8-2010

An investigation of sliced inverse regression with censored data.

Daniel W. Riggs
University of Louisville

Recommended Citation
Riggs, Daniel W., "An investigation of sliced inverse regression with censored data." (2010). *Electronic Theses and Dissertations.* Paper 1206.
https://doi.org/10.18297/etd/1206

AN INVESTIGATION OF SLICED INVERSE REGRESSION WITH CENSORED DATA

By

Daniel W. Riggs
B.A. Mathematics, Keuka College, 2002

A Thesis
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2010

AN INVESTIGATION OF SLICED INVERSE REGRESSION WITH CENSORED DATA

By

Daniel W Riggs

B.A., Keuka College, 2002

A Thesis Approved on

August 6, 2010

by the following Thesis Committee:

____

Dr. Somnath Datta (Mentor)

____

Dr. Guy Brock

____

Dr. Aruni Bhatnagar

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Somnath Datta, for his guidance and dedication to me as a student. I would also like to thank the other committee members, Dr. Guy Brock and Dr. Aruni Bhatnagar for their comments and assistance.

ABSTRACT

An Investigation of Sliced Inverse Regression with Censored Data

Daniel Riggs

August,6 2010

The complexity of high-dimensional data creates a number of concerns when trying to analyze it. This data often consists of a response or survival time and potentially thousands of predictors. These predictors can be highly correlated, and the sample size is often very small and right censored. Sliced inverse regression(SIR) is a method of reducing the dimension of the data, while preserving all the regression information. Sliced inverse regression with regularizations was developed to work when the number of predictors exceeds the sample size, and to deal with highly correlated predictors as well.

In this study we investigated the performance of Sliced inverse regression with regularizations using three different approaches for handling right censored data. The methods of reweighting, mean imputation, and multiple imputation were analyzed. Based on the simulation scenarios, the mean imputation method performs the best in regards to fitting the data as well as prediction. The method of reweighting appears inadequate when combined with SIR.

TABLE OF CONTENTS

PAGE

CHAPTER

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

A. Censored Survival Data

The need to analyze survival data arises in a number of fields, such as biology, medicine, and epidemiology. This type of analysis is focused on data that are generated from the time to a specific event. It is frequently used in cancer studies to measure the time to recurrence or death. Survival data often contains censored observations which occur when the time to an event is only known to have happened during a certain period of time. Data can be censored through various schemes, such as left, interval, and right censoring. Left censoring involves a survival time only known to have occurred before a specific time point. For example, there is a specific level of sensitivity for air pollution monitoring sites. If the level of air pollution falls below this threshold, then the actual value is known only to be less than the sensitivity level. Interval censoring is associated with a survival time occurring at a point between two known time periods. If, for example, you want to analyze the number of days until the recurrence of cancer in a subject that only has monthly checkups, then you may only determine the recurrence as occurring somewhere between the last two monthly checkups. Right censoring occurs when an event is only known to have happened after a certain time point. In human studies, it is common for subjects to drop out for numerous reasons, whether they move away from the study, or simply lose interest. These subjects would be considered right censored. Any combination of these three censoring schemes can occur in a study. It is not uncommon for a study involving left censoring to also contain right censoring. In such a case the lifetimes are considered doubly censored. Right censored survival data will be the focus of this thesis.

Right censoring includes a number of subcategories. Type I right censoring occurs when an event is observed only if it happens before a pre-specified time point. Any individuals that do not have the event observed before this pre-specified time point are considered censored. This typically occurs in animal studies or clinical research due to time and cost concerns of letting an experiment continue indefinitely until all the subjects have failed. In Type II censoring, a study ends when a pre-set number or percentage of subjects have failed. Anything that does not fail is considered right censored. This design is typically used for testing equipment failure. Finally, random censoring, is an example of competing risks censoring. This occurs with an interest in the time to a certain event, but some individuals experience a competing event, leading them to be dropped from the study. Thus, the event of interest is unobservable, and the subject is considered to be random right censored. As with the more general categories of censoring, it is common for a study to contain more than one type of right censoring. In a Type I study design patients could also leave the study, causing both Type I and random censoring (1).

The time to an event of interest is typically assumed to be independent of the censoring time. A right-censoring mechanism is independent if the failure rates of an individual at time t is the same as it would have been without censoring (2). When this independence assumption fails, the basic methods of survival analysis are inapplicable and special techniques must therefore be used (2).

The intent of survival analysis is to make valid inferences on survival times. While right censored data complicates this goal, several methods have been developed to modify this problem. The focus of this thesis will be on the methods proposed by Datta, *et al.* (3) of reweighting, mean imputation and multiple imputation, all of which keep the mean response the same. These were developed, in part, because most methods for linear regression require a full data model. Simply deleting censored data could lead to a loss of power because of the reduced sample size and could also introduce bias if the remaining sample is not representative of the entire population. Imputation methods impute an actual time into the censored time, while reweighting changes the unobserved times to zero and reweighs the actual times. Once done, normal linear regression methods can be performed on the modified data set.

B. Right Censored Data in Microarray

There is an increasing demand for methods of analyzing high-dimensional data. This is especially true in the fields of biomedical and genomic research. With DNA microarray becoming increasingly popular, sophisticated statistical methods are needed to analyze the data yielded from this technology. DNA Microarray consists of an arrayed series of thousands of DNA oligonucleotides or DNA fragments called features (4). This is a short section of either a gene or DNA element, which is then used as a probe to hybridize DNA or RNA from a sample. The hybridization can be quantified fluorescently using fluoro-based samples to determine the relative abundance of nucleic acid sequences in the sample (4). Research has shown that gene expression profiles from microarray can be used to predict various clinical phenotypes, such as tumor class, drug response and survival time. For example, gene expression profiling could be used to identify genes that change expression based on a given treatment.

Gene expression is a process in which information from a gene is used in the synthesis of a functional product, protein or RNA. It is used in all forms of life. Several steps of gene expression can be regulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. The regulation of gene expression refers to the amount of functional product in a gene. It also pertains to the appearance of this functional product. Control of this expression allows a cell to produce the gene products it needs, which in turn allows the cell to adapt to different situations such as environment, external signals and cell death. Gene regulation is also the basis for both cellular differentiation and morphogenesis, and enables the cell to control its structure and function. Gene expression is the most fundamental level at which a genotype gives rise to a phenotype. Thus, it is both natural and beneficial to use gene expression as a predictor for survival.

In this thesis, survival time will be considered as the response variable. In one study, Van De Vijver, *et al.* found gene expression data to be a more powerful predictor of breast cancer survival than existing clinical methods (5). Shedden, *et al.* used gene expression to predict survival in patients with lung adenocarcinoma (6). Most recently, Steidl, *et al.* found that an increased number of tumor-associated macrophages was strongly associated with a shortened survival in patients with classic Hodgkin's lymphoma (7). There are many benefits of predicting survival time based on gene expression data. With a better indication of predicted survival, these findings can lead to better care and treatment of patients. This also results in a better understanding of a specific disease and how it causes death, thus providing ideas for future research and understanding, as well as life saving cures.

However, the complexity of this data creates a number of concerns when trying to model it. Typically, the data consists of a response (survival time) and potentially thousands of predictors (genes). Often these genes can be highly correlated, and the typical sample size can be very small compared to the number of predictors. Further, when using microarray to predict survival time, it is possible that much of the data will be right censored. It is not uncommon for Type I censoring to occur, meaning there is a preset follow-up time in which an individual's lifetimes may be unknown.

There are several methods in survival analysis that can manage this type of censoring, such as the accelerated failure time model and the Cox proportional hazard model. However, to use these methods, some kind of functional form must be specified, which can be increasingly complicated as the number of predictors increases. Standard linear regression requires the sample size to be larger than the number of predictors and is thus inapplicable in this situation. It is often beneficial to do some exploratory data analysis before any modeling is done. Therefore, one can consider dimension reduction prior to model building. Sufficient dimension reduction, developed by Cook, is a way of reducing the predictor dimension while still preserving all regression information of the data set without specifying a parametric model. Sliced inverse regression (8) is one of the most commonly used methods in this area; however, SIR cannot work when the number of predictors is greater than the sample size. And, when a high collinearity among predictors is present, which is often the case in genomic data, SIR also suffers.

To combat this, sliced inverse regression with regularizations was developed (9). Ridge regression with the L_2 regularization was introduced to allow SIR to work with highly correlated data, as well as sample sizes less than the number of predictors. Using the least absolute shrinkage and selection operator (Lasso) idea, the L_1 regularization was further introduced to achieve predictor selection.

C. Purpose of Study

The most common method of handling missing data is complete case analysis, in which all subjects with missing data are removed from the analysis. Almost all Sufficient Dimension Reduction methods, including normal SIR, employ this method when there is missingness or censoring to the data. While few methods have been developed that apply to right censored survival data, Li, *et. al.* used a modified version of SIR to deal with censoring (10). More recently Wen and Cook developed model-free dimension reduction for bivariate regression that can be applied to censored data (11). An alternative to these bivariate dimension reduction methods is the use of imputation or reweighting methods to handle the censoring. Li and Lu found that reweighting using sliced inverse regression gave better results than complete case analysis (12). Yet, it is currently unclear how reweighting and imputation methods perform with SIR with regularizations for right censored data. The main focus of this study is to compare three different proposals for managing right censored survival data, when used with sliced inverse regression with regularizations. The three methods, reweighting, mean imputation and multiple imputation, will be compared to bivariate SIR method for handling censored data as well as uncensored data to test their effectiveness. Using the accelerated failure time model, the overall fit of the data, as well as the accuracy of prediction, will be examined.

Simulation studies will be performed using a randomly-generated training data set to measure the overall fit of the model. Finally, a new testing data set will be generated to measure the prediction accuracy of each method.

CHAPTER II

METHODS FOR RIGHT CENSORED DATA

A. Background

For right censored survival data, let T_i = observed failure time and C_i = fixed right censored time. Then, $X_i = \min(T_i , C_i)$, where i=1,2,...,N is the sample size. Thus, the time, $X_i = T_i$ if the failure is observed, and $X_i = C_i$ if it is censored. Let $\delta_i = 1$ if X_i corresponds to a failure and $\delta_i = 0$ if X_i is censored. For the purpose of this study, T_i and C_i are independent. Thus, the likelihood function can be constructed as follows:

$L = \Pi\ P[x_i, \delta_i] = \Pi\ [f(x_i)]^{\delta i}\ [S(x_i)]^{1-\delta i}$,

where $f(x_i) = P(T=x_i)$, and $S(x_i) = P(X>x_i)$.

To estimate S(x) the Kaplan-Meier or Product-Limit estimator is used. Let $x_1<x_2<...<x_n$ be distinct ordered failure times. The Kaplan-Meier Survival estimate is defined as follows:

$S(x_i)= \Pi\ [1 - (d_i/Y_i)]$ for $i|x_i<x$,

where d_i= number of failures at time x_i, and Y_i= number at risk at time x_i.

The variance is estimated by Greenwood's formula:

$V[S(x)]= S(x)^2\sum d_i/(Y_i(Y_i-d_i))$.

The mean time to event, which will be needed for the imputation methods, is estimated by $\mu=\int_1^n S(x)\ dx$.

However, this estimate is only appropriate when the largest observation corresponds to a failure as the Survival estimate is not defined beyond the largest failure. For the imputation methods, if the largest time is censored it is changed to a failure time.

B. Accelerated Failure-Time Model

Parametric models can provide more accurate estimates and prediction than their semi-parametric counterparts because they rely on fewer parameters. However, if the model is not chosen correctly it can lead to consistently wrong predictions. The accelerated failure-time model is a fully parametric linear model representation of the logarithm of the true survival time. It is a useful alternative to the more commonly used semi-parametric Cox proportional hazard model. The Cox model assumes a multiplicative effect on the hazard function, while the AFT model assumes a multiplicative effect on survival times.

The survival time for the i^{th} individual is

$Y_i = \ln(X_i) = \mu + \beta^t \mathbf{Z_i} + \sigma W_i$ for $i=1,\ldots,n$.

