November 2010

Survival Analysis of Microarray Data With Microarray Measurement Subject to Measurement Error

Juan Xiong
The University of Western Ontario

Supervisor
Dr. Wenqing He
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

SURVIVAL ANALYSIS OF MICROARRAY DATA WITH MICROARRAY MEASUREMENT SUBJECT TO MEASUREMENT ERROR

(Spine title: Survival Analysis of Microarray Data with Measurement Error)

(Thesis format: Monograph)

by

Juan Xiong

Graduate Program
in
Statistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

Dr. Wenqing He

Examiners

Dr. Joseph Beyene

Dr. Duncan Murdoch

Dr. John Koval

Dr. Xingfu Zou

The thesis by

Juan Xiong

entitled:

SURVIVAL ANALYSIS OF MICROARRAY DATA WITH MICROARRAY MEASUREMENT SUBJECT TO MEASUREMENT ERROR

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Date

Chair of the Thesis Examination Board

ABSTRACT

Microarray technology is essentially a measurement tool for measuring expressions of genes, and this measurement is subject to measurement error. Gene expressions could be employed as predictors for patient survival, and the measurement error involved in the gene expression is often ignored in the analysis of microarray data in the literature. Efforts are needed to establish statistical method for analyzing microarray data without ignoring the error in gene expression.

A typical microarray data set has a large number of genes far exceeding the sample size. Proper selection of survival relevant genes contributes to an accurate prediction model. We study the effect of measurement error on survival relevant gene selection under the accelerated failure time (AFT) model setting by regularizing weighted least square estimator with adaptive LASSO penalty. Simulation results and real data analysis show that ignoring measurement error will affect survival relevant gene selection. Simulation-Extrapolation (SIMEX) method is investigated to adjust the impact of measurement error on gene selection. The resulting model after adjustment is more accurate than the model selected by ignoring measurement error.

Microarray experiments are often performed over a long period of time, and samples can be prepared and collected under different conditions. Moreover, different protocols or methodology may be applied in the experiment. All these factors contribute to a possibility of heteroscedastic measurement error associated with microarray data set. It is of practical importance to combine microarray data from different labs or platforms. We construct a prediction AFT model using data with heterogeneous covariate measurement error. Two variations of the SIMEX algorithm are investigated to adjust the effect of the mis-measured covariates. Simulation re-

sults show that the proposed method can achieve better prediction accuracy than the naive method.

In this dissertation, the SIMEX method is used to adjust for the effects of covariate measurement error. This method is superior to other conventional methods in that it is not only more robust to distributional assumptions for error prone covariates, it also offers marked simplicity and flexibility for practical use. To implement this method, we developed an R package for general users.

Keywords: Accelerated failure time model, Measurement error, Microarray, Prediction, Simulation and extrapolation method, Survival analysis, Variable selection.

CO-AUTHORSHIP STATEMENT

All materials presented in this thesis were obtained under the supervision of Dr.Wenqing He. Dr.He provided valuable insight in the ideas behind the materials. The majority of the work associated with implementing this research was done by myself.

This dissertation is dedicated to my family.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my Dissertation supervisor, Dr. Wenqing He. His encouragement, advice and support over the past five years have been tremendous, which I appreciate from the bottom of my heart. This dissertation would not have been possible without his support and guidance. He has helped me develop my scientific acumen and has been a mentor in true sense.

I would also like to thank Dr. Joseph Beyene, Dr. Duncan Murdoch, Dr. John Koval and Dr. Xingfu Zou for serving as members of my dissertation committee, reading my dissertation thoroughly and providing very helpful comments and suggestions on my research. I also want to thank Dr. Hao Yu for his technical and career advice during my graduate study.

I would like to extend my deep appreciation to the Department of Statistical and Actuarial Sciences at the University of Western Ontario for providing great academic and logistic support for my graduate study. I would like to thank knowledgable faculty members for their excellent lectures and thank friendly staff Jennifer, Lisa and Jane for their assistance. I appreciate many colleagues in my department, especially Paul, for his friendship and support.

Personally, I would like to express thanks to my parents for their continued love and encourage throughout my entire life. I would also like to express thanks and gratitude to my uncle for his love and support during my time in graduate school. Finally, I own my thanks to Xin Song and Juanna Yang because I could not make it through without their love, encouragement and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Microarray Technology

Microarray is an innovative technology that facilitates the analysis of thousands of gene expressions simultaneously (Golub et al., 1999; Schena et al., 1995). The use of this technology calls for a multidisciplinary efforts from biological, statistical sciences and bioinformatics community. There have been various microarray studies carried out in recent years.

1.1.1 Microarray Data Examples

1. DLBCL Data

   Diffuse large-B-cell lymphoma (DLBCL) is the most common type of lymphoma in adults. The data set of Rosenwald et al. (2002) consists of 7399 gene expression profiles across 240 patients with untreated DLBCL. Median survival time was 2.8 years and 138 patients died during this follow up period. Gene expression can be used as predictor of patient survival time after chemotherapy. In Rosenwald et al. (2002), the authors used 17 genes to build a Cox regression model to predict the survival time of these DLBCL patients.

2. Golub Data

   The data set of Golub et al. (1999) consists of 72 bone marrow samples obtained from acute leukemia patients. Ribonucleic acid (RNA) prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide

microarray, where 6817 human gene expression profiles were measured. By relying on gene expression, Golub et al. (1999) were able to make distinguishable identification between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without any prior knowledge of these classes.

3. Breast Cancer Data

   Van de Vijver et al. (2002) studied a cohort of 295 young patients with breast cancer. This study utilized only tumor specimens which were less than 5 cm. All patients were treated using modified radical mastectomy or breast-conserving surgery and assessed annually for a period of at least five years. The median follow-up among all 295 patients was 6.7 years. The authors reported that prediction model based on 70 gene expression profiles performed best for the appearance of distant metastases during the first five years after treatment.

These studies give examples of the applications of microarray technology to measure gene expression, and utilize gene expression as covariate to build statistical model for survival prediction, group classification, etc.

Microarray is a breakthrough technology that allows us to analyze completed genetic variations in entire genome level (Golub et al., 1999). Introduced in the mid-1990s (Schena, 1995; Lockhart, 1996; DeRisi, 1997), microarray has ever since been applied to a number of diverse areas, such as nutrition research (DellaPenna, 1999), drug discovery (Debouck, 1999), environmental health research (Nuwaysir, 1999) and cancer diagnostic (Golub, 1999; Dudoit, 2002). With the advent of new technologies and more rapid methods of analysis, microarray technology has the potential to become increasingly popular tool in many new areas in the future. Comprehensive reviews of microarray technology and data analysis can be found in Duggan et al. (1999) and Quackenbush (2001, 2002).

Microarray technology is essentially a measurement tool for measuring biological features such as gene expression. However, the measurement of gene expression may

be subject to error, and those measurement errors are often ignored in the microarray data analysis in the literature. We investigate this issue in this thesis.

1.1.2 Genomes and Microarray Experiment

To provide a better picture of gene expression data, we give a brief background of the gene and procedures of microarray experiment. Genes are hereditary units that are composed of deoxyribonucleic acid (DNA) sequences organized in chromosomes in the cell nucleus (Russell et al., 2009). The DNA sequences control the generation of messenger ribonucleic acids (mRNA) through a process called transcription. mRNA encodes amino acids that subsequently form proteins through a process called translation. The proteins carry out the designated function of a particular gene. Thus, the structural and functional features of cells and tissues are determined by the simultaneous, selective, and differential expressions of thousands of genes. The type and amount of protein present in the cells determine the phenotypes of the cells, such as cancer or normal cells.

A microarray is a solid substrate (usually glass) upon which many different cDNAs have been spotted in specific locations in a grid pattern. mRNA from a tissue sample of interest is extracted and the reverse transcription is applied to synthesize cDNA, and labels the cDNA with fluorescent dye or radioactive nucleotides. This labeled cDNA is then hybridized to cDNA immobilized on the array. The labeled cDNA binds to its complementary sequence approximately proportional to the amount of each mRNA transcript in a sample. The amount of radioactivity or fluorescence can be measured, allowing estimation of the amount of mRNA for each transcript in the sample. Once a microarray experiment has been conducted, the arrays are scanned by a confocal laser microscope. The images from the scanner are processed to extract the spot intensities and background spot intensities. The gene expression levels are measured by the normalized ratio of the fluorescence intensity of the test sample and the reference sample for a certain gene. Although microarray data can be generated

from multiple platforms, it is worth to point out that our methods are not limited to specific platforms, as long as the generated data are continuous variable. In the case of the SNP microarray platform, a log transformation is sufficient to convert discrete variables into continuous variables, so to satisfy prerequisite of our methods.

1.2 Survival Analysis

Survival analysis or time-to-event data analysis is a branch of statistics which emphasizes on developing statistical methods for analyzing the time to an event of interest, often referred to as survival time or failure time (Lawless, 2003). Survival analysis is an important topic in various scientific fields, such as biomedical sciences, economics and engineering. One special characteristic of survival data is that the survival time may be subject to censoring. Censoring generally occurs because subject may be lost in the follow-up during the study period or withdraw from the study due to death or some other reasons. In this thesis, we will focus on right censoring. Suppose that we have a random sample of n subjects, $i = 1, \ldots, n$. Let T_i be the survival time and C_i be the censoring time. Usually, it is assumed that the survival time is independent of the censoring time, or at least that they are independent given certain covariates. We only observe $\min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator, with $\delta_i = 0$ if subject i is censored and $\delta_i = 1$ if the survival time T_i is observed.

The survival function and hazard function are essential to survival analysis. The survival function, $S(t)$, describes the probability that the random variate T exceeds the specified time t and is given by

$$S(t) = Pr(T \geq t) \text{ for } t \geq 0,$$

where $S(t)$ is non-increasing and left continuous. At time $t = 0$, $S(0) = 1$ and at $t = \infty$, $S(\infty) = 0$.

The hazard function, $h(t)$, gives the instantaneous rate of failure at time t on condition that individual surviving up to t and is given by

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

In contrast to the survival function, the hazard function focuses on failure given survival up to a certain time point. The hazard function can be used to identify the form of model. The relationship between $S(t)$ and $h(t)$ is

$$S(t) = \exp\left\{ - \int_0^t h(u) du \right\}.$$

1.2.1 Cox Proportional Hazards Model

Cox proportional hazards (PH) model is one of the most popular semiparametric regression model in survival analysis (Cox, 1972). The Cox PH model is given by

$$h(t|\mathbf{X}_i) = h_0(t) \exp\left(\mathbf{X}_i' \boldsymbol{\beta}_x\right),$$

where $h_0(t)$ is a unspecified baseline hazard function; $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ is p dimensional vector of covariates and $\boldsymbol{\beta}_x$ is the associated vector of unknown regression coefficients, which can be estimated by maximizing the partial likelihood (Cox, 1975)

$$L(\boldsymbol{\beta}_x) = \prod_{i=1}^{n} \left(\frac{e^{\mathbf{X}_i' \boldsymbol{\beta}_x}}{\sum\limits_{j \in R_i} e^{\mathbf{X}_j' \boldsymbol{\beta}_x}} \right)^{\delta_i},$$

where R_i is the risk set of subjects at time t_i given by

$$R_i = \{ j : T_j \geq t_i \}.$$

1.2.2 Accelerated Failure Time Model

Another attractive alternative to the Cox PH model is the accelerated failure time (AFT) model (Kalbfleisch and Prentice, 1980). The AFT model relates the logarithm of the survival time linearly to covariates and is given by

$$Y_i = \beta_0 + \mathbf{X}_i' \boldsymbol{\beta}_x + \epsilon_i, \tag{1.1}$$

where $Y_i = \log(T_i)$, ϵ_i is the error term. The parametric AFT model specifies the distribution of ϵ_i up to parameters $\boldsymbol{\alpha}$. Common choice of distribution include the Weibull, exponential, Gaussian, logistic, log-normal and log-logistic distribution (Lawless, 2003). The semiparametric AFT model does not make any assumption on the distribution of ϵ_i.

The Cox PH model and AFT model are intended for different types of comparisons (i.e., the Cox PH model compares the hazard functions whereas the AFT model compares the survival times). In the Cox PH model, the covariates are multiplicative to the hazard function and remains constant over time, whereas in the AFT model, the covariates are multiplicative to the survival times. Compared to the Cox PH model, the results of AFT models are easier to interpret due to its direct modeling of the survival time (Reid, 1994). Also, when there is no censoring, the AFT models reduce to ordinary generalized linear regression models. AFT models have been studied extensively in the literature: Miller (1976) and Buckley and James (1979) modified the least square estimate equation to account for the censored response variable; Tsiatis (1990) and Ying (1993) proposed the rank based estimator; and Stute (1996) investigate the weighted least square estimator. In this thesis, we will focus on AFT models.

1.3 Survival Analysis with Microarray Data

Central to the application of microarrays in biomedical and genomic research is to reveal different gene expression profiles under different medical or treatment scenarios. For example, in cancer research, gene expression profiles can help further understanding of cancer at the genetic or molecular level.

With the microarray features of each patient, we can have a patient survival prediction model that is biologically meaningful. Microarray experiments generate large data sets, to which many biological researchers may not be accustomed (Page et al., 2003). A typical microarray data set has a large number of genes far exceeding the sample size. Other than the high dimensionality of the genes, the expression levels of genes are often highly correlated. As a pre-procedure, one needs to identify survival relevant genes from a large set of candidates produced by microarray experiments. After identifying a subset of genes with the most predictive power to the survival outcome of the patient, one can combine them with patient specific covariates to build a prediction model for future patients' survival outcomes. Simply put, proper selection contributes to an accurate prediction model that are both clinically and biologically meaningful, and could lead to better treatment choice for patients.

