19] Frank MJ (1979). On the simultaneous associativity of $F(x, y)$ and$x + y - F(x, y)$. *Aequationes Math* 19:194-226.

[20] Genest C, Ghoudi K, Rivest LP (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* 82:543-552.

[21] Godambe VP (1991). *Estimating Functions*. New York: Oxford Press.

[22] Gong G, Samaniego FJ (1981). Pseudo maximum likelihood estimation: Theory and Applications. *The Annals of Statistics* 9:861-869.

[23] Henery RJ (1981). An approximation to certain multivariate normal probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)* 43:81-85.

[24] Herring A, Ibrahim JG (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association* 96(453):292-302.

[25] Horvitz DG, Thompson DJ (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260):663-685.

[26] Hougaard P (1986). A class of multivariate failure time distributions. *Biometrika* 73:671-678.

[27] Hsieh DA, Manski CF, McFadden D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association* 80(391):651-662.

[28] Hsu L, Prentice RL (1996). On assessing the strength of dependency between failure time variates. *Biometrika* 83:491-506.

[29] Joe H (1997). *Multivariate Models and Dependence Concepts*. London, UK: Chapman & Hall.

[30] Kendall M, Gibbons JD (1990). *Rank Correlation Methods fifth edition* (3rd edition is 1962). NY: Oxford University Press

[31] Kish L, Frankel MR (1974). Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)* 36(1):1-37.

[32] Keiding N, Klein JP and Goel PK (eds) (1992). *Survival Analysis: State of the Art*, pp. 309-326. Netherlands: Kluwer Academic Publishers.

[33] Lang W (2000). Applications of Copulas to Repeated Measures Data. Ph.D. Thesis, University of Pittsburgh Biostatistics.

[34] Lawless JF, Kalbfleisch JD, Wild CJ (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 61:413-438.

[35] Liang K-Y, Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73:13-22.

[36] Lin, DY, Ying Z (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* 88(424):1341-1349.

[37] Lipsitz SR, Ibrahim JG, Fitzmaurice GM (1999) Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates. *Biometrics* 55:214-223.

[38] Little, Rubin (2002). *Statistical Analysis with Missing Data.*

[39] McCullagh P, Nelder JA (1984). *Generalized Linear Models.* London: Chapman and Hall.

[40] McLaren CE, Brittenham GM, Hasselblad V (1986). Analysis of the volume of red blood cells: application of the expectation- maximization algorithm to grouped data from the doubly-truncated lognormal distribution. *Biometrics* 42:143-158.

[41] McLaren CE, Wagstaff M, Brittenham GM, Jacobs A (1991). Detection of two-component mixtures of lognormal distributions in grouped, doubly truncated data: analysis of red blood cell volume distributions. *Biometrics* 47:607-622.

[42] Manski CF, Lerman SR (1977). The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society* 45(8):1977-1988.

[43] Mark SD, Gail MH (1994). A comparison of likelihood-based and marginal estimating equation methods for analysing repeated ordered categorical responses with missing data: application to an intervention trial of vitamin prophylaxis for oesophageal dysplasia. *Statistics in Medicine* 13:470-493.

[44] Marshall AW, Olkin I (1988). Families of multivariate distributions. *Journal of the American Statistical Association* 83:834-841.

[45] Nan B, Emond M, and Wellner JA (2002). Information bounds for Cox regression models with missing data. (Revision of Technical Report 378.) Annals of Statistics 32, 2004, 723-753.

[46] Nelsen RB (1999). *An Introduction to Copulas*. New York: Springer-Verlag.

[47] Oakes D (1982). A concordance test for independence in the presence of censoring. *Biometrics* 38:451-455.

[48] Oakes D 1986. Semiparametric Inference in a Model for Association in Bivariate Survival Data. *Biometrika* 73:353-361.

[49] Oakes D (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 84:487-493.

[50] Phelps AJ, Weissfeld LA (1997). A comparison of dependence estimators in bivariate copula models. *Commun Statistica-SimulaA* 26:153-1597.

[51] Prentice RL (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69(2):331-342.

[52] Prentice RL (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73(1):1-11.

[53] Pugh M, Robins JM, Lipsitz SR, Harrington DP (1993). Inference in the Cox proportional hazards model with missing covariate data. *Technical Report*, Harvard School of Public Health, Department of Biostatistics.

[54] Robins JM, Rotnitzky A, Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89:846-866.

[55] Robins JM, Hsieh F, Newey W (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:409-424.

[56] Robins JM, Rotnitzky A, Zhao LP (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90:106-121.

[57] Rotnitzky A, Robins JM (1997). Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine*, 16:81-102.

[58] Rubin DB (1976). Inference and missing data. *Biometrika* 63(3):581-592.

[59] Schafer J:L (1997). *Analysis of incomplete multivariate data.* NY: Chapman and Hall/CRC.

[60] Schweizer B, Wolff EF (1981). On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics* 9:879-885.

[61] Shih JH, Louis TA (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* 51:1384-1399.

[62] Sklar A (1959). Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Statistic Univ. Paris* 8:229-231.

[63] Stram DO, Wei LJ, Ware JH (1988). Analysis or repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates.*Journal of the American Statistical Association* 83:631-637.

[64] Tobin J (1958). Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society* 26:24-36.

[65] Wang W, Wells M (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association* 95:62-76.

[66] Wei LJ and Stram DO (1988). Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Statistics in Medicine* 7:139-148.

[67] Wei LJ, Johnson WE (1985). Combining dependent tests with incomplete repeated measures. *Biometrika* 72(2):359-364.

[68] Wild CJ (1991). Fitting prospective regression models to case-control data. *Biometrika* 78(4):705-717.

[69] Zhang Z, Rockette HE (2004). On maximum likelihood estimation in parametric regression with missing covariates. Journal of Statistical Planning and Inference, In Press, Corrected Proof, Available online 5 August 2004.