Where X_i is the survival time, $\mathbf{Z_i}$ is the covariate vector corresponding to the i^{th} individual, β is a vector of regression coefficients, μ is the intercept, σ is a scale parameter, and W_i is the error distribution. In terms of the survival function this equation can be written as

$S(x|\mathbf{Z_i})= S_o[\exp(\theta^t \mathbf{Z_i})x]$ for all x

where $\theta= -\beta$.

The likelihood function for the AFT model for right censored data is

$L = \Pi\ [(1/\sigma)f_o(y_i - \mu - \beta^t \mathbf{Z_i})/ \sigma)]^{\delta i}\ [S_o(y_i - \mu - \beta^t \mathbf{Z_i})/ \sigma)]^{1-\delta i}$

The estimates for μ, β and σ are found by maximizing the likelihood function of the AFT model (13). A number of distributions are commonly used for W in the AFT model, the most notable being the Weibull, log logistic , exponential, and log normal distributions. In this study, the focus will be on the Weibull and the log normal distributions.

The survival function for the Weibull distribution is given by

$S_X(x) = \exp(-\lambda x^{\alpha})$

where $\alpha>0$ is a shape parameter, and $\lambda>0$ is a scale parameter.

Incorporating covariates, the hazard rate for the Weibull is:

$h(x|\mathbf{Z}) = (\lambda\alpha x^{\alpha-1})\exp(\boldsymbol{\beta}^t\mathbf{Z})$.

Letting Y=lnX, the survival function for the log transform is:

$S_Y(y)=\exp(-\lambda\exp(\alpha y)) \exp(\boldsymbol{\beta}^t\mathbf{Z})$.

By letting $\lambda= \exp(-\mu/\sigma)$ and $\sigma = 1/\alpha$, then Y has the log linear model form of

$Y_i = \ln(X_i) = \mu + \beta^t\mathbf{Z}_i + \sigma W_i$ for $i=1,\ldots,n$

where W is the extreme value distribution.

The survival function for the log normal distribution is given by

$S(x)= 1- \Phi\{[\log(x)- (\mu+ \boldsymbol{\beta}^t\mathbf{Z})]/\sigma\}$,

where $\Phi\{\}$ is the cumulative distribution function of the standard normal distribution. All of these parametric methods can be implemented in R software using the survreg function from the survival package.

C. Reweighting

Reweighting, more commonly known as "inverse probability of censoring weighted" estimation (14), is a commonly used method for handling right censored data in the accelerated failure-time models, as well as the Cox proportional hazards model.

This scheme replaces the censored observations with 0, then reweighs the observed Y_i by the reciprocal of the probability that it is an actual failure time. Thus, we want to replace the observed Y_i, by $\bar{Y}_i$, where

$$\bar{Y}_i = \delta_i h(T_i)/S(T_i^-)$$

and h is the log transform in the AFT model, δ is the censoring indicator, $S(T_i)$ is the Kaplan-Meier estimate of the survival times, and - denotes a left limit. It has been shown that the mean response of $\bar{Y}_i$ is approximately equal to that of Y_i (15). Thus, the new response, $\bar{Y}_i$, is used in the AFT model.

D. Mean Imputation

Under mean imputation the observed Y_i are kept the same but the censored times are replaced by Y^*, their expected value given that the failure time T_i is larger than the censored time C_i. The survival can be estimated using the Kaplan-Meier curve as:

$$Y_i^* = \{S(C_i)\}^{-1} \sum \log(t_j) \Delta S(t_j) \quad \text{for } j=1,...,n$$

where $\Delta S(t_j)$ is the change in S at time t_j and $Y_i = \log(t_j)$. To use this scheme, Efron's tail correction is used, which simply modifies the largest event time to a true failure time. Thus at time t_n, $\delta_n=1$. Therefore under this scheme, $Y_i = Y_i$ if $\delta=1$, and $Y_i = Y_i^*$ if $\delta=0$. Then the accelerated failure time model can be fit with the new response variables.

E. Mulitple Imputation

Multiple imputation is a technique in which the censored observations are replaced by B different simulated versions from the conditional distribution of T, given that T>C. The mass points are estimated by

$$\Delta S(t_j) \;/\; S(C_i), \quad \text{where } t_j>C_j.$$

Let k=1,...,B, then the AFT model fit based on the kth set of imputed values is given by

$Y_k = \boldsymbol{X}\beta_{k.}$

The final answers are given by

$Y=B^{-1}\sum Y_k$ and

$\beta=B^{-1}\sum\beta_k$.

These averages approximate the conditional expectation given the observed data (3). Often times B is taken to be between 3 and 10. Ideally B should be taken as large as possible. In these simulations B =50.

CHAPTER III

SIR MODELS

A. Sufficient Dimension Reduction

When the number of predictors, p, is large statistical modeling often suffers from the curse of dimensionality. It is often beneficial to consider dimension reduction prior to model building. The overall goal of regression analysis is to understand how the conditional distribution of Y given X depends on the values assumed by X. Graphical displays of the data are often a good exploratory tool for the relationship between the response and predictors in regression studies. When the dimension d=1, a simple scatter plot of Y versus X provides information about the relationship of the data. When d=2 a rotating three-dimensional plot can provide information. Useful plots of three predictors can be created by replacing Y with a discrete $\tilde{Y}$, constructed by partitioning the range, then assigning predictors to the axis of a three-dimensional plot and marking the points that correspond to the values of $\tilde{Y}$ (16). However, when the dimension d>3, it is generally not possible to construct a comprehensive display of the data. Thus, in practice it is useful to reduce the dimensions to d=1, 2, or 3. Sufficient dimension reduction (17) has been developed to reduce the predictor dimension without loss of information on the response Y, given the predictors X with $X \in \mathbb{R}^p$. The ultimate goal is to find the smallest number of linear combinations of X, $\beta_1^TX,...,\beta_d^TX$, such that

$Y \perp\!\!\!\perp X \mid (\beta_1^T X,\ldots,\beta_d^T X)$,

where $\perp\!\!\!\perp$ indicates statistical independence and $d \leq p$. This implies that a p-dimensional predictor can be replaced by a d-dimensional $\beta^T X$ because given $\beta^T X$, X contains no additional information of Y. These linear combinations $(\beta_1^T X,\ldots,\beta_d^T X)$ are called sufficient predictors. The minimum number d of sufficient predictors is called the structural dimension of the regression. If regression has a structural dimension of d=0 then Y is independent of X. If d=1, then all the information of Y given X can be contained in a single linear combination, $\beta_1^T X$. The sufficient predictors are not unique, because we can multiply β by a nonzero constant and still have independence with Y. Thus, a linear subspace Span(β) is needed which is spanned by the columns of β. This span is called the dimension reduction subspace (17). It has been shown that the intersection of all the dimension reductions subspaces is itself a dimension reduction subspace under minor conditions (18). This intersection is a unique and parsimonious population parameter, which captures all regression of Y given X. It is called the central subspace, denoted by $S_{Y|X}$, and is the main object of interest in dimension reduction.

There are several methods designed to estimate sufficient predictors, the most notable being sliced inverse regression (SIR), sliced average variance estimation (SAVE) (19) and principal Hessian directions (PHD) (20). A modification of SIR will be the basis for this study.

B. Sliced Inverse Regression

The idea behind Sliced inverse regression is to replace the response Y with a discrete version $\tilde{Y}$, constructed by partitioning the range of Y into h slices, where h is a tuning parameter chosen to be h>d. SIR can find, at most, h-1 sufficient predictors. Thus h is generally chosen to be somewhat larger than d+1. Then regression is based on $\tilde{Y}$ given X. Thus if $\beta^T X$ is a sufficient predictor of $\tilde{Y}$ given X, then it is also a sufficient predictor for the regression of Y on X. Sliced inverse regression operates in the following way for data Y_i, X_i , where i = 1,...,n:

1. Standardize X to get $\tilde{x}$ $=\hat{\Sigma}^{-1/2}(x_i-\bar{x})$, where $\hat{\Sigma}$ is the sample covariance and $\bar{x}$ is the sample mean of x.

2. Divide the range of y into H slices, $I_1,...,I_h$. Let $\hat{p}_h = (1/n)\sum_{i=1}^{n} \delta_h(y_i)$, the proportion of the y_i that falls into slice h, where $\delta_h(y_i)=1$ if y_i falls into the h^{th} slice I_h and $\delta_h(y_i)=0$ if it does not.

3. Within each slice the sample mean of the $\tilde{x}_i$'s is computed, denoted by $\hat{m}_h$, where h= 1,...,H, so that $\hat{m}_h =(1/n\hat{p}_h)\sum_{y\in I} \tilde{x}_i$.

4. Perform a weighted principal component analysis for the data $\hat{m}_h$ (h= 1,...,H) in the following way: Form the weighted covariance matrix $\hat{V} = \sum_{h=1}^{H} \hat{p}_h\, \hat{m}_h\, \hat{m}_h'$, then find the eigenvalues and the eigenvectors for $\hat{V}$.

5. Let the K largest eigenvectors be $\hat{\eta}_k$ where k= 1, ...,K. Then, $\hat{\beta}_k= \hat{\eta}_k\widehat{\Sigma_{xx}^{-1/2}}$ for k=1,...,K.

C. SIR for survival data

The theory behind SIR does not require that Y be univariate, it holds equally for multivariate responses. One useful application for a bivariate response is in the area of survival analysis with censored data. Let T denote the survival time and C denote the censoring time, and let δ =0,1 be a binary indicator variable for the event C>T. Then, $Y_i=T_i\, \delta_i + C_i(1-\delta_i)$, i =1,...,n. Finally, let X be a p x 1 vector of predictors. Ideally we would like to estimate the central subspace of T given X, but T is not fully observable.

Thus we estimate the central subspace of the observable (Y,δ) on X. This is useful because of the condition:

$(T,C) \perp\!\!\!\perp X | B^T X$

This condition requires the sufficient predictors $B^T X$, for the regression of T on X also be sufficient predictors for the bivariate regression of (T,C) on X. This implies that censoring can depend on X but only by the sufficient predictors for the regression of T on X. The previous condition is equivalent to the pair of conditions

$T \perp\!\!\!\perp X | B^T X$ and $C \perp\!\!\!\perp X | (B^T X, T)$

The second condition states that C must be independent of X, given the sufficient predictors and the true survival time. From these conditions if follows that

$(Y,\delta) \perp\!\!\!\perp X | B^T X$

because (Y,δ) is a function of (T,C). This implies that

$\mathcal{S}_{(Y,\delta)|X} \subseteq \mathcal{S}_{T|X}$

This result demonstrates that the sufficient predictors for the observable regression (Y,δ)|X are also sufficient predictors for the regression T|X. The usual independence condition is

$T \perp\!\!\!\perp C | X$. If this condition is not met, then more information about the censoring

mechanism is needed for further analysis.

Bivariate SIR requires slicing on (Y,δ). This is done by partioning Y into h_o slices for the subsample when δ=0. Y must then be partitioned into h_l slices for the subsample when δ=1, for a total of H = h_0+h_1 slices (16). Then the normal SIR algorithm proceeds as described previously.

D. SIR with regularizations

Normal SIR estimation requires the inversion of the predictor covariance matrix Σ_x. In applications such as microarray studies, the number of predictors often exceeds the sample size. In these cases, the estimate of Σ_x is singular and noninvertible and thus normal SIR does not apply. In addition, the predictors may be highly correlated. This collinearity can produce a highly variable sample estimate when using SIR. Methods such as partial inverse regression (21) have been developed to address these problems; however, these methods do not focus on individual predictor selection. Sliced inverse regression with regularizations was developed for simultaneous dimension reduction when n<p as well as predictor selection. Using the least squares formulation of SIR, this regularized SIR method combines both L_1 and L_2 regularizations. The L_2 regularization enables SIR to work with n<p as well as highly correlated predictors. The L_1 regularization achieves reduction estimation as well as predictor selection. Too construct the least squares formulation of SIR suppose there are n independent and identically distributed realizations (X,Y). The sample version of Z is $\hat{Z}$= $\hat{\Sigma}_x^{-1/2}$(X-$\bar{X}$), where $\bar{X}$ is the grand average of X, and $\hat{\Sigma}_x$ is the sample covariance matrix. Suppose the range of the response Y is partitioned into h nonoverlapping slices, with n_y observations in the y^{th} slice, y=1,...,h . Let $\bar{Z}_y$ denote the average of $\hat{Z}$ in the y^{th} slice and $\hat{f}_y = n_y/n$. It has been shown that the normal SIR estimate can be obtained by minimizing

$G(B,C)=\sum_{y=1}^{h} \hat{f}_y \, ||\bar{Z}_y - BC||^2$,

over B ∈ $\mathbb{R}^{pxd}$ and C= (C_1,...,C_h) ∈ $\mathbb{R}^{dxh}$ (22).