1.3.1 Variable Selection

Variable selection is fundamental in statistical modeling and data analysis. Traditional variable selection approaches include Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978), and risk inflation criterion (RIC) (Foster and George, 1994). Recent high throughput technologies have generated data where the number of covariates is significantly larger than the sample size. Typical examples include microarray data, text categorization and image retrieval. New features of these data present a direct challenge to standard variable selection methods.

Literature on variable selection in high dimensional models is growing quickly. Fan and Lv (2010) presented a review on variable selection in high-dimensional modeling. Here we introduce two popular methods: least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and the adaptive LASSO (Zou, 2006).

1.3.1.1 LASSO

Tibshirani (1996) proposed the popular shrinkage regression technique that could select variables and estimate the regression coefficient simultaneously. The LASSO estimate is defined as

$$\widehat{\boldsymbol{\beta}}_x(lasso) = \underset{\boldsymbol{\beta}_x}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_{x_j} \right)^2 + \frac{\gamma}{n} \sum_{j=1}^{p} \left| \beta_{x_j} \right| \right\},$$

where Y_i is the response for subject i, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ is p dimensional covariate vector and $\boldsymbol{\beta}_x$ is the associated vector of unknown regression coefficients. γ is a penalty parameter determined by cross-validation. It shrinks a number of coefficients to zero, thus can be used for variable selection. Efron et al. (2004) published least angle regression algorithm which can be employed to solve LASSO estimate.

1.3.1.2 Adaptive LASSO

Fan and Li (2001) showed that the LASSO penalty produces biased coefficients estimates. To overcome the bias, Zou (2006) proposed the adaptive LASSO that has oracle properties:“ it can not only selects significant variables consistently but also performs as efficient as if the true model was known, a property not enjoyed by the LASSO.” Hence, the adaptive LASSO method is an ideal one for variable selection. The adaptive LASSO estimate is defined as

$$\widehat{\boldsymbol{\beta}}_x = \underset{\boldsymbol{\beta}_x}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_{x_j} \right)^2 + \gamma \sum_{j=1}^{p} v_j \left| \beta_{x_j} \right| \right\},$$

where $\mathbf{v} = (v_1, \ldots, v_p)$ is a known weight vector. The weight vector can be constructed as $\mathbf{v} = 1/|\boldsymbol{\beta}^*|^\tau$, $\tau > 0$, where $\boldsymbol{\beta}^*$ needs to be a root-n-consistent estimator of $\boldsymbol{\beta}_x$, such as the ordinary least square estimate.

1.3.2 Variable Selection for Survival Outcomes

In the past few years, some of the variable selection procedures in linear regression analysis have been extended to the censored survival data analysis in the presence of high dimensional predictors. For example, Tibshirani (1997) developed a regularized Cox regression by minimizing L_1 LASSO penalty to the partial likelihood; Faraggi and Simon (1998) proposed a Bayesian variable selection method for the Cox model; Li and Luan (2003) investigated the L_2 penalized estimation of the Cox model using kernel; and Gui and Li (2005) introduced a threshold gradient descent regularization estimation method. For the AFT model, Schmid and Hothorn (2008) presented a boosting algorithm for fitting the parametric AFT model. For the semiparametric AFT model, Huang et al. (2006) investigated the LASSO regularization for estimation and variable selection in the AFT model based on the inverse probability of censoring weights method; Huang and Harrington (2005), Datta et al. (2007) used the LASSO regularized Buckle-James method for the AFT model; and Cai et al. (2008) developed variable selection for the AFT model by the LASSO regularized rank based estimator.

1.4 Measurement Error Models

In many biomedical studies, it is often the case that some covariates can not be measured accurately, which leads to measurement error models or errors-in-variable models in the literature (Carroll et al., 2006). Measurement error arises for many

reasons. Sometimes it is due to covariate nature, for instant, blood pressure. In other cases the patient consent and cost may prevent precise observation of the covariate. It is well known that ignoring measurement error in covariate leads to biased estimate of the covariate effects and consequently affects inference (Fuller, 1987; Carroll et al., 2006).

In the past several decades, a great deal of research has been done on measurement error models. Fuller (1987) summarized a detailed discussion of statistical method for linear measurement error models while Carroll et al. (2006) provided systematic guide on dealing with nonlinear measurement error models. Recently, the study of measurement error models has become an increasingly popular theme in nonparametric measurement error area (Delaigle and Meister, 2007; Carroll et al., 2009).

1.4.1 Models for the Measurement Error Process

Specifying the model for the measurement error process is fundamental for analyzing measurement error problems. There are a number of measurement error models reported in the literature. The general two models are the classical additive measurement error model and the Berkson error model. Assume we have response Y_i and two types of covariates, $\mathbf{Z}_i$ consists of the covariates measured without error, and $\mathbf{X}_i$ represents those that can not be observed exactly for subjects $i = 1, \ldots, n$. Instead of observing $\mathbf{X}_i$, we observe its contaminated version $\mathbf{W}_i$.

The classical additive model assumes that

$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i, \tag{1.2}$$

the Berkson model assumes that

$$\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i,$$

where measurement error $\mathbf{U}_i$ has multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{\Sigma}_{u_i}$. The measurement errors can be either homoscedastic or heteroscedastic. If the variances $\mathbf{\Sigma}_{u_i}$ are the same for all subjects, it is called homoscedastic measurement error. Otherwise it is heteroscedastic. The measurement errors are mutually independent and are independent of $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i\}$. The use of different model forms is decided by the nature of study.

It is important to identify distinct measurement error mechanisms. Measurement error is non-differential when the distribution of Y_i given $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ is the same as the distribution given $(\mathbf{X}_i, \mathbf{Z}_i)$. In other words, $\mathbf{W}_i$ contains no information about Y_i other than what is available in $(\mathbf{X}_i, \mathbf{Z}_i)$. $\mathbf{W}_i$ is called a surrogate for $\mathbf{X}_i$. Otherwise, measurement error is differential. This thesis focuses on nondifferential measurement error models.

1.4.2 Methods for Measurement Error Analysis

A large number of methods has been proposed to deal with measurement error problems, including likelihood based methods (Stefanski and Carroll, 1990); score function methods (Kukush and Schneeweiss, 2004); Bayesian methods (Clayton, 1992); and semiparametric and nonparametric methods (Huang and Wang, 2000). Two widely applicable methods for measurement error analysis are regression calibration and simulation extrapolation (SIMEX).

1.4.2.1 Regression Calibration

Regression calibration is commonly applied to account for measurement error (Carroll et al., 2006). It reduces bias in the estimate of parameter and enjoys simplicity in practical use. This method replaces the unobserved covariate $\mathbf{X}_i$ by the conditional expectation of $\mathbf{X}_i$ given the observed covariates denoted by $\mathrm{E}(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)$. Then, it runs the standard analysis based on this approximation. Either the bootstrap

or sandwich method is needed to adjust the standard errors of the parameter estimates to account for the variation induced by estimation of parameters in modeling $E(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i)$. The key issue with this method is how to best estimate this expectation.

1.4.2.2 Simulation Extrapolation Method

Cook and Stefanski (1994) introduced a SIMEX approach for estimating and correcting bias due to measurement error. The general idea of the SIMEX method is to generate additional data sets with increasingly larger measurement error, estimate the trend of the effect of the measurement error on the estimation of the parameter of interest. We then extrapolate the trend back to the case of no measurement error. The major advantage of the SIMEX method is its easy implementation and robustness to distributional assumptions for error prone covariates. See section 2.2.3 for the detailed description.

1.5 Survival Analysis with Measurement Error

A well-known challenge associated with survival data analysis is to find an appropriate way of handling measurement errors which are frequently present in covariates. It is known that many biomarkers, such as blood pressure and CD4 counts are often subject to measurement error. Great research effort has already been undertaken to explore effective ways of handling covariate measurement error for survival data.

Prentice (1982) first considered the regression calibration method to adjust the impact of measurement error in covariates for the Cox PH model; Clayton (1992) modified Prentice's approach by using the regression calibration within each risk set; Zhou and Pepe (1995) proposed a nonparametric method for discrete covariates with measurement error; later, Zhou and Wang (2000) extended this method to continuous covariates by applying a kernel smoothing. Hu, Tsiatis and Davidian (1998)

developed a full likelihood approach to account for measurement error in the Cox regression model with a single covariate; Nakamura (1992) and Buzas (1998) applied the corrected score function to the Cox PH model when the measurement errors are additive and normally distributed; Huang and Wang (2000) modified the score function and proposed a nonparametric approach to estimate the parameter of the Cox regression model when replicates of $\mathbf{W}_i$ are available for each subject; and Gimenez et al. (1999, 2006) have applied the corrected score approach in their investigation of inference methods under Weibull regression models.

For multivariate survival analysis with mismeasured covariates, Li and Lin (2000) used the expectation-maximization algorithm to calculate the nonparametric maximum likelihood estimates for clustered survival data with covariates subject to frailty measurement error; Hu and Lin (2004) proposed semiparametric regression methods for multivariate failure times; and Green and Cai (2004) explored the SIMEX method in dealing with measurement error effect on multivariate failure time model.

In the above literature, all the works are focused on the Cox PH models with covariates subject to measurement error. With AFT models, Tseng et al. (2005) considered the joint modeling of failure time and longitudinal data under the AFT assumption when covariates are assumed to follow a linear mixed effects model with measurement errors; He et al. (2007) applied the SIMEX method to adjust the effect of mismeasured covariates in the accelerated failure time model; Yu and Nan (2009) considered the regression calibration estimation method for the semiparametric AFT model with covariates subject to measurement error.

1.6 Objective of This Thesis

Like other measurement tools, the gene expression levels measured from microarrays have measurement error. Throughout the microarray experiment process, measurement error might be produced from various sources, including errors associated with

the fluorescent signal, slide hybridization, image creating and reading, etc. As commonly acknowledged, the presence of measurement error leads to substantially biased and inconsistent parameter estimates. Thus this leads to invalid hypothesis test and mask the feature of the data.

To the best of our knowledge, most existing variable selection procedures are limited to directly observed predictors. Variable selection for measurement error data has not been systematically studied yet. Liang and Li (2009) proposed a class of variable selection procedures for partially linear measurement error models by using penalized least squares and penalized quantile regression. Ma and Li (2010) discussed variable selection for general parametric and semiparametric measurement error models via penalized estimating equations. In this thesis, we study the impact of measurement error on survival relevant gene selection under the AFT model.

Microarray experiments are often performed over a long period of time, and samples can be prepared and collected under different conditions. Moreover, different protocols or methodology may be applied in the experiment. All these factors contribute to a possibility of heteroscedastic measurement error associated with microarray data set. It is of practical importance to combine microarray data from different labs or platforms which presents a natural way to increase sample size so that reliable statistical analysis can be conducted. In this thesis, we will investigate prediction of survival time under AFT model with gene expression subject to heteroscedastic measurement error.

An outline of this thesis is as follows. In Chapter 2, we study the effect of measurement error on survival relevant gene selection under the AFT model setting by regularizing the weighted least square estimator with an adaptive LASSO penalty. In Chapter 3, we consider prediction of AFT model using data with heteroscedastic covariate measurement error. Two variations of the SIMEX algorithm are investigated to adjust the effect of the mis-measured covariates, and a best linear prediction is employed to predict the corresponding value of the unobserved covariates of fu-

ture observation. We develop an R package `simexaft` to adjust biases induced by covariate measurement error under AFT models and illustration is given in Chapter 4. Concluding remarks and discussion on future work are presented in Chapter 5. The source code for the R package and some technical details are included in the appendix.

Chapter 2

Survival Relevant Gene Selection in Microarray Data Analysis with Gene Expression Subject to Measurement Error

2.1 Introduction

Microarray technology has become a very popular tool for investigating molecular features of different clinical outcomes (Golub et al., 1999; Dudoit et al., 2002; Rosenwald et al., 2002). In survival analysis, microarray data are commonly used for building a prediction model of survival outcomes based on the gene expression profiles (e.g., survival times of patients). However, because of their unique features, microarray data must be analyzed carefully. For instance, the number of genes far exceeds the sample size in many microarray data sets. Other than the high dimensionality of the genes, the gene expression levels are often highly correlated. Therefore, we need to identify a subset of genes that are significantly correlated with the survival outcomes, and combine patient specific covariates together to build a prediction model for future patients' survival outcomes (Li, 2008).

There has been extensive research on variable selection and estimation methodologies in the presence of high dimensional predictors. Examples include bridge regression (Frank and Friedman, 1993); non-negative garrote (Breiman, 1995); least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996); smoothly clipped absolute deviation (Fan and Li, 2001); gradient directed regularized method (Friedman and Popescu, 2004); the boosting algorithm (Buhlmann and Yu, 2003); the elastic net (Zou and Hastie, 2005); the adaptive LASSO (Zou, 2006); and the Dantzig selector (Candes and Tao, 2007). Fan and Lv (2010) gave a comprehensive overview of several of these high dimensional variable selection methods.

It becomes more complicated if the goal is to predict survival time with high dimensional gene expressions when the survival time is censored. As a consequence, direct employment of traditional survival analysis techniques is difficult to obtain accurate parameter estimates. See section 1.3.2 for the literature review on variable selection methods for combining high-dimensional covariates to predict failure time outcomes.

Microarray technology allows for the measurement of the expressions of thousands of genes simultaneously. Like many other quantitative tools, gene expressions are subject to measurement errors. It is commonly acknowledged that ignoring measurement error could lead to substantially biased estimates of the regression parameters (Fuller, 1987; Carroll et al., 2006). This leads to incorrect results for statistically identifying survival relevant genes. Consequently, it is essential to investigate survival relevant gene selection when the gene expressions are subject to measurement errors.

Huang et al. (2006) used censoring weights for adjusting the LASSO penalized least squares loss function for variable selection in AFT model. Because of the L_1 penalty structure, this method has the advantage of carrying variable selection and estimating parameters simultaneously. The adaptive LASSO (Zou, 2006) is similar to LASSO in that it has retained the near-minimax optimality and can be solved by the least angle regression algorithm (Efron et al., 2004). Furthermore, it enjoys the oracle properties as mentioned in Section 1.3.1.2.

In this chapter, we study the effect of the measurement error on survival relevant gene selection in the AFT model by regularizing the weighted least square estimator with the adaptive LASSO penalty. The bootstrap method, which samples with replacement from the original observations, is employed to estimate variances. The simulation extrapolation (SIMEX) method is explored to adjust the effect of measurement error on variable selection.

2.2 Model Framework

Survival regression models are mainly used to identify covariates that are significantly related to the survival times. With microarray data, we use gene expressions to build the survival model to find gene expressions that significantly predict the survival time of a patient. Typically, the gene expression measurements from microarray experiments have measurement errors. Let $\mathbf{W}_i$ be the measured gene expression and $\mathbf{X}_i$ be the true gene expression, which is usually unavailable, for the ith subject. The relationship between $\mathbf{W}_i$ and $\mathbf{X}_i$ could be assumed through the most commonly used additive measurement model given by equation (1.2), where measurement error $\mathbf{U}_i$ follows a normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{u_i}$. By ignoring $\mathbf{U}_i$, the AFT model (1.1) will be the naive model given by

$$Y_i = \beta_0 + \mathbf{W}_i'\boldsymbol{\beta}_w + \epsilon_i. \tag{2.1}$$

The estimates for $\boldsymbol{\beta}_w$ will attenuate from the true $\boldsymbol{\beta}_x$ in the model (1.1) (Fuller, 1987); hence, the survival relevant variable selection will be affected (Carroll et al., 2006; He et al., 2007).

2.2.1 The Adaptive LASSO Regularized Inverse Probability of Censoring Weight (IPW) Method

One problem in utilizing the adaptive LASSO for survival relevant gene selection is that the survival times are not available for censored observations. Thus, the least square term in adaptive LASSO has to be modified for survival data. The IPW method is a popular choice to overcome this problem (Huang et al., 2006).

The Kaplan-Meier estimator of the distribution function of survival time changes

only at the uncensored point by the jumps $\pi'_{ni}s$ given by

$$\pi_{n1} = \frac{\delta_{(1)}}{n},$$

and

$$\pi_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, \ i = 2, \ldots, n,$$

where $\delta_{(1)}$, $\delta_{(2)}$, $\ldots$, $\delta_{(n)}$ are the censoring indicators for the corresponding ordered logarithm of survival times $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. The weighted least square estimator defined by Stute (1996) is the set of values for $(\beta_0, \boldsymbol{\beta}_x)$ that minimizes

$$\ell(\beta_0, \boldsymbol{\beta}_x) = \frac{1}{2} \sum_{i=1}^{n} \pi_{ni} \left(Y_{(i)} - \beta_0 - \mathbf{X}'_{(i)} \boldsymbol{\beta}_x \right)^2,$$

where $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$, $\ldots$, $\mathbf{X}_{(n)}$ are covariates of the corresponding ordered $Y_{(i)}$'s.

We first adjust $\mathbf{X}_{(i)}$ and $Y_{(i)}$ by their $\pi_{ni}-$weighted means, respectively,

$$\mathbf{X}_{\pi(i)} = \pi_{ni}^{1/2} \left(\mathbf{X}_{(i)} - \bar{\mathbf{X}}_{\pi} \right)$$

and

$$Y_{\pi(i)} = \pi_{ni}^{1/2} \left(Y_{(i)} - \bar{Y}_{\pi} \right),$$

where

$$\bar{\mathbf{X}}_{\pi} = \sum_{i=1}^{n} \pi_{ni} \mathbf{X}_{(i)} / \sum_{i=1}^{n} \pi_{ni}$$

and

$$\bar{Y}_{\pi} = \sum_{i=1}^{n} \pi_{ni} Y_{(i)} / \sum_{i=1}^{n} \pi_{ni}.$$

By replacing the original sample $(Y_i, \mathbf{X}_i, \delta_i)$ with the weighted centered values $(Y_{\pi(i)}, \mathbf{X}_{\pi(i)})$,

the weighted least squares (LS) objective function becomes

$$\ell(\boldsymbol{\beta}_x) = \frac{1}{2}\sum_{i=1}^{n}\left(Y_{\pi(i)} - \mathbf{X}'_{\pi(i)}\boldsymbol{\beta}_x\right)^2.$$

Then, the adaptive LASSO regularized IPW estimator, $\widehat{\boldsymbol{\beta}}_x$, is the solution that minimizes

$$\ell(\boldsymbol{\beta}_x) = \frac{1}{2}\sum_{i=1}^{n}\left(Y_{\pi(i)} - \mathbf{X}'_{\pi(i)}\boldsymbol{\beta}_x\right)^2 + \gamma\sum_{j=1}^{p} v_j\left|\beta_{x_j}\right|, \tag{2.2}$$

where γ is the adaptive LASSO penalty parameter and $\mathbf{v} = (v_1, \ldots, v_p)$ is a known adaptive LASSO weight component. We use $\mathbf{v} = 1/|\hat{\boldsymbol{\beta}}|$, where $\hat{\boldsymbol{\beta}}$ is the ordinary least square estimate of $\widehat{\boldsymbol{\beta}}_x$, as suggested by Zou (2006).

For the variance estimates of $\widehat{\boldsymbol{\beta}}_x$, we use the bootstrap method, where samples are generated with replacement from the original observations. The bootstrap estimator is computed with the same optimal value of γ as used on the original data. According to Theorem 2 in Zou (2006), $\widehat{\boldsymbol{\beta}}_x$ is asymptotically normal.

2.2.2 Variable Selection with Mismeasured Covariates

Throughout the microarray experiment process, measurement error might be produced from various sources (He et al., 2007). Consider the hypothesis test for evaluating the significance of the covariate, X_{ij}, given by

$$H_0 : \beta_{x_j} = 0 \text{ versus } H_A : \beta_{x_j} \neq 0, j = 1, \ldots, p.$$

The test statistic is given by

$$z_{x_j} = \frac{\widehat{\beta}_{x_j}}{\mathrm{SE}(\widehat{\beta}_{x_j})},$$

where $\widehat{\beta}_{x_j}$ and $\mathrm{SE}(\widehat{\beta}_{x_j})$ are the estimate and standard error estimate of β_{x_j}, respectively. Under H_0, z_{x_j} follows the standard normal distribution. If W_{ij} is observed

but X_{ij} is not observable, the naive test statistic is given by

$$z_{w_j} = \frac{\widehat{\beta}_{w_j}}{\mathrm{SE}(\widehat{\beta}_{w_j})},$$

where $\widehat{\beta}_{w_j}$ and $\mathrm{SE}(\widehat{\beta}_{w_j})$ are the estimate and standard error estimate of β_{w_j}, respectively.

Typically, $\widehat{\beta}_{w_j}$ will be attenuated from $\widehat{\beta}_{x_j}$ and its corresponding $\mathrm{SE}(\widehat{\beta}_{w_j})$ will be smaller than $\mathrm{SE}(\widehat{\beta}_{x_j})$ under certain conditons(Buzas et al., 2005). Hence, its corresponding p-value will be different from the p-value if X_{ij} is observable, often leading to incorrect decisions for the hypothesis test.

A common problem for microarray data involves the analysis of high dimensional gene expression data that are typically characterized by thousands of variables with few observations. That is, these data have a high degree of multicollinearity in the analysis. In general, the naive test statistic which ignores the measurement error and substitutes $\mathbf{W}_i$ for $\mathbf{X}_i$ in a test is not correct (Carroll et al., 2006). Here the $\mathbf{X}_i$ are highly correlated, under the multivariate distribution, the naive test of no effect due to any component of $\mathbf{X}_i$ is not correct. We will use the SIMEX method (Cook and Stefanski, 1994) to adjust the effect of measurement error on survival relevant gene selection.

2.2.3 Simulation Extrapolation Method

Here we describe the SIMEX method to correct the bias due to measurement error. More details are available in Carroll et al. (2006). Although the theory of the SIMEX method is not trivial, an example from simple linear regression can well illustrate the idea of this method. Suppose the response variable Y and the covariate X is modeled as

$$Y = \beta_0 + X\beta_x + \epsilon,$$

where ϵ has mean 0. Here X is normal distributed with variance σ_x^2. If replacing X with its observed measurement W, modeled by $W = X + U$, where U follows normal distribution with mean 0 and variance σ_u^2 , and is independent of ϵ and X, then the resulting least squares estimator $\hat{\beta}_x^*$ for β_x converges in probability to

$$\beta_x^* = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta_x,$$

instead of β_x. To see how the bias may be related to the degree of measurement error in X, we perturb W by adding additional error to create $W(b, \lambda) = W + \sqrt{\lambda}U_b$ where U_b is independently generated from a $\mathrm{N}(0, \sigma_u^2)$ distribution. Intuitively, if regressing Y over the perturbed version $W(b, \lambda)$, then the resulting estimator $\hat{\beta}_x(b, \lambda)$ would converge in probability to

$$\beta_x^*(b, \lambda) = \frac{\sigma_x^2}{\sigma_x^2 + (1 + \lambda)\sigma_u^2}\beta_x.$$

This expression indicates the dependence of the asymptotic bias on the magnitude of measurement error - the less degree of measurement error (equivalently, a smaller λ), the smaller asymptotic bias. In particular, if λ shrinks to 0, $\hat{\beta}_x(b, 0)$ recovers the naive estimator $\hat{\beta}_x^*$; if λ takes value -1, then the limit $\beta_x^*(b, -1)$ is identical to the true parameter β_x.

We consider two practical cases for the parameters in $\mathbf{\Sigma}_{u_i}$: (i) the parameters in $\mathbf{\Sigma}_{u_i}$ are given as fixed values; and (ii) the parameters in $\mathbf{\Sigma}_{u_i}$ are not known, but replicate measurements of $\mathbf{W}_i$ are available.

Suppose that one could estimate regression parameter $\boldsymbol{\theta}$ by solving an estimating equation given by

$$\mathbf{0} = \sum_{i=1}^{n} \varphi(\boldsymbol{\theta}; Y_i, \mathbf{X}_i). \tag{2.3}$$

Given an integer B and a sequence $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, ..., \lambda_M\}$ taken from $[0, \lambda_M]$,

the estimates and associated variances of the parameters $\boldsymbol{\theta}$ can be obtained by the following SIMEX algorithm.

2.2.3.1 Case I: The parameters in $\boldsymbol{\Sigma}_{u_i}$ are given as fixed values.

1. Simulation Step

   We generate the pseudo errors $\mathbf{U}_{bi} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{u_i})$ for $i = 1, \ldots, n$, $b = 1, \ldots, B$. For each $\lambda \in \boldsymbol{\Lambda}$, let the pseudo data be given by

   $$\mathbf{W}_i(b, \lambda) = \mathbf{W}_i + \lambda^{\frac{1}{2}} \mathbf{U}_{bi}.$$

   Note that $\text{E}(\mathbf{W}_i(b, \lambda)|\mathbf{X}_i) = \mathbf{X}_i$ and $\text{Var}(\mathbf{W}_i(b, \lambda)) = \boldsymbol{\Sigma}_{w_i}(b, \lambda) = \boldsymbol{\Sigma}_x + (1 + \lambda)\boldsymbol{\Sigma}_{u_i}$. When $\lambda = -1$, $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_x$. Hence, the mean square error, $\text{E}[(\mathbf{W}_i(b, \lambda) - \mathbf{X}_i)^2|\mathbf{X}_i]$, converges to zero as $\lambda \to -1$. This implies that $\mathbf{W}_i(b, \lambda)$ is the best measurement of $\mathbf{X}_i$. This is the most important property of the pseudo data (Carroll et al., 2006).

2. Estimation Step

   Given λ and b, we obtain the estimate $\widehat{\boldsymbol{\theta}}(b, \lambda)$ by solving equation (2.3) with $\mathbf{X}_i$ replaced by $\mathbf{W}_i(b, \lambda)$. The corresponding variance estimate, $\widehat{\boldsymbol{\Omega}}(b, \lambda)$, is the diagonal elements of the observed information matrix given by

   $$\left[-\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}'} \varphi(\boldsymbol{\theta}; Y_i, \mathbf{W}_i(b, \lambda))|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(b, \lambda)} \right]^{-1}.$$

   By averaging over b, we obtain

   $$\widehat{\boldsymbol{\theta}}(\lambda) = B^{-1} \sum_{b=1}^{B} \widehat{\boldsymbol{\theta}}(b, \lambda)$$

and

$$\widehat{\mathbf{\Omega}}(\lambda) = B^{-1} \sum_{b=1}^{B} \widehat{\mathbf{\Omega}}(b, \lambda).$$

Define $\widehat{\mathbf{S}}(\lambda) = (B-1)^{-1} \sum_{b=1}^{B} \left(\widehat{\boldsymbol{\theta}}(b, \lambda) - \widehat{\boldsymbol{\theta}}(\lambda) \right)^2$, the variance for the estimator $\widehat{\boldsymbol{\theta}}(b, \lambda)$ is given by

$$\widehat{\mathbf{\Omega}}(\lambda) - \widehat{\mathbf{S}}(\lambda).$$

3. Extrapolation Step

   Fit these average estimates to an extrapolation function of lambda and extrapolate $\widehat{\boldsymbol{\theta}}(\lambda)$ back to the case of no measurement error (i.e., $\lambda = -1$) to yield the SIMEX estimate $\widehat{\boldsymbol{\theta}}_{simex}$. The variance estimate for $\widehat{\boldsymbol{\theta}}_{simex}$ can be obtained by extrapolating $\widehat{\mathbf{\Omega}}(\lambda) - \widehat{\mathbf{S}}(\lambda)$ back to $\lambda = -1$.