Then, the solution $\hat{B}$ forms an estimation of the basis of $\mathcal{S}_{Y|X}$. SIR does not impose any model assumption on the conditional distribution of Y|X, but instead requires a condition on the marginal distribution of X. This condition is called the linearity condition and states that for any b∈ $\mathbb{R}^p$, $E(b^TX|\eta^TX)= c_0+c_1\eta_1{}^TX+...+c_d\eta_d{}^TX$, for some constants $c_0,...,c_d$, where $\eta=(\eta_1,...,\eta_d)$ forms a basis of $\mathcal{S}_{Y|X}$ (9). When X is elliptically symmetrically distributed, the linearity condition holds (23). Predictor transformation is commonly performed if the condition is not met (18).

The standardized predictor Z involves the inverse of Σ_x , and thus is not applicable when the sample covariance matrix $\hat{\Sigma}_x$ is singular. Hence, ridge regression with L_2 regularization is used to address this issue through the ordinary least squares setup. First, the least squares formulation is derived in the original X scale. Thus G(B,C) becomes

$$G(A,C)= \sum_{y=1}^{h} \hat{f}_y\{(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y\}^T \times \hat{\Sigma}_x^{-1} \{(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y\}$$

where $\bar{X}_y$ denotes the average of X in the y^{th} slice, and $A = \hat{\Sigma}_x^{-1/2} B$.

Then $\hat{A}$ is the value that minimizes G(A,C), and Span($\hat{A}$) estimates the central subspace $S_{Y|X}$. An equivalent form of G(A,C), which drops the $\hat{\Sigma}_x^{-1}$ term is

$$\tilde{G}(A,C)= \sum_{y=1}^{h} \hat{f}_y ||(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y ||^2$$

The equivalence requires the existence of $\hat{\Sigma}_x^{-1}$; however, by dropping the term it can be easily extended to incorporate the regularization parameter. Based on this, we get the following ridge SIR estimator:

$$G_\tau(A,C) = \sum_{y=1}^{h} \hat{f}_y ||(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x AC_y ||^2 + \tau\, vec(A)^T vec(A)$$

where τ is a nonnegative constant, and vec() is a matrix operator that stacks all columns of the matrix into a single vector (9). Let $(\hat{A}, \hat{C})= \text{argmin}_{A,C} G_\tau(A,C)$. Then Span($\hat{A}$) is called the ridge SIR estimator of the central subspace $S_{Y|X}$. When $\hat{\Sigma}_x^{-1}$ exists and $\tau = 0$, then $G_\tau(A,C)$ reduces to the usual SIR estimator. When $\hat{\Sigma}_x$ is not invertible, then a positive τ is incorporated to address the issue of singularity.

An alternating least squares algorithm as proposed by Li, is used to minimize $G_\tau(A,C)$ for a fixed τ. For a given A, C is obtained by h usual least squares:

$\hat{C} = (\hat{C}_1,\ldots,\hat{C}_h)$, with

$\hat{C}_y = (A^T\ \hat{\Sigma}_x{}^2 A)^{-1}\ A^T\hat{\Sigma}_x(\bar{X}_y - \bar{X})$,

for $y = 1,\ldots,h$.

Rewriting $G_\tau(A,C)$ in the least-squares regression form,

$$G_\tau(A,C) = \sum_{y=1}^{h} \hat{f}_y \|(\bar{X}_y - \bar{X}) - (C^T \otimes\ \hat{\Sigma}_x)\mathrm{vec}(A)\|^2 + \tau\ \mathrm{vec}(A)^T\mathrm{vec}(A)$$

$$= \|\widetilde{W}^{1/2}\ \tilde{Y} - \widetilde{W}^{1/2}(C^T \otimes \hat{\Sigma}_x)\ \mathrm{vec}(A)\|^2 + \tau\ \mathrm{vec}(A)^T\mathrm{vec}(A),$$

where $\otimes$ is the Kronecker product, $\tilde{Y} = \mathrm{vec}(\bar{X}_1 - \bar{X},\ldots, \bar{X}_h - \bar{X})$, $\widetilde{W}^{1/2} = D_f^{1/2} \otimes I_p$, and $D_f = \mathrm{diag}(\hat{f}_1,\ldots,\hat{f}_h)$. Then given C, the solution of A is found by

$\mathrm{vec}(\hat{A}) = (CD_f\, C^T \otimes \hat{\Sigma}_x{}^2 + \tau\, I_{pd})^{-1}\ (CD_f \otimes \hat{\Sigma}_x)\tilde{Y}$.

The solution to $G_\tau(A,C)$ is found by cycling between minimizing A and C until convergence.

The ridge parameter τ, for $G_\tau(A,C)$ is selected by minimizing a generalized crossvalidation criterion (GCV) which follows Golub's method (24).

$\mathrm{GCV} = \|(I_{ph} - S_\tau)\widetilde{W}^{1/2}\tilde{Y}\|^2/\ ph\{1 - \mathrm{trace}(S_\tau)/ph\}^2$,

where $S_\tau = (D_f^{1/2}\hat{C}^T \otimes \hat{\Sigma}_x)(\hat{C}D_f\hat{C}^T \otimes \hat{\Sigma}_x{}^2 + \tau\, I_{pd})^{-1}(\hat{C}D_f^{1/2} \otimes \hat{\Sigma}_x)$, and $d = \dim(S_{Y|X})$ is assumed to be known.

The ridge SIR estimates are linear combinations of all of the predictors. To achieve variable selection, the L_1 regularization is introduced from the least absolute shrinkage and selection operator idea (25). Let $(\hat{A}, \hat{C}) = \mathrm{argmin}_{A,C}\ G_\tau(A,C)$ denote the ridge SIR estimator. The sparse ridge SIR estimator of the central subspace $S_{Y|X}$ is then defined as Span(diag($\hat{\alpha}$)$\hat{A}$), where the shrinkage vector $\hat{\alpha}= (\hat{\alpha}_1, ..., \hat{\alpha}_p)^T \in \mathbb{R}^p$ is obtained by minimizing

$G_\lambda(\alpha) = \sum_{y=1}^{h} \hat{f}_y \|(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x \text{diag}(\alpha)\hat{A}\hat{C}_y\|^2$

over α, subject to $\sum_{j=1}^{p} |\alpha_j| \leq \lambda$ for $\lambda \geq 0$ (9). Optimization of $G_\lambda(\alpha)$ is done using the standard lasso algorithm. Noting that $\text{diag}(\alpha)\hat{A}\hat{C}_y = \text{diag}(\hat{A}\hat{C}_y)\alpha$,

$G_\lambda(\alpha) = \sum_{y=1}^{h} \hat{f}_y \|(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x \text{diag}(\hat{A}\hat{C}_y)\alpha\|^2$.

Then write:

$\tilde{Y} = \text{vec}(\bar{X}_1 - \bar{X},\dots, \bar{X}_h - \bar{X}) \in \mathbb{R}^{ph}$,

$\tilde{X} = (\text{diag}(\hat{A}\hat{C}_1)\, \hat{\Sigma}_x,\dots,\text{diag}(\hat{A}\hat{C}_h)\, \hat{\Sigma}_x)^T \in \mathbb{R}^{ph \times p}$,

where the shrinkage vector α is the Lasso estimator for the regression of $\tilde{Y}|\tilde{X}$ with *ph* observations. As the Lasso parameter λ decreases, some coefficients α_j are shrunk to zero, indicating that the corresponding predictors are not needed given the other predictors (9). When $\lambda \geq$ p, then $\hat{\alpha}_j = 1$, for j=1,...,p and $G_\lambda(\alpha)$ is equal to the ridge SIR estimator.

The family of information criteria is used for the selection of λ, including Akaike information criterion (AIC) (26), Bayesian information criterion (BIC) (27) and residual information criterion (RIC) (28):

$\text{AIC} = ph \log(G_\lambda(\hat{\alpha})/ph) + 2p_\lambda$,

$\text{BIC} = ph \log(G_\lambda(\hat{\alpha})/ph) + \log(ph)p_\lambda$,

$\text{RIC} = (ph - p_\lambda) \log \{G_\lambda(\hat{\alpha})/(ph - p_\lambda)\} + p_\lambda(\log(ph) - 1) + 4/(ph - p_\lambda - 2)$,

where p_λ denotes the number of parameters in the Lasso estimator, which is approximated by the number on nonzero components in the estimated $\hat{\alpha}$ (29).

In the estimation procedure of the regularized SIR, d=dim($\mathcal{S}_{Y|X}$) is assumed to be known. In practice, d needs to be estimated from the real data. A criterion proposed by Zhu, *et al.* will be used which estimates d by the number of nonzero eigenvalues of the matrix Cov(E(X|Y)), which is equivalent to the number of eigenvalues of the matrix Ω = Cov(E(X|Y)) + I_p that are greater than one (30). The suggested estimator of d is

$\hat{d}$=argmax$\{\frac{n}{2}\sum_{1|min(\kappa,m)}^{p}$ $(\log(\hat{\delta}_i) + 1 - \hat{\delta}_i) - (C_n m(2p - m + 1))/2\}$,

for m$\in\{0,1,..., p-1\}$,

where $\hat{\delta}_1, \dots, \hat{\delta}_p$ denotes the eigenvalues of the sample estimate $\hat{\Omega}$ of Ω, κ denotes the number of $\hat{\delta}_i$'s greater than one, and C_n denotes a penalty constant. For the current simulations, a penalty constant of C_n = log(n)h/n was used.

CHAPTER IV

SIMULATION STUDIES

A. Simulation Scenarios

A variety of settings were used to test the performance of the competing proposals for dealing with right censored data in the SIR with regularizations model. In all cases the accelerated failure time model was used, with the transformation function h as the natural logarithm. In all cases N=50 simulations were averaged for the final results. The various design parameters used are as follows:

(i) Covariate dimension: The covariate dimension p was taken to be 100.

(ii) Sample sizes:n=25 and 50 were considered as sample sizes.

(iii)Parameter values: Three different choices for the β coefficients were considered in order to cover a range of situations. The first case, $\beta_j=1$, corresponds to the situation in which all covariates have equal contribution to the regression function. The second case, $\beta_j= 1/j$, for $1\leq j \leq p$, corresponds to covariates that are decaying, and only a portion of these covariates contribute to the regression function. In the last case, $\beta_j = \frac{1}{\Sigma(\beta j)}$ for $1\leq j \leq p/2$, and $\beta_j= 0$ for $p/2 \leq j \leq p$, the first half of the covariates have an equal contribution to the regression, while the last half have zero contribution.

(iv) Design matrix: The rows X were generated from a multivariate normal distribution with mean zero and variance of one.

(v) Errors: Two types of error distributions were considered in the accelerated failure time model. The first choice the normal distribution, implies log-normally distributed failure times. Second we took the errors to be logarithms of the Weibull distribution, which leads to Weibull distributed failure times.

(vi) Censoring: The averaging censoring rates C_0 were chosen to be 0% for no censoring, 25% for low censoring and 50% for medium censoring.

B. Measures of Error

Measures of fit: The following measure of fit was computed to measure the fit in the training sample:

$\mathrm{MSE_f} = 1/(n_R\sigma^2) \sum_{i=1}^{n} \delta_i (\hat{Y}_i - \log T_i)^2$,

where $n_R = \sum_{i=1}^{n} \delta_i$.

This measure compares the fitted values with the true values corresponding to the uncensored units. For each design this measure was averaged over 50 independently generated training data sets.

Measures of prediction: For each training data set, a test data set $Y_{new} = \log T_{new}$ of the same size and design parameters was generated. The SIR with regularizations model was fitted using the training data set. The fitted model was used with the X-matrix of the test data to get predicted values $\hat{Y}_{new}$. The following measure was computed to determine the accuracy of prediction:

$\mathrm{MSE_p} = 1/(n_R\sigma^2) \sum_{i=1}^{n} (\hat{Y}_{new,i} - Y_{new,i})^2$.

For each design this measure was averaged over 50 independent replications of the entire process (3).

C. Simulation Results and Discussion

1. Tables (1-6)

To perform the simulations, a ridge parameter, τ, and a lasso parameter, λ, must be found to optimize the SIR models. As described previously a generalized crossvalidation criterion (GCV) is minimized to find the optimal ridge parameter. To select this parameter, the family of information criterion, AIC, BIC, and RIC are minimized. From Table 1, it is shown that the 25% censoring rate has smaller GCV values than 50% censoring rate in the corresponding methods. An increase in the sample size from 25 to 50 also shows a decrease in the GCV values. On average, the reweighting method had the smallest GCV, followed by the censored bivariate method. The following tables contain the minimized values used to select the ridge parameters.

TABLE 1

Criterion used to select ridge and lasso parameters (β_j=1, error = lognormal)

Model	Method	GCV	AIC	BIC	RIC
Censoring=0% n=25	Univariate	0.01901959	-1590.081	-1499.075	-1406.585
Censoring=25% n=25	Bivariate	0.010592	-1834.45	-1702.58	-1546.85
	Reweighting	0.005649	-1848.03	-1775.85	-1591.87
	Mean Imputation	0.013783	-1297.09	-1266.53	-1231.63
	Multiple Imputation	0.010606	-1199.22	-1167.83	-1129.1
Censoring=50% n=25	Bivariate	0.010629	-1833.34	-1701.42	-1545.87
	Reweighting	0.007992	-1681.3	-1607.14	-1449.14
	Mean Imputation	0.01498	-1320.42	-1289.55	-1255.32
	Multiple Imputation	0.012068	-1238.76	-1208.56	-1172.26
Censoring=0% n=50	Univariate	0.01169017	-1801.58	-1639.606	-1450.699
Censoring=25% n=50	Bivariate	0.007291	-2040.41	-1842.44	-1580.31
	Reweighting	0.003409	-1511.56	-1404.13	-1190.66
	Mean Imputation	0.010142	-1603.39	-1552.36	-1490.41
	Multiple Imputation	0.00731	-1391.41	-1341.23	-1272.88
Censoring=50% n=50	Bivariate	0.007319	-2038.19	-1840.2	-1578.34
	Reweighting	0.006366	-1498.92	-1387.17	-1203.88
	Mean Imputation	0.010692	-1650.86	-1599.85	-1539.38
	Multiple Imputation	0.008971	-1483.31	-1433.23	-1369.45

GCV: Generalized Crossvalidation Criterion; AIC:Akaike Information Criterion; BIC:Bayesian Information Criterion; RIC: Residual Information Criterion

TABLE 2

Criterion used to select ridge and lasso parameters (β_j=1/j, error = lognormal)

Model	Method	GCV	AIC	BIC	RIC
Censoring=0% n=25	Univariate	0.019742	-1577.98	-1493.36	-1408.06
Censoring=25% n=25	Bivariate	0.00940306	-1867.81	-1735.566	-1570.176
	Reweighting	0.001905	-1795.47	-1724.42	-1539.6
	Mean Imputation	0.01407	-1300.51	-1268.48	-1232.25
	Multiple Imputation	0.006865	-1133.88	-1104.1	-1064.26
Censoring=50% n=25	Bivariate	0.01035	-1843.25	-1710.14	-1552.13
	Reweighting	0.011353	-1475.89	-1399.06	-1278.8
	Mean Imputation	0.019598	-1491.03	-1459.5	-1428.35
	Multiple Imputation	0.015416	-1330.01	-1298.61	-1264.61
Censoring=0% n=50	Univariate	0.011451	-1811.28	-1651.86	-1465.06
Censoring=25% n=50	Bivariate	0.007113	-2050.45	-1845.17	-1571.76
	Reweighting	0.002432	-1573.15	-1459.1	-1215.72
	Mean Imputation	0.010566	-1682.65	-1627.66	-1562.16
	Multiple Imputation	0.005085	-1303.83	-1250.7	-1172.39
Censoring=50% n=50	Bivariate	0.007207	-2045.58	-1839.97	-1566.7
	Reweighting	0.008765	-1512.99	-1390.69	-1222.55
	Mean Imputation	0.011722	-1787.59	-1733.4	-1671.91
	Multiple Imputation	0.011659	-1792.48	-1738.6	-1677.34

GCV: Generalized Crossvalidation Criterion; AIC:Akaike Information Criterion; BIC:Bayesian Information Criterion; RIC: Residual Information Criterion

TABLE 3

Criterion used to select ridge and lasso parameters ($\beta_j = \frac{1}{\Sigma(\beta j)}$ for $1 \leq j \leq p/2$, and $\beta_j = 0$ for $p/2 \leq j \leq p$, error = lognormal)

Model	Method	GCV	AIC	BIC	RIC
Censoring=0% n=25	Univariate	0.019991	-1566.52	-1470.23	-1373.57
Censoring=25% n=25	Bivariate	0.007913	-1915.261	-1785.024	-1606.631
	Reweighting	0.001798	-1798.55	-1727.45	-1537.03
	Mean Imputation	0.012317	-1255.28	-1223.98	-1187.54
	Multiple Imputation	0.0132	-1279.3	-1247.75	-1210.99
Censoring=50% n=25	Bivariate	0.010253	-1845.79	-1711.51	-1552.02
	Reweighting	0.010998	-1457.47	-1379.38	-1257.4
	Mean Imputation	0.019775	-1507.62	-1475.87	-1444.69
	Multiple Imputation	0.017103	-1407.81	-1375.66	-1341.86
Censoring=0% n=50	Univariate	0.011434	-1810.34	-1649.01	-1460.01
Censoring=25% n=50	Bivariate	0.006864	-2060.12	-1856.55	-1584.26
	Reweighting	0.003319	-1544.67	-1427.66	-1186.14
	Mean Imputation	0.000854	-1522.66	-1468.83	-1399.42
	Multiple Imputation	0.006383	-1359.13	-1306.19	-1230.31
Censoring=50% n=50	Bivariate	0.006936	-2059.26	-1854.81	-1581.36
	Reweighting	0.008566	-1510.25	-1386.81	-1218.99
	Mean Imputation	0.011182	-1807.23	-1752.47	-1689.74
	Multiple Imputation	0.011068	-1733.86	-1680.24	-1618.04

GCV: Generalized Crossvalidation Criterion; AIC:Akaike Information Criterion; BIC:Bayesian Information Criterion; RIC: Residual Information Criterion

TABLE 4

Criterion used to select ridge and lasso parameters (β_j=1, error = Weibull)

Model	Method	GCV	AIC	BIC	RIC
Censoring=0% n=25	Univariate	0.020705	-1557.79	-1475.56	-1393.72
Censoring=25% n=25	Bivariate	0.010466	-1838.51	-1707.37	-1552.33
	Reweighting	0.007282	-1530.633	-1458.763	-1318.266
	Mean Imputation	0.013266	-1260.92	-1230.37	-1194.9
	Multiple Imputation	0.005213	-1115.07	-1083.09	-1037.55
Censoring=50% n=25	Bivariate	0.010452	-1838.19	-1706.79	-1551.44
	Reweighting	0.00955341	-1454.442	-1379.755	-1254.412
	Mean Imputation	0.014694	-1302.48	-1272.18	-1238.22
	Multiple Imputation	0.012445	-1255.16	-1225.61	-1191.55
Censoring=0% n=50	Univariate	0.011893	-1806.274	-1676.552	-1526.02
Censoring=25% n=50	Bivariate	0.007322	-2039.26	-1840.01	-1576.29
	Reweighting	0.002869	-1520.408	-1413.192	-1191.202
	Mean Imputation	0.010196	-1578.09	-1527.07	-1464.85
	Multiple Imputation	0.008487	-1408.83	-1357.32	-1290.7
Censoring=50% n=50	Bivariate	0.007218	-2043.93	-1845.54	-1582.54
	Reweighting	0.006627	-1507.849	-1395.255	-1213.223
	Mean Imputation	0.011069	-1673.38	-1623.07	-1564.18
	Multiple Imputation	0.008308	-1470.75	-1420.75	-1355.38

GCV: Generalized Crossvalidation Criterion; AIC:Akaike Information Criterion; BIC:Bayesian Information Criterion; RIC: Residual Information Criterion

TABLE 5

Criterion used to select ridge and lasso parameters (β_j=1/j, error = Weibull)

Model	Method	GCV	AIC	BIC	RIC
Censoring=0% n=25	Univariate	0.019329	-1587.94	-1509.71	-1430.45
Censoring=25% n=25	Bivariate	0.00957	-1863.414	-1730.258	-1566.594
	Reweighting	1.14E-06	-1804.04	-1736.43	-1548.35
	Mean Imputation	0.014681	-1334.49	-1303	-1267.73
	Multiple Imputation	0.00711	-1120.67	-1090.6	-1050.6
Censoring=50% n=25	Bivariate	0.010316	-1844.39	-1711.15	-1553.06
	Reweighting	0.01069076	-1663.065	-1586.759	-1441.508
	Mean Imputation	0.019422	-1490.01	-1458.3	-1426.96
	Multiple Imputation	0.017196	-1391.168	-1359.558	-1326.133
Censoring=0% n=50	Univariate	0.01149	-1807.8	-1645.19	-1454.92
Censoring=25% n=50	Bivariate	0.007117	-2048.63	-1843.27	-1569.35
	Reweighting	1.13E-05	-1606.724	-1500.518	-1233.792
	Mean Imputation	0.010764	-1685.98	-1632.72	-1569.73
	Multiple Imputation	0.006494	-1370.56	-1321.05	-1251.1
Censoring=50% n=50	Bivariate	0.007203	-2046.61	-1843.64	-1573.88
	Reweighting	0.008881	-1533.77	-1412.02	-1245.42
	Mean Imputation	0.011708	-1785.93	-1733.3	-1673.63
	Multiple Imputation	0.010503	-1684.5	-1629.19	-1563.45

GCV: Generalized Crossvalidation Criterion; AIC:Akaike Information Criterion; BIC:Bayesian Information Criterion; RIC: Residual Information Criterion

TABLE 6

Criterion used to select ridge and lasso parameters ($\beta_j = \frac{1}{\Sigma(\beta j)}$ for $1 \le j \le p/2$, and $\beta_j = 0$ for $p/2 \le j \le p$, error = Weibull)

Model	Method	GCV	AIC	BIC	RIC
Censoring=0% n=25	Univariate	0.019991	-1566.52	-1470.23	-1373.57
Censoring=25% n=25	Bivariate	0.007876	-1922.6	-1791.57	-1610.48
	Reweighting	0.003473	-1757.02	-1683.15	-1498.45
	Mean Imputation	0.01308	-1292.42	-1262.84	-1227.88
	Multiple Imputation	0.015691	-1323.6	-1291.29	-1255.38
Censoring=50% n=25	Bivariate	0.010231	-1846.46	-1712.01	-1552.25
	Reweighting	0.01145	-1417.68	-1340.63	-1228.26
	Mean Imputation	0.019351	-1509.66	-1477.64	-1446.07
	Multiple Imputation	0.017943	-1435.41	-1404.35	-1372.59
Censoring=0% n=50	Univariate	0.011434	-1810.34	-1649.01	-1460.01
Censoring=25% n=50	Bivariate	0.006689	-2061.01	-1854.65	-1574.61
	Reweighting	0.003486	-1509.63	-1389.33	-1151.54
	Mean Imputation	0.009867	-1600.5	-1547.1	-1481.52
	Multiple Imputation	0.006952	-1452.08	-1397.01	-1320.77
Censoring=50% n=50	Bivariate	0.006951	-2055.41	-1849.09	-1573.56
	Reweighting	0.008192	-1501.53	-1378.27	-1208.84
	Mean Imputation	0.011559	-1794.14	-1739.56	-1677.46
	Multiple Imputation	0.011298	-1771.02	-1716.04	-1652.96

GCV: Generalized Crossvalidation Criterion; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; RIC: Residual Information Criterion

2. Table 7

A variety of Ridge parameters were tried, ranging from $\tau = .001,...,10$, to minimize the GCV values provided in Tables 1-6. Once the ridge parameter was selected, the lasso parameter was varied using $\lambda = 10,...,50$ to minimize the family of information criterion used. Table 7 shows the final parameters used for the various methods. The # predictors column represents the average number of effective predictors used in finding the central subspace. The larger values for the Lasso parameter correspond to a greater number of predictors used. The average residual sum of squares (RSS) for the various methods is also provided, indicating the discrepancy between the data and the fit of the regularized SIR. Reweighting has the smallest RSS value which would indicate the best fit, while the mean imputation method has the highest RSS value. The univariate method, which is used on the uncensored data, has the second highest RSS at 4.569871, which should theoretically be the best fit of the data.