   The most commonly used extrapolation functions are linear extrapolation function, quadratic extrapolation function and rational linear extrapolation function. The quadratic extrapolant function in the SIMEX method is generally a safe choice for most regression models (He et al., 2007).

2.2.3.2 Case II: The parameters in $\mathbf{\Sigma}_u$ are not known, but replicate measurements of $\mathbf{W}_i$ are available.

Consider the case where the measurement error covariance matrices are not known but replicate measurements of $\mathbf{W}_i$ are available. The procedures of the SIMEX algorithm described above can be applied except for the data simulation step. Devanarayan and Stefanski (2002) introduced this variation and named it empirical SIMEX.

Suppose we have $k_i \geq 2$ replicate measurements denoted by $\{\mathbf{W}_{i1}, \ldots, \mathbf{W}_{ik_i}\}$ for every subject i such that

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}, \quad i = 1, \ldots, n; \; j = 1, \ldots, k_i,$$

where $\mathbf{U}_{ij}$ are mutually independent from each other and independent of $\{Y_i, \mathbf{X}_i\}$. For fixed i, $\mathbf{U}_{ij}$ are independent and identically distributed random errors such that $\mathbf{U}_{ij} \overset{\text{iid}}{\sim} \text{MVN}(\mathbf{0}, \mathbf{\Sigma}_{u_i})$. Without knowing the covariance matrix $\mathbf{\Sigma}_{u_i}$, we cannot generate pseudo errors directly. However, we can obtain them by taking linear combination of the replicated measurements.

The empirical SIMEX algorithm is as follows

1. We generate $z_{b,i,j} \overset{\text{iid}}{\sim} N(0,1)$, $i = 1, \ldots, n$, $j = 1, \ldots, k_i$, $b = 1, \ldots, B$. We define $\bar{z}_{b,i,.} = \sum_{j=1}^{k_i} z_{b,i,j}/k_i$ and

$$c_{b,i,j} = \frac{z_{b,i,j} - \bar{z}_{b,i,.}}{\sqrt{\sum_{j=1}^{k_i} \left(z_{b,i,j} - \bar{z}_{b,i,.}\right)^2}},$$

   where $\sum_{j=1}^{k_i} c_{b,i,j} = 0$ and $\sum_{j=1}^{k_i} c_{b,i,j}^2 = 1$.

2. For $i = 1, \ldots, n$, $j = 1, \ldots, k_i$, $b = 1, \ldots, B$ and each $\lambda \in \mathbf{\Lambda}$, we define

$$\mathbf{W}_i(b, \lambda) = \bar{\mathbf{W}}_{i.} + \left(\frac{\lambda}{k_i}\right)^{\frac{1}{2}} \sum_{j=1}^{k_i} c_{b,i,j} \mathbf{W}_{ij},$$

   where $\bar{\mathbf{W}}_{i.} = \sum_{j=1}^{k_i} \mathbf{W}_{ij}/k_i$. It is easy to see that $\text{E}(\mathbf{W}_i(b,\lambda)|\mathbf{X}_i) = \mathbf{X}_i$ and $\text{Var}(\mathbf{W}_i(b,\lambda)) = \mathbf{\Sigma}_{w_i}(b,\lambda) = \mathbf{\Sigma}_x + (1+\lambda)\mathbf{\Sigma}_{u_i}/k_i$.

   This pseudo data have the same important property as the SIMEX algorithm defined in the previous section. That is, as $\lambda \to -1$, $\text{E}[(\mathbf{W}_i(b,\lambda) - \mathbf{X}_i)^2|\mathbf{X}_i]$ converges to zero. We estimate $\widehat{\boldsymbol{\theta}}(b,\lambda)$ by solving equation (2.3) with $\mathbf{X}_i$ replaced by $\mathbf{W}_i(b,\lambda)$ for each b and λ. Then, we averaged over b to obtain $\widehat{\boldsymbol{\theta}}(\lambda) = \sum_{b=1}^{B} \widehat{\boldsymbol{\theta}}(b,\lambda)/B$.

3. We extrapolate $\widehat{\boldsymbol{\theta}}(\lambda)$ back to $\lambda = -1$ to obtain $\widehat{\boldsymbol{\theta}}_{simex}$.

For practical use, choosing $B = 50$, 100 or 200, and taking $\mathbf{\Lambda}$ to be equally cut points of interval $[0, 1]$ or $[0, 2]$ with $M = 5$, 10 or 20 can often lead to fairly reasonable SIMEX estimates (Carroll et al., 2006).

2.2.3.3 Asymptotic Normality of SIMEX Estimate

The adaptive LASSO regularized IPW method is utilized to select the variable and estimate the covariate coefficients under the AFT setting (2.1). Here we assume that all the covariates are subject to measurement error and the SIMEX method is applied to adjust the effect of measurement error on variable selection on AFT models. For each λ and b. We obtain estimates, $\widehat{\boldsymbol{\beta}}_x(b, \lambda)$, by minimizing

$$\ell(\boldsymbol{\beta}_x(b, \lambda)) = \frac{1}{2} \sum_{i=1}^{n} \left(Y_{\pi(i)} - \mathbf{W}_i(b, \lambda)'_{\pi(i)} \boldsymbol{\beta}_x(b, \lambda) \right)^2 + \gamma \sum_{j=1}^{p} v_j \left| \beta_{x_j}(b, \lambda) \right|.$$

For any given λ, by applying Theorem 2 of Zou (2006), we have

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}}_x(b, \lambda) - \boldsymbol{\beta}_x(b, \lambda) \right) \rightarrow_d \text{MVN}(\mathbf{0}, \mathbf{C}(b, \lambda)),$$

where $\boldsymbol{\beta}_x(b, \lambda)$ is the true unknown adaptive LASSO regularized IPW parameter. $\mathbf{C}(b, \lambda)$ is the variance parameter. We calculate the average of these B estimators, $\widehat{\boldsymbol{\beta}}_x(\lambda) = B^{-1} \sum_{b=1}^{B} \widehat{\boldsymbol{\beta}}_x(b, \lambda)$, and let $\mathbf{C}(\lambda) = B^{-1} \sum_{b=1}^{B} \mathbf{C}(b, \lambda))$. According to Slutsky's theorem, we have

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}}_x(\lambda) - \boldsymbol{\beta}_x(\lambda) \right) \rightarrow_d \text{MVN}(\mathbf{0}, \mathbf{C}(\lambda)).$$

Let $\widehat{\boldsymbol{\beta}}_x(\mathbf{\Lambda}) = \text{vec}\{\widehat{\boldsymbol{\beta}}_x(\lambda); \lambda \in \mathbf{\Lambda}\}$, $\boldsymbol{\beta}_x(\mathbf{\Lambda}) = \text{vec}\{\boldsymbol{\beta}_x(\lambda); \lambda \in \mathbf{\Lambda}\}$ and $\mathbf{\Gamma}(\mathbf{\Lambda}) = \text{diag}\{\mathbf{C}(\lambda), \lambda \in \mathbf{\Lambda}\}$, we have

$$\sqrt{n} \left(\widehat{\boldsymbol{\beta}}_x(\mathbf{\Lambda}) - \boldsymbol{\beta}_x(\mathbf{\Lambda}) \right) \rightarrow_d \text{MVN}(\mathbf{0}, \mathbf{\Gamma}(\mathbf{\Lambda})).$$

Assume that the exact extrapolation function $\mathbf{G}(\boldsymbol{\zeta};\lambda)$ is known in the extrapolation step to fit $\widehat{\boldsymbol{\beta}}_x(\lambda)$, where $\boldsymbol{\zeta}$ is d-dimensional vector of parameters. Let $\widehat{\boldsymbol{\zeta}}$ be the least squares estimator. Let $\mathbf{G}_\zeta(\boldsymbol{\zeta};\lambda) = (\partial/\partial\boldsymbol{\zeta})\mathbf{G}'(\boldsymbol{\zeta};\lambda)$,

$$\mathbf{s}(\boldsymbol{\zeta}) = (\mathbf{G}_\zeta(\boldsymbol{\zeta};\lambda_1)\mathbf{G}_\zeta(\boldsymbol{\zeta};\lambda_2)\cdots\mathbf{G}_\zeta(\boldsymbol{\zeta};\lambda_M))$$

and $\mathbf{A}(\boldsymbol{\zeta}) = \mathbf{s}(\boldsymbol{\zeta})\mathbf{s}'(\boldsymbol{\zeta})$ be the $d \times d$ matrix. Then by the similar argument to that of Carroll et al. (1996), we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \to_d \text{MVN}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\zeta})),$$

where $\mathbf{Q}(\boldsymbol{\zeta}) = [\mathbf{A}(\boldsymbol{\zeta})]^{-1}\mathbf{s}(\boldsymbol{\zeta})\boldsymbol{\Gamma}(\boldsymbol{\Lambda})\mathbf{s}'(\boldsymbol{\zeta})[\mathbf{A}(\boldsymbol{\zeta})]^{-1}$ is a $d \times d$ matrix. Letting $\lambda = -1$ leads to the SIMEX estimator $\widehat{\boldsymbol{\beta}}_x = \mathbf{G}(\widehat{\boldsymbol{\zeta}}; -1)$. Therefore, obtain

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_x - \boldsymbol{\beta}_x) \to_d \text{MVN}(\mathbf{0}, \mathbf{G}'_\zeta(\boldsymbol{\zeta};-1)\mathbf{Q}\left(\boldsymbol{\zeta}\right)\mathbf{G}_\zeta(\boldsymbol{\zeta};-1)\big).$$

2.3 Simulation Studies

This section investigates the impact of ignoring measurement error on survival relevant gene selection in the AFT model and exploring how the SIMEX method can adjust the selection when measurement error is shown. Each simulation study consists of 100 data sets of size $n = 200$. The survival times are generated from the model (1.1) using $\beta_0 = 0$, $\boldsymbol{\beta}_x = (0.7, 0.7, 0, 0, 0, 0.7, 0, 0, 0)'$ and ϵ_i follows the standard extreme value distribution with its scale parameter, α, set to 0.5 and 1.5. The censoring times are generated such that the censoring rates are approximately 30% and 50%. The true covariates $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{i9})'$ are generated from MVN$(\mathbf{1}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is defined by

Scenario 1: Independent Covariance Matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}_{9\times 9}$$

Scenario 2: Exchangeable Covariance Matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho & & \\ \rho & 1 & \ddots & \\ & \ddots & \ddots & \rho \\ & & \rho & 1 \end{pmatrix}_{9\times 9}$$

where ρ, which is set to be 0.5 to account for moderate correlation, is the pairwise correlation between X_{ii} and $X_{i(i+1)}$. The pseudo errors $\mathbf{U}_{bi}$ are generated from MVN$(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix and σ is set to be the following four contamination levels: 0.25, 0.50, 0.75 and 1.00 to feature various degrees of measurement error.

Using the naive and SIMEX methods, Tables 2.1 and 2.2 report the results of the parameter estimates, $\widehat{\boldsymbol{\beta}}$, its standard errors, SE$(\widehat{\boldsymbol{\beta}})$, which is computed using 1000 bootstraps and the corresponding p-values under different degrees of measurement error and $\alpha = 1.5$. Tables 2.5 and 2.6 reports the results of the same settings but $\alpha = 0.5$. Since the gene expressions from the microarray may be correlated with each other, we assume that the genes have been rearranged based on their functional similarity. We generated gene expressions using the exchangeable covariance matrix, where gene expressions are only correlated with the ones right beside them. Tables 2.3 and 2.4 contain the results of using the exchangeable covariance matrix.

When the measurement error is minor (i.e., $\sigma = 0.25$) or moderate (i.e., $\sigma = 0.50$), the impact of measurement error is not noticeable. The naive method, which uses the adaptive LASSO regularized IPW method by solving equation (2.2) with

$\mathbf{X}_{\pi(i)}$ replaced by $\mathbf{W}_{\pi(i)}$, can select the true model. However, when the measurement error becomes increasingly larger, the effect of measurement error is more obvious. The bias of the $\widehat{\boldsymbol{\beta}}$ increases, while the SE($\widehat{\boldsymbol{\beta}}$) decreases. The covariates, X_{i1}, X_{i2} and X_{i6}, which should be in the true model, can be selected. However, as the measurement error becomes severe, those covariates that are not actually in the true model were improperly selected since their p-values were typically smaller than the 5% level.

Afterwards, the SIMEX approach was employed with $B = 50$, $\lambda_M = 2$ and $M = 11$ to adjust the impact of measurement error to the selection of variables. In this chapter, we choose to use quadratic extrapolation for the SIMEX method as used in He et al. (2007). The results are listed in Tables 2.1 to 2.4. These results show that the SIMEX approach improves the performance of variable selection when measurement errors are present. The SIMEX method gives good estimates of $\boldsymbol{\beta}_x$, especially those that are not present in the true model, since the biases are smaller and the p-values provide the correct conclusion at the 5% significance level. However, when the measurement error is severe, the SIMEX method seems to perform less satisfactorily. The bias tends to increase as the degree of measurement error increases. By comparing the estimates reported in Tables 2.1 to 2.6, we find that the proportion of censoring could also affect the estimation of $\boldsymbol{\beta}_x$ since the censoring rate is highly related to the Kaplan Meier weights used in the IPW step.

Table 2.1: Survival relevant gene selection: Simulation results of independent covariance matrix with α=1.5 and censoring rate=30%.

Method	Predictor	$\sigma = 0.25$			$\sigma = 0.50$			$\sigma = 0.75$			$\sigma = 1.00$		
		$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p
NAIVE	$X_1 = 0.7$	0.645	0.110	0	0.594	0.110	0	0.543	0.107	0	0.482	0.104	0
	$X_2 = 0.7$	0.645	0.110	0	0.597	0.109	0	0.531	0.106	0	0.476	0.102	0
	$X_3 = 0$	0.159	0.095	0.095	0.173	0.094	0.065	0.217	0.095	0.022	0.220	0.090	0.015
	$X_4 = 0$	0.147	0.095	0.123	0.199	0.097	0.041	0.219	0.094	0.020	0.214	0.090	0.018
	$X_5 = 0$	0.175	0.097	0.072	0.161	0.090	0.075	0.185	0.093	0.045	0.210	0.088	0.017
	$X_6 = 0.7$	0.658	0.110	0	0.597	0.108	0	0.547	0.107	0	0.481	0.103	0
	$X_7 = 0$	0.139	0.092	0.130	0.187	0.096	0.052	0.180	0.089	0.044	0.231	0.091	0.011
	$X_8 = 0$	0.161	0.096	0.092	0.192	0.095	0.044	0.203	0.092	0.028	0.219	0.092	0.017
	$X_9 = 0$	0.154	0.097	0.111	0.182	0.094	0.053	0.193	0.093	0.038	0.219	0.095	0.021
SIMEX	$X_1 = 0.7$	0.690	0.122	0	0.687	0.130	0	0.632	0.140	0	0.579	0.138	0
	$X_2 = 0.7$	0.671	0.119	0	0.629	0.129	0	0.634	0.144	0	0.544	0.139	0
	$X_3 = 0$	0.140	0.104	0.178	0.144	0.122	0.239	0.166	0.125	0.184	0.218	0.128	0.088
	$X_4 = 0$	0.167	0.107	0.118	0.156	0.116	0.176	0.133	0.122	0.275	0.200	0.130	0.125
	$X_5 = 0$	0.164	0.106	0.121	0.186	0.120	0.120	0.193	0.126	0.125	0.208	0.132	0.116
	$X_6 = 0.7$	0.662	0.124	0	0.645	0.129	0	0.649	0.137	0	0.594	0.141	0
	$X_7 = 0$	0.163	0.109	0.137	0.169	0.119	0.156	0.174	0.131	0.184	0.174	0.126	0.167
	$X_8 = 0$	0.152	0.108	0.161	0.154	0.116	0.186	0.192	0.135	0.156	0.212	0.129	0.101
	$X_9 = 0$	0.129	0.104	0.217	0.169	0.117	0.148	0.161	0.128	0.210	0.229	0.129	0.077

a. $\hat{\beta}$: coefficient estimate; SE($\hat{\beta}$): standard error; p: p-value of the hypothesis test; σ: standard deviation of the measurement error.

Table 2.2: Survival relevant gene selection: Simulation results of independent covariance matrix with α=1.5 and censoring rate=50%.

Method	Predictor	$\sigma = 0.25$			$\sigma = 0.50$			$\sigma = 0.75$			$\sigma = 1.00$		
		$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p
NAIVE	$X_1 = 0.7$	0.642	0.140	0	0.627	0.137	0	0.533	0.131	0	0.517	0.128	0
	$X_2 = 0.7$	0.664	0.143	0	0.592	0.135	0	0.542	0.133	0	0.509	0.132	0
	$X_3 = 0$	0.227	0.128	0.075	0.240	0.121	0.048	0.249	0.119	0.036	0.251	0.116	0.031
	$X_4 = 0$	0.199	0.122	0.103	0.231	0.122	0.058	0.228	0.116	0.049	0.238	0.111	0.032
	$X_5 = 0$	0.209	0.126	0.097	0.220	0.118	0.062	0.260	0.121	0.031	0.252	0.113	0.025
	$X_6 = 0.7$	0.638	0.137	0	0.594	0.137	0	0.568	0.132	0	0.522	0.126	0
	$X_7 = 0$	0.210	0.124	0.090	0.238	0.121	0.049	0.250	0.122	0.041	0.231	0.117	0.049
	$X_8 = 0$	0.214	0.126	0.088	0.224	0.121	0.064	0.261	0.118	0.028	0.242	0.114	0.034
	$X_9 = 0$	0.190	0.127	0.135	0.205	0.119	0.084	0.245	0.121	0.043	0.253	0.113	0.025
SIMEX	$X_1 = 0.7$	0.638	0.143	0	0.658	0.169	0	0.586	0.171	0.001	0.560	0.178	0.002
	$X_2 = 0.7$	0.665	0.142	0	0.649	0.163	0	0.678	0.169	0	0.620	0.180	0.001
	$X_3 = 0$	0.209	0.133	0.118	0.192	0.153	0.209	0.256	0.162	0.114	0.238	0.170	0.161
	$X_4 = 0$	0.201	0.127	0.113	0.183	0.149	0.218	0.227	0.167	0.174	0.260	0.164	0.114
	$X_5 = 0$	0.189	0.127	0.137	0.237	0.153	0.121	0.177	0.156	0.257	0.200	0.161	0.215
	$X_6 = 0.7$	0.650	0.144	0	0.642	0.170	0	0.620	0.177	0	0.577	0.178	0.001
	$X_7 = 0$	0.224	0.140	0.108	0.224	0.155	0.148	0.212	0.170	0.212	0.250	0.169	0.139
	$X_8 = 0$	0.199	0.129	0.122	0.201	0.159	0.206	0.183	0.159	0.248	0.270	0.171	0.115
	$X_9 = 0$	0.209	0.128	0.102	0.231	0.153	0.132	0.227	0.160	0.156	0.235	0.174	0.177

a. $\hat{\beta}$: coefficient estimate; SE($\hat{\beta}$): standard error; p: p-value of the hypothesis test; σ: standard deviation of the measurement error.

Table 2.3: Survival relevant gene selection: Simulation results of exchangeable covariance matrix with α=1.5 and censoring rate=30%.

Method	Predictor	$\sigma = 0.25$			$\sigma = 0.50$			$\sigma = 0.75$			$\sigma = 1.00$		
		$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p
NAIVE	$X_1 = 0.7$	0.744	0.133	0	0.680	0.121	0	0.632	0.114	0	0.542	0.105	0
	$X_2 = 0.7$	0.593	0.161	0	0.585	0.139	0	0.498	0.124	0	0.459	0.111	0
	$X_3 = 0$	0.195	0.143	0.172	0.212	0.126	0.093	0.231	0.113	0.041	0.228	0.098	0.020
	$X_4 = 0$	0.058	0.134	0.666	0.057	0.114	0.618	0.098	0.106	0.357	0.117	0.088	0.182
	$X_5 = 0$	0.175	0.148	0.235	0.229	0.129	0.076	0.225	0.113	0.045	0.247	0.103	0.016
	$X_6 = 0.7$	0.575	0.163	0	0.500	0.138	0	0.458	0.124	0	0.408	0.110	0
	$X_7 = 0$	0.188	0.148	0.205	0.206	0.126	0.103	0.219	0.114	0.055	0.266	0.103	0.010
	$X_8 = 0$	0.037	0.133	0.781	0.088	0.114	0.439	0.114	0.100	0.255	0.114	0.089	0.199
	$X_9 = 0$	0.193	0.118	0.102	0.185	0.110	0.094	0.185	0.100	0.066	0.220	0.091	0.016
SIMEX	$X_1 = 0.7$	0.779	0.147	0	0.792	0.150	0	0.729	0.149	0	0.728	0.145	0
	X_2 =0.7	0.590	0.181	0.001	0.585	0.182	0.001	0.574	0.170	0.001	0.539	0.157	0.001
	$X_3 = 0$	0.189	0.174	0.278	0.192	0.173	0.268	0.184	0.155	0.234	0.209	0.142	0.142
	$X_4 = 0$	0.004	0.179	0.984	0.001	0.166	0.996	0.030	0.147	0.840	0.078	0.138	0.570
	$X_5 = 0$	0.223	0.189	0.238	0.218	0.176	0.214	0.257	0.157	0.102	0.254	0.148	0.085
	$X_6 = 0.7$	0.573	0.202	0.005	0.552	0.183	0.003	0.497	0.169	0.003	0.468	0.153	0.002
	$X_7 = 0$	0.225	0.185	0.223	0.206	0.174	0.236	0.205	0.157	0.191	0.262	0.138	0.057
	$X_8 = 0$	0.026	0.167	0.877	0.009	0.160	0.954	0.061	0.145	0.676	0.024	0.130	0.853
	$X_9 = 0$	0.173	0.139	0.212	0.203	0.150	0.176	0.184	0.143	0.200	0.193	0.135	0.154

a. $\hat{\beta}$: coefficient estimate; SE($\hat{\beta}$): standard error; p: p-value of the hypothesis test; σ: standard deviation of the measurement error.

Table 2.4: Survival relevant gene selection: Simulation results of exchangeable covariance matrix with α=1.5 and censoring rate=50%.

Method	Predictor	$\sigma = 0.25$			$\sigma = 0.50$			$\sigma = 0.75$			$\sigma = 1.00$		
		$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p
NAIVE	$X_1 = 0.7$	0.814	0.168	0	0.714	0.150	0	0.660	0.141	0	0.591	0.132	0
	$X_2 = 0.7$	0.517	0.202	0.011	0.578	0.171	0.001	0.503	0.153	0.001	0.458	0.136	0.001
	$X_3 = 0$	0.241	0.189	0.203	0.216	0.156	0.166	0.271	0.144	0.059	0.277	0.130	0.033
	$X_4 = 0$	0.047	0.178	0.793	0.098	0.146	0.502	0.122	0.130	0.345	0.135	0.114	0.237
	$X_5 = 0$	0.227	0.194	0.243	0.260	0.161	0.106	0.257	0.140	0.067	0.262	0.128	0.041
	$X_6 = 0.7$	0.573	0.206	0.005	0.515	0.169	0.002	0.469	0.153	0.002	0.434	0.139	0.002
	$X_7 = 0$	0.248	0.187	0.185	0.236	0.158	0.137	0.298	0.144	0.038	0.262	0.128	0.041
	$X_8 = 0$	0.036	0.168	0.828	0.091	0.144	0.528	0.084	0.125	0.502	0.149	0.114	0.193
	$X_9 = 0$	0.236	0.157	0.134	0.229	0.139	0.100	0.244	0.127	0.056	0.267	0.119	0.025
SIMEX	$X_1 = 0.7$	0.772	0.197	0	0.816	0.190	0	0.700	0.196	0	0.707	0.183	0
	$X_2 = 0.7$	0.565	0.260	0.030	0.506	0.227	0.026	0.628	0.217	0.004	0.592	0.198	0.003
	$X_3 = 0$	0.277	0.251	0.270	0.308	0.224	0.169	0.211	0.196	0.282	0.199	0.182	0.273
	$X_4 = 0$	-0.013	0.239	0.956	-0.038	0.208	0.857	0.071	0.175	0.686	0.103	0.179	0.567
	$X_5 = 0$	0.274	0.249	0.270	0.301	0.217	0.165	0.217	0.197	0.269	0.284	0.177	0.108
	$X_6 = 0.7$	0.559	0.253	0.027	0.503	0.227	0.027	0.522	0.213	0.014	0.489	0.202	0.015
	$X_7 = 0$	0.223	0.231	0.333	0.281	0.229	0.219	0.275	0.200	0.169	0.295	0.187	0.115
	$X_8 = 0$	0.029	0.218	0.893	-0.029	0.213	0.891	0.109	0.187	0.560	0.107	0.178	0.548
	$X_9 = 0$	0.248	0.177	0.162	0.311	0.187	0.096	0.265	0.178	0.136	0.221	0.173	0.200

a. $\hat{\beta}$: coefficient estimate; SE($\hat{\beta}$): standard error; p: p-value of the hypothesis test; σ: standard deviation of the measurement error.

Table 2.5: Survival relevant gene selection: Simulation results of independent covariance matrix with α=0.5 and censoring rate=30%.

Method	Predictor	$\sigma = 0.25$			$\sigma = 0.50$			$\sigma = 0.75$			$\sigma = 1.00$		
		$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p
NAIVE	$X_1 = 0.7$	0.647	0.109	0	0.585	0.107	0	0.517	0.107	0	0.480	0.102	0
	$X_2 = 0.7$	0.635	0.107	0	0.599	0.109	0	0.533	0.106	0	0.494	0.099	0
	$X_3 = 0$	0.176	0.097	0.069	0.192	0.093	0.038	0.202	0.090	0.025	0.207	0.087	0.017
	$X_4 = 0$	0.148	0.094	0.116	0.175	0.092	0.058	0.220	0.093	0.019	0.219	0.088	0.013
	$X_5 = 0$	0.149	0.092	0.103	0.185	0.094	0.049	0.189	0.087	0.030	0.232	0.091	0.011
	$X_6 = 0.7$	0.654	0.106	0	0.569	0.107	0	0.535	0.105	0	0.473	0.101	0
	$X_7 = 0$	0.158	0.093	0.091	0.187	0.094	0.046	0.203	0.089	0.022	0.197	0.085	0.021
	$X_8 = 0$	0.162	0.093	0.080	0.198	0.094	0.034	0.207	0.090	0.022	0.209	0.087	0.016
	$X_9 = 0$	0.161	0.098	0.099	0.176	0.093	0.057	0.206	0.091	0.023	0.214	0.087	0.014
SIMEX	$X_1 = 0.7$	0.664	0.120	0	0.659	0.134	0	0.627	0.143	0	0.580	0.145	0
	$X_2 = 0.7$	0.678	0.116	0	0.666	0.128	0	0.632	0.135	0	0.587	0.136	0
	$X_3 = 0$	0.155	0.106	0.145	0.160	0.120	0.184	0.178	0.128	0.165	0.201	0.129	0.121
	$X_4 = 0$	0.124	0.105	0.238	0.132	0.120	0.273	0.164	0.127	0.197	0.190	0.126	0.130
	$X_5 = 0$	0.135	0.105	0.196	0.139	0.118	0.240	0.157	0.129	0.223	0.182	0.130	0.160
	$X_6 = 0.7$	0.683	0.117	0	0.679	0.131	0	0.651	0.138	0	0.608	0.140	0
	$X_7 = 0$	0.139	0.107	0.194	0.147	0.119	0.217	0.174	0.127	0.173	0.202	0.130	0.120
	$X_8 = 0$	0.139	0.105	0.185	0.138	0.118	0.242	0.157	0.124	0.208	0.184	0.125	0.141
	$X_9 = 0$	0.154	0.107	0.152	0.161	0.123	0.190	0.175	0.130	0.178	0.200	0.130	0.124

a. $\hat{\beta}$: coefficient estimate; SE($\hat{\beta}$): standard error; p: p-value of the hypothesis test; σ: standard deviation of the measurement error.