TABLE 7

Average values for regularized SIR for various methods

Method	**Ridge parameter**	**Lasso parameter**	**# predictors**	**RSS**
Univariate	.001	40	35.764	4.569871
Bivariate	.001	40	42.052	2.701443
Reweighting	.05	30	26.821	1.563724
Mean Imputation	1	10	10.881	4.967454
Multiple Imputation	1	10	11.230	3.886914

RSS: Residual Sum of Squares

3. Tables (8-11)

Once the estimate of the central subspace is found, that estimate, B^TZ, is used as the new variable to fit the accelerated failure time model. Tables 8-13 provide coefficients and intercepts for the different model scenarios, along with the corresponding p-value from the Wald test. The method labeled univariate corresponds to the uncensored data. In Table 8, with n=25 and error =lognormal, the mean imputation method in most cases has significant values for both the intercept and B^TZ, with p-values $< .05$. The reweighting values are mostly non-significant with the exception of the $B_j=1/j$ parameter values. With the smaller sample size of n=25, the majority of B^TZ parameters are insignificant. However, as the sample size increases to n=50, all of the parameters become significant. Decreasing the censoring rate from 50% to 25% also contributes to more significant values as expected. There does not appear to be much of a difference between the lognormal errors, as compared to the Weibull distributed error terms. The uncensored univariate method would be expected to produce the smallest p-values, although this is not the case. In summary, the mean imputation method produces the most significant central subspace when fit in the accelerated failure time model. A sample size of n=25 appears to be too small to produce any significant models when parameters of $B_j=1$ are used. The reweighting method appears to be the least effective at generating a significant central subspace.

TABLE 8

Coefficients and Wald test for accelerated failure time model(n=25, error = lognormal)

Censoring rate		$B_j = 1$		$B_j = 1/j$		$B_j = 1,0$	
	Method	Intercept	B^TZ	Intercept	B^TZ	Intercept	B^TZ
C=0%	**Univariate**	-1.89	7.24	-.584	-.271	-.313	-.191
	SE	.860	.702	.334	.246	.184	.130
	P-value	.0282	5.74E-25	.0806	.271	.0893	.141
C=25%	**Bivariate**	-2.94	0.558	-0.702	0.69	-.373	.433
	SE	1.512	1.303	0.315	0.135	.555	.198
	P-value	.0518	.668	.026	2.94E-07	.502	.0283
	Reweighting	-3.09	-0.86	-1.18	1.335		
	SE	2.632	1.883	0.143	0.111		
	P-value	.241	.648	1.94E-16	2.47E-33		
	Mean Imputation	-6.84	2.64	-2.098	0.126	-1.558	0.108
	SE	.601	.508	0.0489	0.0378	0.0699	0.0552
	P-value	5.11E-30	2.02E-07	0.00	8.32E-04	4.95E-110	.0501
	Multiple Imputation	-7.274	0.418	-2.421	0.377	-1.5161	0.0348
	SE	1.092	0.83	0.0773	0.0589	0.119	0.0946
	P-value	2.71E-11	.614	2.24E-215	1.49E-10	3.42E-37	.713
C=50%	**Bivariate**	-3.92	2.75	0.0913	1.1179	-0.1679	0.1285
	SE	.991	.786	0.347	0.409	0.289	0.225
	P-value	7.48E-05	4.68E-04	0.79259	0.00628	.562	.567
	Reweighting	-3.1562	-0.0264	-0.7074	0.8276	-0.1802	0.2334
	SE	2.542	1.846	0.213	0.149	0.225	0.183
	P-value	.214	.989	8.94E-04	2.82E-08	.424	.202
	Mean Imputation	-5.83	2.27	-1.142	0.138	-0.96	0.055
	SE	0.865	0.612	0.0792	0.0589	0.0962	0.0771
	P-value	1.61E-11	2.16E-04	3.90E-47	.0191	1.96E-23	.476
	Multiple Imputation	-5.238	0.624	-1.098	-0.254	-0.863	0.125
	SE	1.259	0.939	0.0748	0.0645	0.1301	0.0886
	P-value	3.19E-05	.506	8.95E-49	8.30E-05	3.29E-11	.160

SE: Standard error

TABLE 9

Coefficients and Wald test for accelerated failure time model (n=50, error = lognormal)

Censoring rate		$B_j = 1$		$B_j = 1/j$		$B_j = 1,0$	
	Method	Intercept	B^TZ	Intercept	B^TZ	Intercept	B^TZ
C=0%	**Univariate**	-4.16	7.66	-.413	.0266	-.439	.780
	SE	.863	.826	.213	.204	.0768	.0754
	P-value	1.42E-06	1.76E-20	.0522	.896	1.08E-08	4.47E-25
C=25%	**Bivariate**	-0.708	4.498	-0.32241	1.33532	-0.15	-0.38
	SE	1.583	1.521	0.39	0.329	0.424	0.234
	P-value	.655	3.11E-03	.408	4.86E-05	.724	.105
	Reweighting	-2.66	6.45	-0.485	1.4	1.76	-1.76
	SE	2.71	2.05	0.276	0.267	.841	1.02
	P-value	.326	1.67E-03	.0793	1.56E-07	.0361	.0850
	Mean Imputation	-5.51	1.90	-1.729	0.23	-1.6495	0.0332
	SE	.621	.436	0.0693	0.0598	0.0496	0.0384
	P-value	6.77E-19	1.32E-05	2.14E-137	1.20E-04	2.41E-242	.387
	Multiple Imputation	-5.8	-1.98	-1.782	-0.171	-1.6789	-0.0198
	SE	0.72	0.587	0.11	0.104	0.0877	0.0733
	P-value	8.13E-16	7.38E-04	4.45E-59	.100	1.17E-81	.787
C=50%	**Bivariate**	-1.08	4.85	0.5486	1.6144	-0.2529	0.0166
	SE	1.038	.896	0.301	0.334	0.187	0.17
	P-value	.298	6.13E-08	.0688	1.31E-06	.177	.922
	Reweighting	-3.21	5.75	-0.505	0.427	-0.15	0.187
	SE	2.62	1.9	0.321	0.269	0.388	0.386
	P-value	.221	2.53E-03	.115	.112	.698	.628
	Mean Imputation	-4.42	-2.14	-0.595	0.196	-0.533	0.065
	SE	0.656	0.544	0.132	0.109	0.104	0.104
	P-value	1.69E-11	8.25E-05	6.23E-06	.0715	2.87E-07	.531
	Multiple Imputation	-4.7	-2.29	-0.7842	0.1694	-0.578	-0.549
	SE	0.672	0.565	0.146	0.118	0.0959	0.0873
	P-value	2.72E-12	5.08E-05	8.63E-08	.153	1.72E-09	3.22E-10

SE: Standard Error

TABLE 10

Coefficients and Wald test for accelerated failure time model (n=25, error = weibull)

Censoring rate		$B_j = 1$		$B_j = 1/j$		$B_j = 1,0$	
	Method	Intercept	B^TZ	Intercept	B^TZ	Intercept	B^TZ
C=0%	**Univariate**	.374	7.52	.155	-.629	.165	-.316
	SE	.965	.616	.265	.173	.187	.101
	P-value	.698	2.63E-34	.558	2.83E-04	.377	1.72E-03
C=25%	**Bivariate**	-1.34	1.44E-03	-.693	.643	-.303	.394
	SE	1.18	1.01	.258	.108	.458	.136
	P-value	.257	.999	7.31E-03	2.64E-09	.508	3.68E-03
	Reweighting	4.93	-3.76	-.850	1.17		
	SE	4.43	3.15	.115	.0904		
	P-value	.266	.234	1.29E-13	1.78E-38		
	Mean Imputation	-5.39	2.17	-1.98	.183	-1.41	.151
	SE	.527	.402	.0576	.0368	.059	.041
	P-value	1.53E-24	7.32E-08	8.07E-259	6.76E-07	4.99E-126	2.32E-04
	Multiple Imputation	-4.63	-.107	-2.23	.38	-1.22	-.0174
	SE	.891	.493	.0978	.0635	.113	.126
	P-value	2.05E-07	.829	2.04E-115	2.03E-09	2.73E-27	.891
C=50%	**Bivariate**	-2.448	1.84	.624	1.39	.191	.203
	SE	.616	.473	.448	.525	.304	.231
	P-value	7.02E-05	9.64E-05	.164	8.12E-03	.529	.380
	Reweighting	4.19	3.16	-.127	.898	.242	-.402
	SE	4.22	3.22	.287	.185	.308	.304
	P-value	.320	.327	.658	1.24E-06	.431	.185
	Mean Imputation	-3.73	1.78	-.909	.199	-.753	-.132
	SE	.798	.579	.109	.106	.101	.0842
	P-value	3.01E-06	2.14E-03	5.57E-17	.0605	1.07E-13	.116
	Multiple Imputation	-2.709	-.504	-.891	-.254	-.505	-.0459
	SE	.951	.842	.107	.121	.125	.0853
	P-value	4.37E-03	.550	7.70E-17	.0363	5.47E-05	.591

SE: Standard Error

TABLE 11

Coefficients and Wald test for accelerated failure time model (n=50, error = weibull)

Censoring rate		$B_j = 1$		$B_j = 1/j$		$B_j = 1,0$	
	Method	Intercept	B^TZ	Intercept	B^TZ	Intercept	B^TZ
C=0%	**Univariate**	-1.35	.735	.312	.228	-.167	.840
	SE	.736	.595	.213	.215	.0865	.0729
	P-value	.0664	4.65E-35	.143	.291	.0532	9.39E-31
C=25%	**Bivariate**	1.90E-03	2.40	.220	1.32	.334	-.548
	SE	1.09	.845	.404	.315	.384	.201
	P-value	.999	4.45E-03	.587	2.58E-05	.384	6.34E-03
	Reweighting	10.5	17.2	.993	2.43		
	SE	5.63	3.98	5.86E-01	.482		
	P-value	.0615	1.46E-05	.0899	4.78E-07		
	Mean Imputation	-3.85	.527	-1.52	.208	-1.47	.115
	SE	.348	.313	.0536	.0416	.0465	.0254
	P-value	1.89E-28	.0929	3.41E-177	5.98E-07	5.64E-219	6.23E-06
	Multiple Imputation	-3.67	-.413	-1.42	-.146	-1.36	.0528
	SE	.379	.345	.0938	.0802	.0891	.0536
	P-value	3.53E-22	.232	1.30E-51	.0682	1.31E-52	.325
C=50%	**Bivariate**	-0.24	3.16	1.11	1.76	.224	.278
	SE	.698	.569	.297	.319	.184	.165
	P-value	.731	2.95E-08	1.86E-04	3.16E-08	.223	.0913
	Reweighting	8.78	15.0	1.03	.123	1.76	-1.76
	SE	5.53	3.95	.676	.706	.841	1.02
	P-value	.112	1.51E-04	.129	.862	.0361	.085
	Mean Imputation	-2.812	-.649	-.130	.0233	-.181	.149
	SE	.403	.372	.131	.117	.0912	.0927
	P-value	3.03E-12	.081	.322	.841	.0477	.107
	Multiple Imputation	-2.90	-.698	-.275	.200	-.229	-.536
	SE	.425	.387	.146	.129	.101	.105
	P-value	8.65E-12	.0711	.0586	.120	.0237	3.67E-07

SE: Standard Error

4. Tables 12-13

Tables 12 and 13 contain the scale value and the p-value for the likelihood ratio test comparing the accelerated failure time model with the intercept only, to the model including B^TZ. Again in most cases the mean imputation has the smallest p-value, indicating the best fit for the model. With the exception of B_j= 1/j, the univariate uncensored case also produces a highly significant p-values. The multiple imputation method most commonly has the highest p-value. The mean imputation also has the lowest scale(standard deviation) value in most cases along with the univariate method. The reweighting method tends to have the highest scale value, indicating a greater degree of variation in the central subspace estimate.