Table 2.6: Survival relevant gene selection: Simulation results of independent covariance matrix with α=0.5 and censoring rate=50%.

Method	Predictor	$\sigma = 0.25$			$\sigma = 0.50$			$\sigma = 0.75$			$\sigma = 1.00$		
		$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p	$\hat{\beta}$	SE($\hat{\beta}$)	p
NAIVE	$X_1 = 0.7$	0.657	0.132	0	0.625	0.135	0	0.550	0.130	0	0.506	0.125	0
	$X_2 = 0.7$	0.625	0.134	0	0.595	0.131	0	0.558	0.131	0	0.528	0.128	0
	$X_3 = 0$	0.204	0.122	0.095	0.224	0.117	0.055	0.245	0.114	0.031	0.233	0.108	0.031
	$X_4 = 0$	0.206	0.119	0.083	0.212	0.114	0.062	0.224	0.110	0.041	0.254	0.113	0.025
	$X_5 = 0$	0.208	0.118	0.078	0.214	0.115	0.064	0.232	0.110	0.035	0.262	0.110	0.017
	$X_6 = 0.7$	0.645	0.132	0	0.574	0.131	0	0.573	0.130	0	0.518	0.121	0
	$X_7 = 0$	0.234	0.129	0.069	0.218	0.116	0.061	0.227	0.115	0.048	0.259	0.113	0.022
	$X_8 = 0$	0.184	0.118	0.118	0.247	0.121	0.041	0.224	0.112	0.046	0.241	0.109	0.028
	$X_9 = 0$	0.203	0.118	0.085	0.233	0.120	0.052	0.243	0.111	0.029	0.228	0.106	0.032
SIMEX	$X_1 = 0.7$	0.643	0.149	0	0.631	0.167	0	0.604	0.177	0.001	0.565	0.179	0.002
	$X_2 = 0.7$	0.682	0.142	0	0.673	0.158	0	0.649	0.168	0	0.617	0.171	0
	$X_3 = 0$	0.195	0.132	0.139	0.203	0.147	0.167	0.219	0.158	0.165	0.231	0.161	0.151
	$X_4 = 0$	0.178	0.130	0.169	0.185	0.150	0.220	0.214	0.167	0.201	0.238	0.167	0.153
	$X_5 = 0$	0.204	0.127	0.109	0.199	0.144	0.167	0.213	0.154	0.168	0.230	0.157	0.142
	$X_6 = 0.7$	0.677	0.148	0.000	0.689	0.166	0	0.680	0.176	0	0.657	0.178	0
	$X_7 = 0$	0.191	0.127	0.134	0.196	0.149	0.188	0.223	0.162	0.168	0.248	0.166	0.135
	$X_8 = 0$	0.191	0.123	0.122	0.189	0.142	0.184	0.211	0.153	0.169	0.232	0.158	0.142
	$X_9 = 0$	0.203	0.129	0.115	0.198	0.144	0.169	0.207	0.160	0.195	0.224	0.161	0.165

a. $\hat{\beta}$: coefficient estimate; SE($\hat{\beta}$): standard error; p: p-value of the hypothesis test; σ: standard deviation of the measurement error.

2.4 Real Data Analysis

2.4.1 PBC Data

The PBC data were collected from the Mayo Clinic trial conducted between 1974 and 1984. A wide range of health-related covariates were collected for 312 randomized participants. See Tibshirani (1997) for the detailed description of those covariates. We first transform the survival time and some covariates on 276 complete observations of the PBC data according to the recommendations by Huang et al (2006). These 276 patients were followed from diagnosis until death or censoring, where the censoring rate was 59.78%. We fit both the AFT model with the Weibull distribution and AFT model by regularizing IPW estimate with adaptive LASSO penalty on these data. Standard error estimates are obtained by using the bootstrap with 1000 replications.

Table 2.7 provides the results for the estimate, the standard error and the p-value for all covariates used in both methods. It seems that the AFT model using the adaptive LASSO regularized IPW method yields a smaller model than the AFT model using the Weibull distribution. Therefore, we will use only the adaptive LASSO regularized IPW method for the investigation of the impact of the measurement error.

We conduct sensitivity analysis by adding different levels of measurement error to the covariates to assess the impact of measurement error on variable selection in the PBC data. Measurement errors are randomly and independently added to the continuous covariates. They are generated from $\mathrm{N}(0, \sigma^2)$, where the standard deviation, σ, is proportional to the standard deviation, σ_x, of the corresponding covariate. The proportions are set to be 10%, 30% and 50% to represent various degrees of measurement error.

Table 2.8 shows the results of the estimate, associated standard error and corresponding p-value from the sensitivity analysis. The error free covariates do not seem to be greatly affected by the error prone covariates. In the naive method, as measurement error becomes severe, the bias of the estimate increases while the standard error

estimate decreases. The impact of measurement error on estimation of the log(bili) is very noticeable. For the original PBC data, the adaptive LASSO regularized IPW p-value of log(bili) in Table 2.7 is 0.043 (i.e., log(bili) is significantly associated with the survival time). The naive method computes p-values greater than 5% under different degrees of measurement error; this concludes that there is no evidence that log(bili) is a survival related covariate, which conflicts with the result from the original data. On the other hand, the SIMEX method makes adjustment to the effect of measurement error. The estimates from the SIMEX method have smaller bias compared to the naive method. The corresponding p-values are more consistent with the p-values given by the adaptive LASSO method of the original PBC data set.

Table 2.7: Fit the AFT model to PBC data using adaptive LASSO regularzied IPW method. $\hat{\beta}_x$: estimate of coefficient; SE($\hat{\beta}_x$): the bootstrap standard error; p: the corresponding p-value.

Predictor	AFT			Adaptive LASSO		
	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
Age	-0.020	0.007	0.005	-0.012	0.009	0.221
Alb	0.305	0.178	0.087	0.393	0.228	0.084
Log(alkphos)	-0.054	0.089	0.548	0.147	0.098	0.133
Ascites	-0.216	0.225	0.336	-0.464	0.442	0.294
Log(bili)	-0.346	0.104	0.001	-0.234	0.115	0.043
Log(chol)	-0.099	0.180	0.580	0.138	0.166	0.406
Edtrt	-0.639	0.226	0.005	-0.612	0.508	0.228
Hepmeg	0.047	0.155	0.761	0.158	0.153	0.302
Log(platelet)	-0.062	0.176	0.727	0	0.172	1
Log(protime)	-1.626	0.832	0.051	0.577	0.874	0.509
Sex	-0.091	0.195	0.640	0	0.182	1
Log(sgot)	-0.281	0.187	0.134	-0.081	0.209	0.697
Spiders	-0.002	0.147	0.987	-0.189	0.167	0.256
Stage	-0.220	0.107	0.039	-0.114	0.095	0.231
Trt	-0.010	0.128	0.661	0	0.136	1
Log(trig)	0.072	0.153	0.636	-0.063	0.185	0.735
Log(copper)	-0.170	0.107	0.112	-0.137	0.105	0.192

Table 2.8: PBC data: Sensitivity analysis with all of the quantitative covariates subject to measurement error

Method	Predictor	$\sigma = 10\%\sigma_x$			$\sigma = 30\%\sigma_x$			$\sigma = 50\%\sigma_x$		
		$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
NAIVE	Age	-0.012	0.009	0.205	-0.013	0.009	0.178	-0.013	0.009	0.152
	Alb	0.387	0.235	0.099	0.347	0.218	0.112	0.298	0.198	0.132
	Log(alkphos)	0.145	0.100	0.148	0.123	0.094	0.191	0.100	0.087	0.250
	Ascites	-0.474	0.430	0.271	-0.492	0.426	0.248	-0.520	0.429	0.226
	Log(bili)	-0.227	0.121	0.060	-0.193	0.105	0.065	-0.157	0.090	0.082
	Log(chol)	0.141	0.169	0.403	0.102	0.148	0.493	0.068	0.136	0.616
	Edtrt	-0.617	0.507	0.223	-0.649	0.501	0.195	-0.688	0.502	0.170
	Hepmeg	0.156	0.157	0.320	0.130	0.151	0.388	0.111	0.151	0.463
	Log(platelet)	-0.015	0.174	0.931	-0.031	0.163	0.850	-0.037	0.157	0.814
	Log(protime)	0.572	0.888	0.519	0.346	0.807	0.668	0.242	0.763	0.751
	Sex	0.000	0.187	1.000	0.010	0.178	0.957	0.013	0.178	0.940
	Log(sgot)	-0.106	0.223	0.634	-0.155	0.204	0.447	-0.193	0.185	0.296
	Spiders	-0.197	0.167	0.238	-0.200	0.164	0.224	-0.213	0.167	0.203
	Stage	-0.117	0.100	0.240	-0.130	0.100	0.194	-0.146	0.101	0.148
	Trt	-0.005	0.140	0.974	-0.010	0.133	0.938	-0.017	0.136	0.902
	Log(trig)	-0.068	0.185	0.713	-0.064	0.168	0.701	-0.064	0.155	0.677
	Log(copper)	-0.135	0.105	0.197	-0.115	0.098	0.239	-0.100	0.089	0.260
SIMEX	Age	-0.011	0.009	0.224	-0.011	0.009	0.201	-0.010	0.008	0.222
	Alb	0.550	0.271	0.042	0.342	0.199	0.086	0.213	0.156	0.173
	Log(alkphos)	0.207	0.113	0.068	0.136	0.086	0.115	0.101	0.070	0.150
	Ascites	-0.337	0.407	0.408	-0.396	0.408	0.332	-0.454	0.418	0.278
	Log(bili)	-0.316	0.127	0.013	-0.197	0.084	0.019	-0.134	0.063	0.032
	Log(chol)	0.160	0.173	0.356	-0.005	0.119	0.963	-0.016	0.093	0.867
	Edtrt	-0.574	0.497	0.248	-0.801	0.499	0.108	-0.897	0.504	0.075
	Hepmeg	0.225	0.167	0.179	0.152	0.156	0.330	0.101	0.147	0.493
	Log(platelet)	-0.042	0.197	0.830	-0.022	0.151	0.884	-0.022	0.124	0.861
	Log(protime)	0.457	0.981	0.641	-0.051	0.734	0.944	-0.149	0.636	0.815
	Sex	-0.001	0.191	0.996	0.007	0.170	0.965	0.012	0.157	0.940
	Log(sgot)	-0.170	0.249	0.496	-0.171	0.184	0.351	-0.119	0.148	0.420
	Spiders	-0.152	0.159	0.339	-0.135	0.158	0.392	-0.147	0.162	0.362
	Stage	-0.074	0.090	0.411	-0.122	0.089	0.169	-0.140	0.090	0.119
	Trt	-0.001	0.138	0.993	-0.011	0.132	0.934	-0.008	0.130	0.951
	Log(trig)	-0.115	0.207	0.578	-0.106	0.155	0.493	-0.086	0.129	0.506
	Log(copper)	-0.147	0.122	0.230	-0.056	0.093	0.546	-0.012	0.076	0.870

2.4.2 DLBCL Data

The DLBCL data consists of 7399 gene expression profiles across 240 patients with untreated diffuse large-B-cell lymphoma (Rosenwald et al., 2002). The outcome is the survival time, which is either observed or right censored. The median survival time is 2.8 years. During the follow up period, 138 patient deaths were observed (i.e., the censoring rate is 42.5%). These patients with zero survival time are excluded.

We are interested in finding survival relevant genes. First, we reduced the dimension of genes. Many authors have analyzed survival times based on gene expression profiles. We used the 74 genes reported by He and Yi (2009) plus 26 other randomly selected genes as potentially survival relevant genes. We fit the AFT model with the Weibull distribution by regularizing IPW estimate with adaptive LASSO penalty. The weights used in the adaptive LASSO step are computed using the LASSO estimates. The bootstrap method with 1000 replications is applied to calculate the standard error estimates with the same optimal adaptive LASSO penalty parameter. Table 2.9 summarizes the results for the estimate, the standard error and the p-value for each of the 100 genes.

Sensitivity analysis is applied by adding a sequence of values of measurement error to the gene expressions to assess the impact of measurement error on the DLBCL data set. Measurement error are randomly and independently added to the true gene expressions with standard deviation, σ_{me}, proportional to the corresponding gene's standard deviation respectively. The proportions are set to be 10%, 20%, 30% and 50%. The results from the naive method are reported in Tables 2.10 and 2.11. In these tables, the biases of the estimates are attenuated to zero. That is, as the standard deviation of the measurement error increases, the attenuations become severe. Tables 2.14 and 2.15 report the rank of the genes according to the descending level of significance. Gene 92 is ranked 37th in terms of significant correlation with the patient's survival time on the original DLBCL data set. However, the naive method ranked it the first survival relevant when σ_{me}=10% or 20% of the corresponding gene's

standard deviation respectively. which conflict with the results from the original data. As the measurement error becomes severe, the effect becomes more noticeable. For example, genes 47, 61 and 63 are reported as being the top 3 survival relevant genes from the original data. However, these 3 genes are ranked much lower under the naive method with different degrees of measurement error. On the other hand, the SIMEX method makes adjustment to the effect of measurement error. It provids estimates with smaller bias than those from the naive method. Meanwhile, the corresponding rankings of p-values are more consistent in value with those given by the adaptive LASSO IPW method of the original DLBCL data set. Hence, the model selected by the SIMEX method is more accurate than the one selected by the naive method.

Table 2.9: Fit the AFT model to DLBCL data: $\hat{\beta}_x$ is the estimate of coefficient, SE($\hat{\beta}_x$) is the bootstrap standard error and p is the corresponding p-value.

Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
1	0.240	0.450	0.594	26	0.244	0.225	0.277	51	0.000	0.334	1.000	76	-0.124	0.182	0.496
2	-0.128	0.239	0.593	27	-0.271	0.179	0.130	52	-0.106	0.246	0.666	77	0.014	0.111	0.899
3	-0.036	0.259	0.890	28	-0.017	0.105	0.869	53	0.314	0.276	0.255	78	0.108	0.132	0.414
4	-0.281	0.285	0.324	29	0.280	0.170	0.100	54	0.000	0.152	1.000	79	0.000	0.164	1.000
5	-0.372	0.417	0.372	30	-0.521	0.296	0.079	55	0.097	0.283	0.733	80	-0.162	0.160	0.311
6	-0.038	0.397	0.924	31	-0.112	0.107	0.295	56	-0.461	0.303	0.128	81	0.476	0.300	0.112
7	0.441	0.277	0.111	32	0.000	0.146	1.000	57	0.507	0.261	0.052	82	-0.150	0.246	0.540
8	0.000	0.303	1.000	33	0.013	0.159	0.937	58	0.221	0.267	0.408	83	0.285	0.175	0.104
9	-0.539	0.410	0.189	34	0.008	0.250	0.974	59	0.000	0.173	1.000	84	0.113	0.286	0.694
10	0.226	0.417	0.588	35	-0.304	0.197	0.123	60	-0.104	0.229	0.649	85	0.000	0.405	1.000
11	-0.411	0.477	0.389	36	0.272	0.211	0.197	61	0.640	0.255	0.012	86	0.036	0.150	0.808
12	-0.049	0.417	0.906	37	-0.269	0.201	0.181	62	-0.128	0.208	0.540	87	0.000	0.169	1.000
13	0.141	0.334	0.674	38	0.095	0.142	0.502	63	-0.616	0.234	0.008	88	-0.182	0.211	0.387
14	-0.153	0.452	0.735	39	0.000	0.196	1.000	64	-0.029	0.193	0.881	89	0.042	0.211	0.842
15	-0.313	0.203	0.123	40	-0.247	0.248	0.319	65	0.332	0.223	0.136	90	0.052	0.154	0.737
16	0.059	0.173	0.732	41	0.202	0.193	0.294	66	0.019	0.221	0.933	91	0.065	0.262	0.804
17	0.036	0.165	0.827	42	-0.079	0.189	0.677	67	-0.396	0.301	0.188	92	0.238	0.230	0.300
18	-0.074	0.127	0.559	43	0.064	0.238	0.787	68	0.363	0.219	0.098	93	0.000	0.183	1.000
19	0.347	0.303	0.252	44	-0.198	0.186	0.288	69	-0.192	0.184	0.297	94	-0.236	0.222	0.290
20	0.092	0.120	0.447	45	0.000	0.262	1.000	70	0.037	0.171	0.828	95	-0.450	0.381	0.237
21	-0.082	0.305	0.789	46	0.418	0.296	0.158	71	0.195	0.169	0.249	96	-0.344	0.224	0.126
22	0.000	0.183	1.000	47	-0.941	0.379	0.013	72	0.000	0.182	1.000	97	-0.270	0.258	0.296
23	-0.010	0.196	0.960	48	0.217	0.198	0.273	73	-0.106	0.285	0.710	98	-0.282	0.296	0.342
24	0.448	0.236	0.058	49	-0.064	0.232	0.783	74	0.049	0.231	0.833	99	-0.596	0.264	0.024
25	0.000	0.280	1.000	50	0.003	0.153	0.982	75	-0.157	0.147	0.284	100	0.253	0.259	0.329

Table 2.10: NAIVE Method: Sensitivity Analysis on DLBCL Data Set (1)

Predictor	$\sigma_{me} = 10\%\sigma_x$			$\sigma_{me} = 25\%\sigma_x$			$\sigma_{me} = 30\%\sigma_x$			$\sigma_{me} = 50\%\sigma_x$		
Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
1	0.071	0.296	0.810	0.022	0.208	0.917	0.012	0.185	0.949	-0.001	0.124	0.992
2	0.045	0.159	0.776	0.034	0.111	0.757	0.028	0.097	0.770	0.016	0.063	0.795
3	0.174	0.179	0.332	0.136	0.121	0.260	0.118	0.105	0.262	0.074	0.068	0.279
4	-0.209	0.207	0.312	-0.099	0.140	0.480	-0.077	0.125	0.537	-0.025	0.084	0.765
5	-0.177	0.283	0.532	-0.069	0.183	0.706	-0.038	0.160	0.810	-0.003	0.106	0.979
6	0.030	0.317	0.925	0.080	0.225	0.722	0.067	0.201	0.738	0.044	0.140	0.751
7	0.158	0.188	0.403	0.086	0.125	0.492	0.072	0.109	0.510	0.036	0.072	0.617
8	-0.090	0.226	0.690	-0.127	0.164	0.440	-0.110	0.146	0.452	-0.070	0.102	0.489
9	-0.159	0.255	0.532	-0.100	0.168	0.553	-0.084	0.147	0.570	-0.044	0.096	0.646
10	-0.132	0.305	0.666	-0.064	0.213	0.764	-0.040	0.186	0.830	-0.015	0.123	0.902
11	0.091	0.326	0.780	0.008	0.222	0.970	-0.022	0.194	0.909	-0.030	0.128	0.813
12	-0.326	0.310	0.292	-0.250	0.214	0.241	-0.214	0.187	0.254	-0.131	0.122	0.285
13	-0.013	0.234	0.957	0.028	0.159	0.859	0.029	0.138	0.835	0.015	0.091	0.866
14	-0.099	0.303	0.745	-0.005	0.210	0.982	0.004	0.185	0.982	0.014	0.124	0.911
15	0.044	0.134	0.744	0.013	0.095	0.892	-0.002	0.084	0.978	-0.013	0.057	0.818
16	0.012	0.103	0.908	0.023	0.074	0.752	0.023	0.066	0.726	0.014	0.045	0.758
17	-0.023	0.123	0.851	-0.023	0.094	0.807	-0.019	0.084	0.819	-0.009	0.059	0.873
18	0.136	0.103	0.188	0.050	0.068	0.463	0.039	0.059	0.505	0.021	0.039	0.600
19	0.299	0.229	0.191	0.236	0.167	0.157	0.192	0.146	0.187	0.106	0.098	0.279
20	-0.097	0.103	0.348	-0.112	0.070	0.108	-0.093	0.060	0.120	-0.045	0.038	0.242
21	-0.108	0.220	0.622	-0.158	0.163	0.333	-0.129	0.143	0.367	-0.061	0.095	0.520
22	0.110	0.147	0.453	0.036	0.097	0.707	0.025	0.085	0.764	0.004	0.054	0.937
23	-0.194	0.165	0.240	-0.052	0.107	0.624	-0.044	0.093	0.639	-0.026	0.060	0.662
24	0.231	0.154	0.134	0.048	0.101	0.632	0.029	0.089	0.740	0.007	0.060	0.905
25	-0.071	0.204	0.728	0.010	0.140	0.945	0.004	0.124	0.977	-0.003	0.084	0.975
26	0.119	0.171	0.484	0.002	0.118	0.984	-0.008	0.104	0.939	-0.016	0.069	0.813
27	-0.029	0.121	0.809	0.105	0.091	0.252	0.101	0.081	0.212	0.064	0.055	0.241
28	-0.040	0.081	0.625	-0.031	0.058	0.584	-0.026	0.050	0.609	-0.009	0.033	0.791
29	0.168	0.116	0.147	0.084	0.075	0.261	0.070	0.066	0.283	0.048	0.044	0.273
30	-0.334	0.215	0.120	-0.118	0.130	0.364	-0.090	0.113	0.424	-0.041	0.073	0.578
31	0.024	0.070	0.731	0.008	0.048	0.869	0.005	0.041	0.901	0.002	0.027	0.945
32	0.011	0.097	0.914	-0.030	0.068	0.660	-0.028	0.059	0.635	-0.024	0.039	0.532
33	0.020	0.108	0.854	0.013	0.070	0.858	0.009	0.060	0.886	0.002	0.038	0.962
34	0.121	0.199	0.544	0.036	0.132	0.784	0.022	0.115	0.849	-0.014	0.077	0.857
35	0.002	0.149	0.989	0.052	0.106	0.621	0.041	0.093	0.656	0.019	0.063	0.767
36	-0.151	0.175	0.388	-0.039	0.114	0.732	-0.029	0.100	0.774	-0.021	0.066	0.748
37	-0.153	0.126	0.226	-0.075	0.087	0.389	-0.063	0.076	0.408	-0.036	0.050	0.473
38	0.009	0.098	0.924	-0.017	0.065	0.790	-0.014	0.057	0.800	-0.012	0.037	0.748
39	0.014	0.142	0.919	-0.013	0.099	0.897	-0.019	0.087	0.824	-0.030	0.060	0.617
40	-0.044	0.172	0.798	0.023	0.120	0.849	0.026	0.106	0.809	0.034	0.074	0.641
41	0.169	0.146	0.247	0.061	0.101	0.544	0.044	0.088	0.618	0.017	0.059	0.775
42	0.008	0.142	0.955	-0.039	0.100	0.697	-0.040	0.089	0.652	-0.023	0.062	0.707
43	-0.063	0.147	0.668	0.008	0.104	0.936	0.012	0.092	0.893	0.010	0.062	0.875
44	-0.163	0.144	0.258	-0.129	0.099	0.193	-0.105	0.087	0.225	-0.062	0.058	0.286
45	-0.068	0.181	0.706	-0.060	0.134	0.656	-0.062	0.121	0.609	-0.050	0.085	0.551
46	0.092	0.181	0.611	-0.003	0.121	0.979	-0.004	0.104	0.968	-0.001	0.067	0.994
47	-0.535	0.241	0.027	-0.122	0.147	0.407	-0.079	0.127	0.535	-0.011	0.081	0.887
48	0.127	0.128	0.319	0.021	0.076	0.779	0.012	0.067	0.859	0.003	0.044	0.941
49	0.276	0.202	0.172	0.059	0.123	0.633	0.031	0.106	0.772	-0.012	0.069	0.861
50	-0.055	0.107	0.607	-0.060	0.075	0.424	-0.049	0.065	0.448	-0.026	0.042	0.536

Table 2.11: NAIVE Method: Sensitivity Analysis on DLBCL Data Set (2)