TABLE 12

Scale and Log-Likelihood test for accelerated failure time model(error = lognormal)

			n=25		n=50	
Censoring rate	Parameter values	Method	Scale	P-value	Scale	P-value
0%	B_j=1/j	Univariate	1.67	.280	1.49	.900
	B_j=1		4.29	1.20E-10	5.89	1.50E-12
	B_j=1,0		0.91	.150	0.536	3.90E-14
25%	B_j=1/j	Bivariate	.263	1.20E-09	1	9.20E-10
		Reweighting	.715	4.50E-12	1.94	2.80E-06
		Mean imputation	.241	.0024	.488	3.20E-04
		Multiple imputation	.34	8.30E-07	.777	.100
	B_j=1	Bivariate	5.98	.68	7.7	.00076
		Reweighting	13.2	.65	19.1	.0027
		Mean imputation	3.01	1.90E-05	4.39	6.00E-05
		Multiple imputation	5.44	.62	5.04	.0014
	B_j=1,0	Bivariate	1.06	6.00E-03	1.48	.088
		Reweighting				
		Mean imputation	.349	.059	.345	.39
		Multiple imputation	.594	.71	.62	.79
50%	B_j=1/j	Bivariate	.699	.0058	.967	1.20E-07
		Reweighting	1.06	7.40E-06	2.26	.12
		Mean imputation	.385	.026	.928	.076
		Multiple imputation	.367	.00052	1.03	.16
	B_j=1	Bivariate	4.4	.003	5.67	2.00E-07
		Reweighting	12.7	.99	18.4	.0038
		Mean imputation	4.33	.00095	4.62	.00024
		Multiple imputation	6.13	.51	4.75	.00016
	B_j=1,0	Bivariate	.97	.54	1.08	.92
		Reweighting	1.08	.21	2.74	.63
		Mean imputation	.473	.48	.72	.53
		Multiple imputation	.648	.17	.676	6.80E-08

TABLE 13

Scale and Log-Likelihood test for accelerated failure time model (error = weibull)

			n=25		n=50	
Model	Parameter values	Method	Scale	P-value	Scale	P-value
0%	$B_j=1/j$	Univariate	1.26	4.60E-03	1.41	.310
	$B_j=1$		4.53	3.80E-10	4.83	7.00E-15
	$B_j=1,0$		0.87	8.00E-03	0.572	7.60E-14
25%	$B_j=1/j$	Bivariate	.178	1.30E-10	.852	1.60E-10
		Reweighting	.544	1.50E-10	3.8	7.60E-05
		Mean imputation	.271	2.00E-04	.36	1.70E-05
		Multiple imputation	.419	3.80E-06	.629	.078
	$B_j=1$	Bivariate	3.5	1	4.23	8.80E-04
		Reweighting	20.7	.27	37.2	2.50E-04
		Mean imputation	2.5	4.60E-05	2.38	.084
		Multiple imputation	4.22	.83	2.55	.22
	$B_j=1,0$	Bivariate	.537	5.60E-04	1.03	.0089
		Reweighting				
		Mean imputation	.279	2.90E-03	.305	2.90E-04
		Multiple imputation	.533	.89	.594	.32
50%	$B_j=1/j$	Bivariate	.667	8.10E-03	.81	8.80E-09
		Reweighting	1.33	3.20E-04	4.41	.86
		Mean imputation	.473	.066	.87	.84
		Multiple imputation	.482	.037	.971	.13
	$B_j=1$	Bivariate	2.45	.0024	3.32	6.10E-07
		Reweighting	19.7	.36	36.7	9.70E-04
		Mean imputation	3.78	.0047	2.74	.075
		Multiple imputation	4.43	.56	2.89	.065
	$B_j=1,0$	Bivariate	.698	.32	.927	.093
		Reweighting	1.42	.21	5.56	.089
		Mean imputation	.476	.13	.596	.12
		Multiple imputation	.585	.60	.669	5.30E-06

5. Tables 14-17

Tables 14-17 show the mean squared error for the fitted values (mse_f) as well as the predicted values(mse_p). As can be seen from the tables, the mean imputation method consistently has the best fit among the various methods. The multiple imputation and bivariate methods were very similar in there fits. The reweighting method appears to be inadequate in fitting the data, having extremely high mse values. There is a small improvement in fit for the various methods when decreasing the censoring from 50% to 25%. It is unclear why the full data set (uncensored) has higher mse_f values than the mean imputation method. Theoretically it should have the lowest values. Increasing the sample size from 25 to 50 had minimal effect on the fit. The different choices for B_j coefficients appear to be the most important determinant of fit.

The accuracy of prediction is an important aspect of performance which is reflected by the mean squared error of prediction. The mean imputation and multiple imputation method appear to be the most accurate at prediction, having the lowest mse_p values. The reweighting approach is clearly the worst at prediction. Overall, none of the methods are very precise in predicting the survival times.

TABLE 14

Mean squared error for fit and prediction of different methods (N=50, n=25, error = lognormal).

Censoring rate	Parameter values	Method	MSE$_f$	MSE$_p$
0%	B_j=1/j	Univariate	3.616304	62.30974
	B_j=1		2425293784	3.27E+16
	B_j=1,0		0.93109	8.313126
25%	B_i=1/j	Bivariate	1.205127	2614730763
		Reweighting	1.56E+12	24332.26
		Mean imputation	0.093836	112.2441
		Multiple imputation	4.240646	11.93516
	B_i=1	Bivariate	1.93E+15	1.52E+56
		Reweighting	5.41E+17	4.60E+33
		Mean imputation	1.119058	1.16E+26
		Multiple imputation	111.4531	126.1125
	B_j=1,0	Bivariate	2.68E+15	1.08E+129
		Reweighting	17.15487	5.320607
		Mean imputation	0.09626	9.566693
		Multiple imputation	1.983821	59.66738
50%	B_i=1/j	Bivariate	0.922607	10359.14
		Reweighting	998563.9	1554.474
		Mean imputation	0.743674	189.5635
		Multiple imputation	3.873195	20.36823
	B_i=1	Bivariate	94227.67	3.09E+49
		Reweighting	3.04E+39	2.74E+38
		Mean imputation	13.76206	1.10E+29
		Multiple imputation	2513.54	59903.26
	B_i=1,0	Bivariate	0.560521	2189.774
		Reweighting	1456.561	8.476804
		Mean imputation	0.636887	6.023066
		Multiple imputation	2.156252	54.31242

MSE$_f$: $1/(n_R\sigma^2) \sum_{i=1}^{n} \delta_i (\hat{Y}_i - \log T_i)^2$; MSE$_p$: $1/(n_R\sigma^2) \sum_{i=1}^{n} (\hat{Y}_{new,i} - Y_{new,i})^2$

TABLE 15

Means squared error for fit and prediction of different methods (N=50, n=50, error = lognormal).

Censoring rate	Parameter values	Method	MSE_f	MSE_p
0%	B_j=1/j	Univariate	4.482463	123.1845
	B_j=1		6.87343E+11	1.50E+29
	B_j=1,0		0.902584	7.858615
25%	B_j=1/j	Bivariate	1.023037	5075.941
		Reweighting	50895.03	154.9116
		Mean imputation	0.110841	84.85015
		Multiple imputation	5.044459	16.69947
	B_j=1	Bivariate	3344.053	1.38E+25
		Reweighting	5.07E+22	5.60E+33
		Mean imputation	183.9164	4.63E+30
		Multiple imputation	117.4351	22537.52
	B_j=1,0	Bivariate	1.489582	15.22566
		Reweighting	2717858	808.3046
		Mean imputation	0.328347	6.564865
		Multiple imputation	2.235019	6.393538
50%	B_j=1/j	Bivariate	0.86449	660.3088
		Reweighting	4.03E+13	44747216
		Mean imputation	0.69669	153.7396
		Multiple imputation	3.691885	20.87405
	B_j=1	Bivariate	7495.397	6.14E+31
		Reweighting	1.61E+94	1.23E+66
		Mean imputation	9903.056	4.23E+33
		Multiple imputation	133861.4	14295875
	B_j=1,0	Bivariate	0.526039	16.93448
		Reweighting	4329281	18.15942
		Mean imputation	0.562776	5.170296
		Multiple imputation	2.145936	68.1274

MSE_f: $1/(n_R\sigma^2)\sum_{i=1}^{n}\delta_i(\hat{Y}_i - \log T_i)^2$; MSE_p: $1/(n_R\sigma^2)\sum_{i=1}^{n}(\hat{Y}_{new,i} - Y_{new,i})^2$

TABLE 16

Means squared error for fit and prediction of different methods (N=50, n=25, error=weibull).

Censoring rate	Parameter values	Method	MSE$_f$	MSE$_p$
0%	B_j=1/j	Univariate	13.11628	36.25056
	B_j=1		5.95E+15	4.41E+15
	B_j=1,0		2.67601	8.606185
25%	B_i=1/j	Bivariate	3.32E+20	6.47E+212
		Reweighting	2.04E+05	1.31E+08
		Mean imputation	0.130834	194.4398
		Multiple imputation	4.396683	11.75878
	B_j=1	Bivariate	4.15E+13	2.07E+30
		Reweighting	1.94E+99	6.34E+143
		Mean imputation	.401	1.72E+25
		Multiple imputation	200	4.83E+03
	B_j=1,0	Bivariate	539.9086	3.55E+36
		Reweighting	2.502239	16.62973
		Mean imputation	0.288499	8.160324
		Multiple imputation	2.054805	4.320406
50%	B_i=1/j	Bivariate	1.825946	352.9547
		Reweighting	4.63E+14	28264791746
		Mean imputation	0.989726	112.1048
		Multiple imputation	1.31E+297	2.89E+299
	B_i=1	Bivariate	511699.7	1.04E+26
		Reweighting	4.96E+81	7.50E+123
		Mean imputation	1937.766	2.04E+26
		Multiple imputation	27535118	5.21E+14
	B_i=1,0	Bivariate	1.259783	15822.79
		Reweighting	6943643	1002579
		Mean imputation	0.661466	5.769323
		Multiple imputation	2.724909	114.5771

MSE$_f$: $1/(n_R\sigma^2)\sum_{i=1}^{n}\delta_i(\hat{Y}_i - \log T_i)^2$; MSE$_p$: $1/(n_R\sigma^2)\sum_{i=1}^{n}(\hat{Y}_{new,i} - Y_{new,i})^2$

TABLE 17

Means squared error for fit and prediction of different methods (N=50, n=50, error = weibull).

Censoring rate	Parameter values	Method	MSE_f	MSE_p
0%	B_j=1/j	Univariate	19.25244	244.4261
	B_j=1		1.29E+16	6.12E+18
	B_j=1,0		2.356578	9.324723
25%	B_i=1/j	Bivariate	1.635221	280041.7
		Reweighting	6.02528E+11	1.66E+23
		Mean imputation	0.133127	342.1746
		Multiple imputation	5.191936	20.37365
	B_i=1	Bivariate	619.9674	9.57E+24
		Reweighting	3.69E+86	5.31E+132
		Mean imputation	2.848023	4.50E+27
		Multiple imputation	137	240
	B_i=1,0	Bivariate	3.084173	7299432
		Reweighting	4.36E+13	68278966
		Mean imputation	0.200152	6.857888
		Multiple imputation	2.977191	11.87037
50%	B_i=1/j	Bivariate	1.826705	178.8969
		Reweighting	1.76E+60	1.37E+43
		Mean imputation	0.90592	143.9906
		Multiple imputation	5.279777	105.136
	B_i=1	Bivariate	600.1246	9.25E+27
		Reweighting	2.82E+266	1.92E+192
		Mean imputation	421561.3	2.37E+31
		Multiple imputation	7.50E+21	1.42E+23
	B_i=1,0	Bivariate	1.257951	68.85456
		Reweighting	8.12E+18	1.27E+15
		Mean imputation	0.583706	5.5258
		Multiple imputation	2.685952	73.69357

MSE_f: $1/(n_R\sigma^2)\sum_{i=1}^{n}\delta_i(\hat{Y}_i - \log T_i)^2$; MSE_p: $1/(n_R\sigma^2)\sum_{i=1}^{n}(\hat{Y}_{new,i} - Y_{new,i})^2$

6. Figures 1-5

Figures 1-5 illustrate the actual survival times plotted against the predicted times using the various methods. In each case the line $y = x$ is also plotted indicating the theoretical value which would minimize the mse_p. In all the plots the data was simulated using the model coefficient $B_j=1/j$ for $1\leq j \leq p$, censoring rate =25%, sample size= 25, and lognormally distributed errors. As seen in the plots, the accuracy of prediction is not very encouraging for the various methods. The uncensored data set, and the reweighting method appear the least precise in there prediction accuracy. More work needs to be done to improve the prediction accuracy.

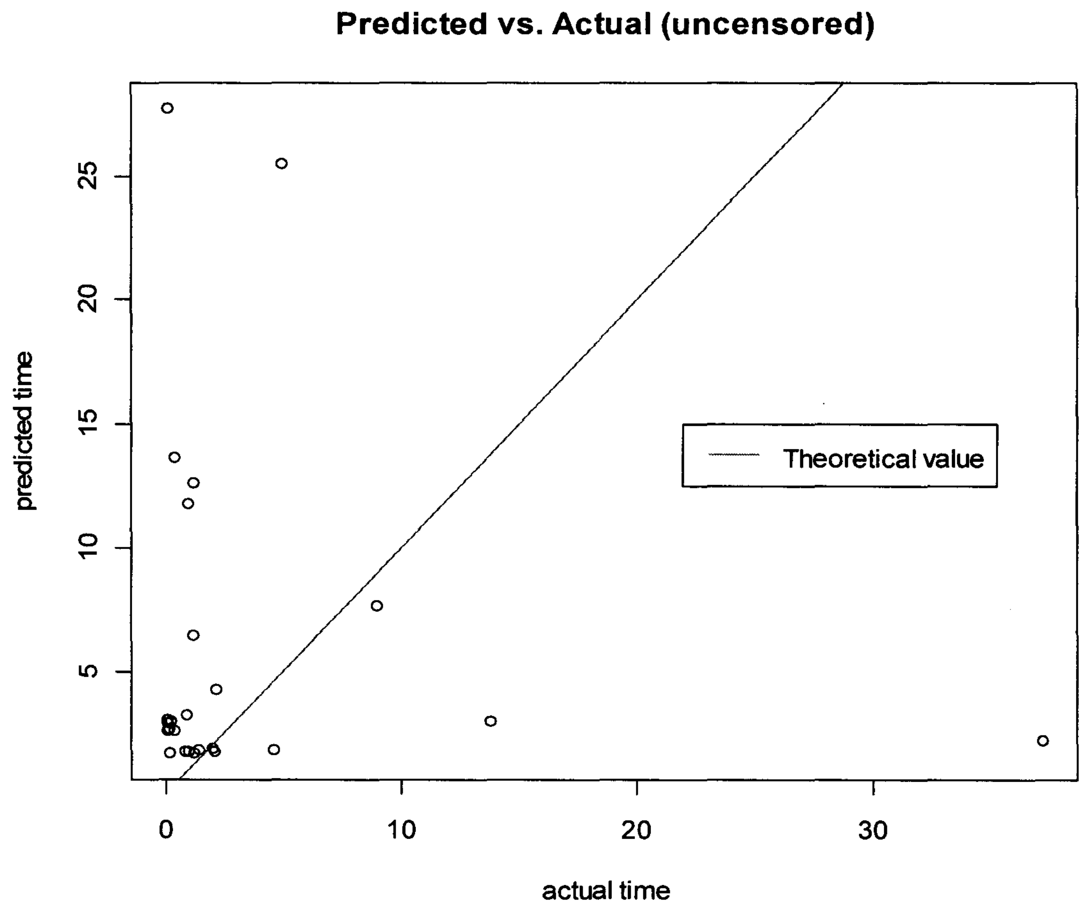

Figure 1. Predicted vs. Actual time points for $B_j=1/j$, censoring rate=25%, n= 25

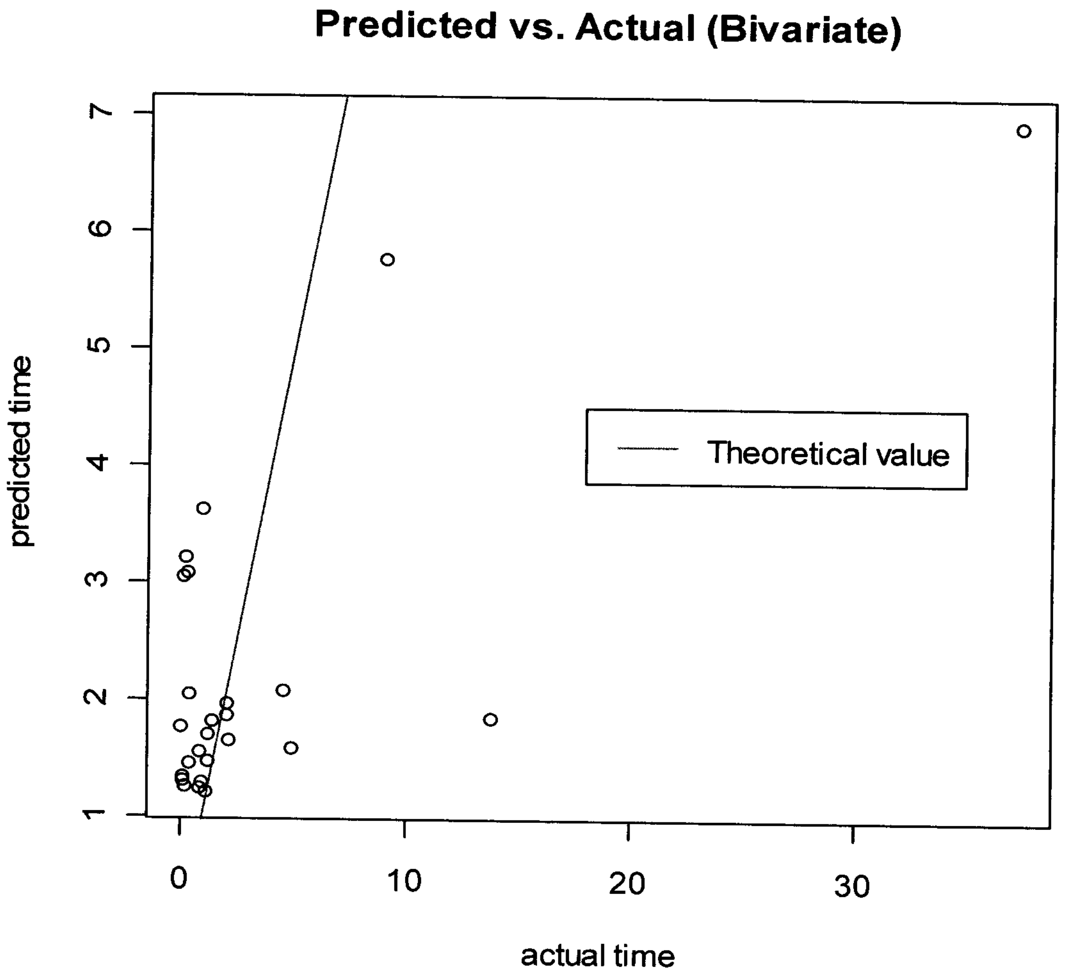

Figure 2. Predicted vs. Actual time points for $B_j=1/j$, censoring rate=25%, n= 25

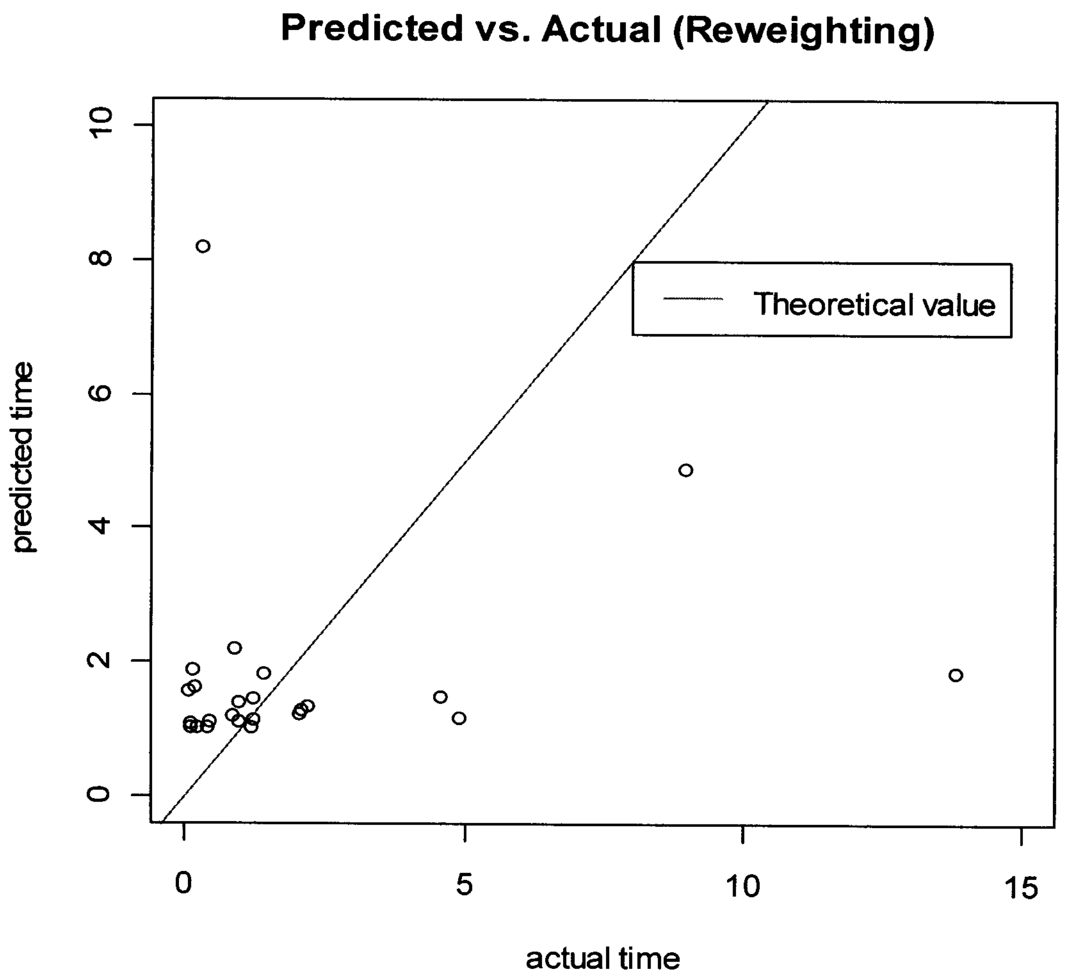

Figure 3. Predicted vs. Actual time points for B_j=1/j, censoring rate=25%, n= 25

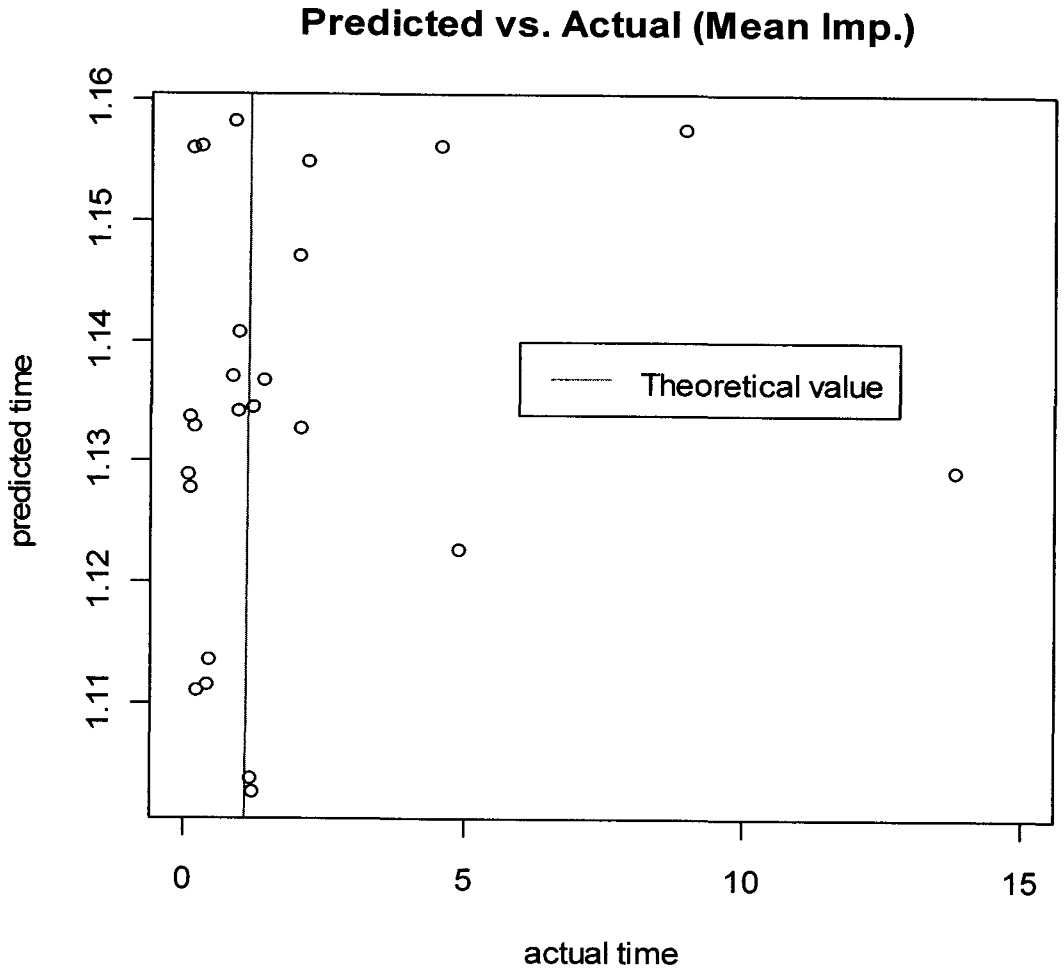

Figure 4. Predicted vs. Actual time points for B_j=1/j, censoring rate=25%, n= 25

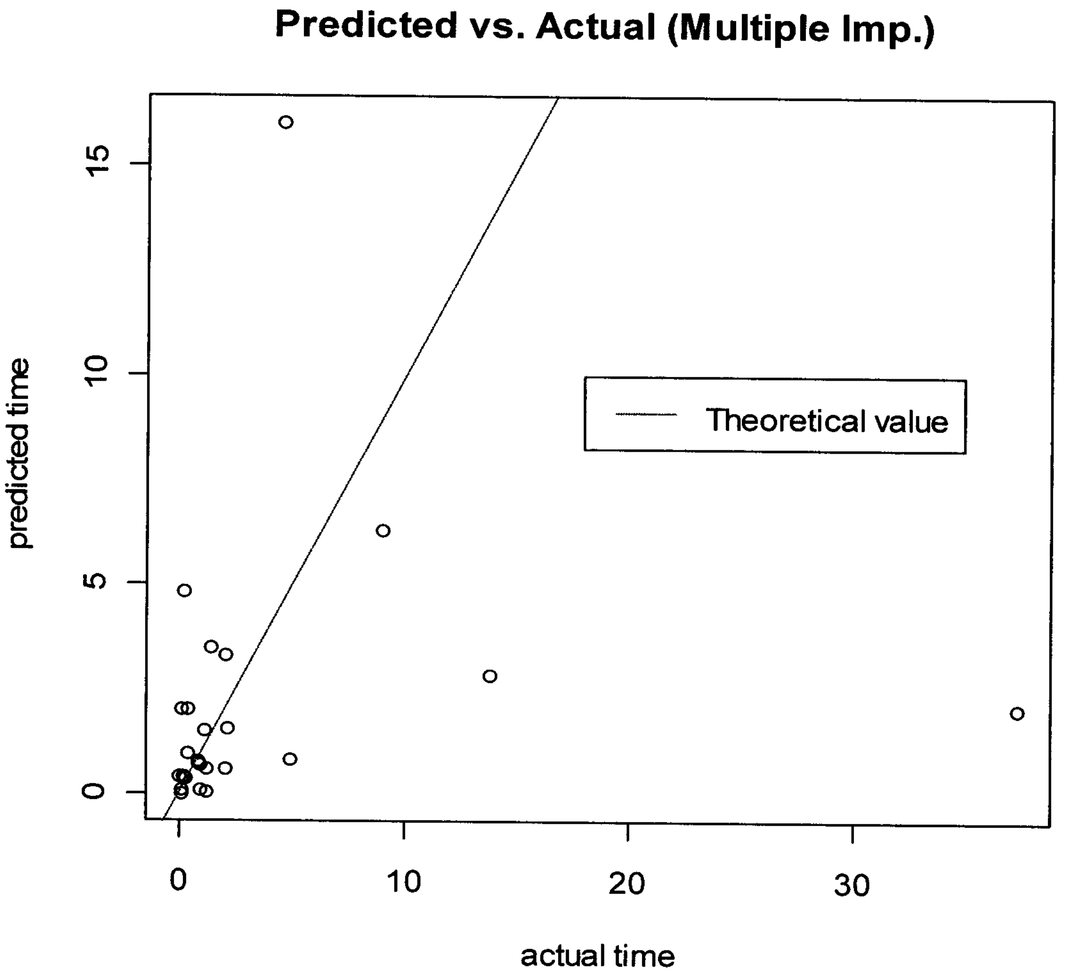

Figure 5. Predicted vs. Actual time points for B_j=1/j, censoring rate=25%, n= 25

CHAPTER V

FUTURE WORK

Mean imputation appears to be an efficient method for fitting right censored data when used with SIR with regularizations. However, more simulations need to be done to test this. A larger amount of predictors, such as p=10,000, needs to be tested to simulate real life microarray data. We also need to examine the performance of the sparse ridge estimator with correlated predictors using the various methods for handling right censored data. There are various censoring schemes that need to be studied as well. Finally, a real life data set needs to be examined.

More work needs to be done on the SIR with regularizations model to improve the prediction accuracy. It appears from the simulations that mean imputation method provided better results than that of the uncensored data. The tests for measures of fit and prediction between uncensored data and censored data may also deserve further attention to clear this up.

The multiple imputation method could also be improved. The current method keeps the mean function Y using the marginal distribution. Better results may be achieved by using the conditional distribution instead. This could possibly be achieved by imputing the residual instead of actual values.

REFERENCES

1. Klein JP, Moeschberger ML. Survival Analysis. Second ed. New York: Springer; 2003.

2. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. Second ed. New York: Wiley; 2002.

3. Datta S, Le-Rademacher J. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics. 2007;63(1):259-71.

4. Teng X, Xiao H. Perspectives of DNA microarray and next-generation DNA sequencing technologies. Sci China C Life Sci. 2009;52(1):7-16.

5. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002;347(25):1999-2009.

6. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008;14(8):822-7. PMCID: 2667337.

7. Steidl C, Lee T, Shah SP, Farinha P, Han G, Nayar T, et al. Tumor-associated macrophages and survival in classic Hodgkin's lymphoma. N Engl J Med. 2010;362(10):875-85.

8. Li K. Sliced Inverse Regression for Dimension Reduction. Journal of the American Statistical Association. 1991;86(414):316-27.

9. Li L, Yin X. Sliced inverse regression with regularizations. Biometrics. 2008;64(1):124-31.

10. Li K, Wang J. Dimension reduction for censored regression data. The Annals of Statistics. 1999;27(1):1-23.

11. Wen X, Cook RD. New approaches to model-free dimension reduction for bivariate regression. Journal of Statistical Planning and Inference. 2009;139:734-48.

12. Li L, Lu W. Sufficient dimension reduction with missing predictors. Journal of the American Statistical Association. 2008;103(482):822-31.

13. Nguyen TS, Rojo J. Dimension reduction of microarray gene expression data: the accelerated failure time model. J Bioinform Comput Biol. 2009;7(6):939-54. PMCID: 2796584.

14. Robins JM, Rotnitzky A, editors. Recovery of information and adjustment for dependent censoring using surrogate markers. Boston: Birkhauser; 1992.

15. Koul H, Susarla V. Regression analysis with randomly right-censored data. Annals of Statistics. 1981;9:1276-88.

16. Cook RD. Dimension reduction and graphical exploration in regression including survival analysis. Stat Med. 2003;22(9):1399-413.

17. Cook RD. Regression Graphics: Ideas for Studying Regressions Through Graphics. New York: Wiley; 1998.

18. Cook RD. Re-weighting to achieve elliptically contoured covariates in regression. Journal of the American Statistical Association. 1994;89:592-600.

19. Cook RD, S. W. Discussion of Li. Journal of the American Statistical Association. 1991;86:328-32.

20. Li KC. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's Lemma. Annals of Statistics. 1992;87:1025-39.

21. Li L, Cook RD. Partial inverse regression. Biometrika. 2007;94(3):615-25.

22. Cook RD. Testing predictor contributions in sufficient dimension reduction. Annals of Statistics. 2004;32:1061-92.

23. Eaton M. A characterization of spherical distributions. Journal of Multivariate Analysis. 1986;20:272-6.

24. Golub GH, Heath M. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics. 1979;21:215-23.

25. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B. 1996;58:267-88.

26. Akaike H, editor. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory; 1973; Budapest.

27. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978;6:461-4.

28. Shi P, Tsai C-L. Regression model selection- A residual likelihood approach. Journal of the Royal Statistical Society, Series B. 2002;64:237-52.

29. Zou H, Hastie T, Tibshirani R. On the "Degrees of Freedom" of the Lasso: Department of Statistics, Stanford University; 2004 Contract No.: Document Number|.

30. Zhu L, Miao B, Peng H. On sliced inverse regression with large dimensional covariates. Journal of the American Statistical Association. 2006;101:630-43.

CURRICULUM VITAE

Name: **Daniel W. Riggs**

Address: Department of Bioinformatics and Biostatistics
School of Public Health, University of Louisville
Louisville, Kentucky 40202

Email: dwrigg01@louisville.edu

Education: B.A. in Mathematics, Minor in Chemistry
Keuka College, Keuka Park, NY

Publications:

1. Wang G, Hamid T, Keith RJ, Zhou G, Partridge CR, Xiang X, Kingery JR, Lewis RK, Li Q, Rokosh DG, Ford R, Spinale FG, **Riggs DW**, Srivastava S, Bhatnagar A, Bolli R, Prabhu SD. Cardioprotective and antiapoptotic effects of heme oxygenase-1 in the failing heart. Circulation. 121(17): 1912-25, 2010.

2. Sithu SD, Srivastava S, Siddiqui MA, Vladykovskaya E, **Riggs DW**, Conklin DJ, Haberzettl P, O'Toole TE, Bhatnagar A, D'Souza SE. Exposure to acrolein by inhalation causes platelet activation. Toxicol Appl Pharmacol. (In Press) 2010.