Predictor	$\sigma_{me} = 10\%\sigma_x$			$\sigma_{me} = 25\%\sigma_x$			$\sigma_{me} = 30\%\sigma_x$			$\sigma_{me} = 50\%\sigma_x$		
Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
51	-0.007	0.249	0.977	-0.012	0.175	0.947	-0.019	0.154	0.899	-0.022	0.104	0.832
52	0.048	0.157	0.758	-0.012	0.105	0.907	-0.016	0.092	0.863	-0.012	0.060	0.839
53	0.145	0.172	0.398	0.028	0.106	0.795	0.012	0.092	0.894	-0.003	0.059	0.955
54	-0.031	0.111	0.783	-0.055	0.076	0.473	-0.051	0.068	0.452	-0.040	0.046	0.381
55	0.284	0.220	0.198	0.079	0.136	0.562	0.065	0.118	0.583	0.032	0.077	0.676
56	0.086	0.142	0.546	0.005	0.084	0.953	0.002	0.072	0.979	0.003	0.044	0.947
57	0.173	0.151	0.252	0.070	0.090	0.436	0.057	0.077	0.460	0.035	0.048	0.463
58	-0.057	0.171	0.741	-0.055	0.122	0.652	-0.046	0.107	0.666	-0.026	0.069	0.707
59	0.076	0.109	0.488	0.045	0.068	0.505	0.040	0.059	0.493	0.026	0.038	0.488
60	0.082	0.151	0.586	0.071	0.102	0.485	0.056	0.089	0.528	0.023	0.058	0.698
61	0.037	0.126	0.771	0.007	0.085	0.938	0.004	0.073	0.956	-0.004	0.046	0.936
62	0.039	0.128	0.763	-0.006	0.087	0.945	-0.004	0.076	0.961	0.007	0.047	0.884
63	-0.211	0.141	0.134	-0.040	0.079	0.611	-0.022	0.067	0.738	0.003	0.041	0.941
64	-0.119	0.154	0.442	-0.023	0.094	0.809	-0.024	0.081	0.764	-0.020	0.051	0.703
65	0.092	0.148	0.534	0.063	0.107	0.556	0.053	0.094	0.570	0.037	0.062	0.548
66	0.069	0.146	0.637	-0.004	0.101	0.965	-0.012	0.089	0.895	-0.020	0.057	0.732
67	0.001	0.184	0.994	-0.027	0.128	0.832	-0.035	0.113	0.758	-0.034	0.073	0.639
68	0.080	0.117	0.493	0.039	0.075	0.602	0.035	0.066	0.593	0.024	0.041	0.558
69	-0.062	0.103	0.545	0.001	0.066	0.988	0.008	0.058	0.884	0.015	0.038	0.693
70	-0.086	0.118	0.465	-0.060	0.080	0.452	-0.047	0.070	0.501	-0.028	0.045	0.539
71	0.052	0.108	0.627	0.019	0.068	0.783	0.013	0.060	0.830	0.001	0.038	0.989
72	0.078	0.137	0.569	0.084	0.096	0.382	0.071	0.084	0.398	0.044	0.052	0.403
73	0.145	0.192	0.451	0.091	0.128	0.475	0.076	0.112	0.496	0.044	0.073	0.549
74	-0.161	0.173	0.353	-0.076	0.116	0.510	-0.071	0.102	0.489	-0.056	0.068	0.410
75	0.050	0.105	0.634	0.088	0.075	0.240	0.085	0.066	0.201	0.065	0.044	0.146
76	-0.188	0.133	0.156	-0.068	0.086	0.428	-0.056	0.076	0.458	-0.038	0.051	0.458
77	0.002	0.079	0.984	0.011	0.057	0.844	0.010	0.050	0.849	0.004	0.034	0.898
78	0.005	0.090	0.953	-0.005	0.064	0.932	-0.002	0.057	0.976	0.002	0.039	0.952
79	0.022	0.124	0.862	0.024	0.090	0.788	0.020	0.080	0.799	0.006	0.055	0.907
80	-0.104	0.126	0.413	-0.056	0.095	0.556	-0.057	0.086	0.506	-0.047	0.060	0.427
81	0.369	0.211	0.080	0.157	0.135	0.244	0.119	0.117	0.309	0.045	0.075	0.551
82	0.067	0.185	0.718	-0.013	0.129	0.917	-0.015	0.114	0.893	-0.016	0.076	0.831
83	0.140	0.128	0.275	0.079	0.091	0.385	0.067	0.080	0.404	0.045	0.055	0.411
84	0.080	0.184	0.662	0.017	0.122	0.890	0.012	0.105	0.908	-0.000	0.066	0.997
85	0.217	0.331	0.513	0.103	0.240	0.670	0.080	0.212	0.705	0.008	0.144	0.954
86	-0.069	0.107	0.517	-0.071	0.076	0.351	-0.056	0.066	0.395	-0.022	0.042	0.590
87	-0.078	0.125	0.533	-0.017	0.088	0.850	-0.012	0.079	0.883	-0.003	0.054	0.960
88	-0.025	0.145	0.861	0.020	0.103	0.849	0.018	0.092	0.847	0.033	0.064	0.606
89	-0.252	0.178	0.157	-0.142	0.119	0.230	-0.108	0.103	0.296	-0.052	0.068	0.446
90	-0.029	0.133	0.828	0.005	0.094	0.955	0.012	0.084	0.887	0.025	0.058	0.669
91	-0.484	0.243	0.047	-0.178	0.148	0.228	-0.142	0.128	0.268	-0.090	0.084	0.283
92	0.556	0.211	0.008	0.206	0.123	0.093	0.161	0.104	0.124	0.085	0.066	0.195
93	0.127	0.143	0.374	0.016	0.100	0.872	0.006	0.087	0.942	0.004	0.059	0.943
94	0.080	0.184	0.666	0.060	0.139	0.664	0.056	0.125	0.652	0.045	0.089	0.615
95	-0.058	0.255	0.820	0.000	0.174	0.999	-0.019	0.155	0.902	-0.027	0.104	0.796
96	-0.101	0.151	0.501	-0.104	0.107	0.329	-0.088	0.094	0.350	-0.056	0.063	0.374
97	-0.079	0.155	0.610	-0.029	0.111	0.793	-0.030	0.098	0.762	-0.015	0.065	0.818
98	-0.139	0.185	0.453	0.006	0.113	0.956	0.014	0.099	0.891	0.019	0.064	0.767
99	-0.292	0.171	0.087	-0.061	0.103	0.557	-0.035	0.089	0.691	-0.010	0.058	0.860
100	0.045	0.169	0.788	0.072	0.118	0.545	0.066	0.105	0.528	0.049	0.069	0.482

Table 2.12: SIMEX Method: Sensitivity Analysis on DLBCL Data Set (1)

Predictor	$\sigma_{me} = 10\%\sigma_x$			$\sigma_{me} = 25\%\sigma_x$			$\sigma_{me} = 30\%\sigma_x$			$\sigma_{me} = 50\%\sigma_x$		
Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
1	1.092	0.688	0.113	1.108	0.682	0.104	0.375	0.610	0.539	-0.185	0.448	0.679
2	-0.543	0.416	0.191	-0.850	0.476	0.074	-0.535	0.420	0.202	0.019	0.228	0.934
3	-0.269	0.418	0.520	-0.338	0.416	0.416	-0.128	0.345	0.710	0.038	0.207	0.853
4	-0.334	0.403	0.407	-0.432	0.425	0.309	-0.170	0.379	0.653	0.223	0.264	0.397
5	-0.140	0.612	0.819	0.194	0.653	0.766	0.299	0.574	0.603	0.341	0.408	0.402
6	0.018	0.700	0.979	0.021	0.747	0.977	-0.085	0.656	0.897	-0.020	0.417	0.962
7	0.444	0.376	0.238	0.338	0.396	0.393	0.282	0.367	0.442	0.360	0.278	0.195
8	0.191	0.454	0.673	0.187	0.480	0.697	0.068	0.441	0.877	-0.235	0.332	0.479
9	-0.801	0.607	0.187	-0.979	0.644	0.129	-0.720	0.540	0.182	-0.185	0.334	0.579
10	0.024	0.680	0.972	0.212	0.722	0.769	0.664	0.640	0.300	0.748	0.442	0.091
11	-0.582	0.719	0.418	-0.659	0.756	0.383	-0.483	0.689	0.483	-0.129	0.485	0.790
12	-0.428	0.660	0.517	-0.256	0.660	0.698	0.009	0.565	0.987	0.026	0.364	0.944
13	0.185	0.493	0.707	0.274	0.515	0.595	0.278	0.439	0.526	0.009	0.292	0.976
14	-0.277	0.697	0.691	-0.110	0.719	0.879	-0.044	0.631	0.944	-0.162	0.445	0.716
15	-0.382	0.311	0.219	-0.271	0.317	0.393	-0.230	0.280	0.412	-0.321	0.205	0.118
16	-0.088	0.255	0.732	-0.272	0.266	0.306	-0.275	0.245	0.263	-0.311	0.184	0.090
17	0.211	0.255	0.409	0.195	0.268	0.467	0.037	0.252	0.883	0.033	0.183	0.856
18	0.039	0.191	0.838	0.084	0.195	0.665	0.041	0.175	0.817	0.052	0.124	0.673
19	0.263	0.458	0.566	0.596	0.500	0.233	0.503	0.468	0.283	0.279	0.335	0.405
20	0.133	0.187	0.476	0.150	0.190	0.428	0.098	0.163	0.546	0.099	0.112	0.374
21	-0.086	0.496	0.862	-0.020	0.508	0.968	0.059	0.444	0.894	0.089	0.287	0.757
22	-0.671	0.407	0.099	-0.970	0.454	0.033	-0.639	0.369	0.083	-0.178	0.208	0.392
23	0.645	0.411	0.116	0.929	0.474	0.050	0.774	0.404	0.055	0.388	0.258	0.133
24	0.440	0.338	0.192	0.324	0.360	0.369	0.434	0.331	0.189	0.568	0.260	0.029
25	0.003	0.404	0.995	-0.104	0.411	0.800	-0.064	0.362	0.861	-0.052	0.229	0.821
26	0.688	0.400	0.085	0.446	0.381	0.242	-0.016	0.341	0.962	-0.171	0.234	0.464
27	-0.501	0.272	0.065	-0.478	0.282	0.091	-0.419	0.257	0.102	-0.349	0.186	0.060
28	-0.035	0.162	0.832	-0.038	0.176	0.829	-0.026	0.153	0.866	-0.028	0.096	0.773
29	0.414	0.242	0.087	0.290	0.234	0.214	0.082	0.198	0.679	0.057	0.136	0.674
30	-0.500	0.442	0.257	-0.462	0.477	0.333	-0.656	0.450	0.145	-0.853	0.358	0.017
31	-0.275	0.171	0.108	-0.310	0.179	0.083	-0.170	0.157	0.279	-0.025	0.100	0.806
32	0.328	0.269	0.222	0.392	0.269	0.145	0.094	0.229	0.683	-0.141	0.147	0.338
33	-0.332	0.264	0.208	-0.365	0.250	0.144	-0.137	0.211	0.517	-0.047	0.138	0.733
34	-0.110	0.379	0.773	-0.310	0.382	0.417	-0.402	0.331	0.223	-0.440	0.256	0.085
35	-0.428	0.334	0.201	-0.424	0.341	0.214	-0.386	0.296	0.192	-0.273	0.201	0.175
36	0.546	0.357	0.126	0.586	0.365	0.108	0.426	0.320	0.183	0.231	0.214	0.280
37	-0.223	0.361	0.537	-0.202	0.360	0.574	-0.254	0.290	0.383	-0.175	0.171	0.305
38	0.094	0.267	0.726	0.061	0.264	0.818	0.048	0.214	0.824	-0.079	0.130	0.544
39	-0.048	0.274	0.860	-0.053	0.274	0.846	0.068	0.242	0.777	0.138	0.175	0.428
40	-0.626	0.370	0.091	-0.885	0.400	0.027	-0.664	0.360	0.065	-0.539	0.275	0.050
41	0.172	0.296	0.562	0.173	0.317	0.586	0.357	0.306	0.243	0.577	0.247	0.019
42	-0.328	0.328	0.318	-0.230	0.318	0.469	-0.073	0.273	0.790	0.040	0.195	0.837
43	0.357	0.374	0.340	0.206	0.365	0.573	0.035	0.318	0.913	0.039	0.221	0.861
44	-0.389	0.277	0.160	-0.566	0.285	0.047	-0.522	0.252	0.038	-0.506	0.188	0.007
45	0.216	0.405	0.593	0.236	0.408	0.562	-0.002	0.359	0.995	-0.046	0.254	0.857
46	1.332	0.558	0.017	1.413	0.513	0.006	0.668	0.385	0.083	0.118	0.240	0.623
47	-2.510	0.784	0.001	-2.691	0.715	0.000	-1.689	0.566	0.003	-1.049	0.388	0.007
48	0.251	0.272	0.355	0.268	0.291	0.358	0.279	0.266	0.295	0.097	0.186	0.604
49	-0.518	0.462	0.262	-0.453	0.483	0.349	0.120	0.452	0.791	0.387	0.339	0.254
50	-0.132	0.222	0.553	-0.199	0.234	0.395	-0.213	0.204	0.297	-0.176	0.134	0.188

Table 2.13: SIMEX Method: Sensitivity Analysis on DLBCL Data Set (2)

Predictor	$\sigma_{me} = 10\%\sigma_x$			$\sigma_{me} = 25\%\sigma_x$			$\sigma_{me} = 30\%\sigma_x$			$\sigma_{me} = 50\%\sigma_x$		
Gene	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p	$\hat{\beta}_x$	SE($\hat{\beta}_x$)	p
51	-0.205	0.521	0.694	-0.047	0.524	0.929	0.042	0.444	0.925	0.024	0.293	0.935
52	0.002	0.380	0.996	0.105	0.393	0.789	0.073	0.344	0.831	0.043	0.222	0.846
53	0.198	0.392	0.613	0.206	0.401	0.608	0.287	0.366	0.434	0.421	0.270	0.119
54	-0.055	0.226	0.809	-0.203	0.237	0.392	-0.246	0.211	0.242	-0.199	0.152	0.192
55	0.129	0.432	0.765	-0.104	0.451	0.818	-0.111	0.404	0.783	0.172	0.285	0.546
56	-0.682	0.438	0.119	-0.387	0.399	0.332	0.059	0.318	0.852	0.178	0.165	0.281
57	0.531	0.419	0.205	0.315	0.397	0.428	0.366	0.320	0.253	0.427	0.199	0.032
58	0.287	0.403	0.476	0.397	0.411	0.334	0.100	0.353	0.777	-0.164	0.211	0.437
59	-0.118	0.257	0.644	-0.141	0.268	0.600	0.005	0.234	0.981	0.197	0.152	0.196
60	-0.471	0.386	0.222	-0.538	0.383	0.160	-0.443	0.341	0.193	-0.434	0.237	0.067
61	1.467	0.545	0.007	1.387	0.468	0.003	0.422	0.342	0.218	-0.030	0.186	0.874
62	-0.177	0.428	0.679	-0.264	0.428	0.538	-0.237	0.345	0.492	0.006	0.185	0.973
63	-1.446	0.523	0.006	-1.316	0.424	0.002	-0.376	0.327	0.249	0.018	0.183	0.920
64	0.255	0.289	0.377	0.226	0.277	0.414	0.033	0.240	0.890	-0.080	0.149	0.590
65	0.689	0.378	0.068	0.868	0.412	0.035	0.780	0.380	0.040	0.530	0.264	0.045
66	-0.017	0.321	0.957	-0.046	0.325	0.888	-0.076	0.275	0.782	-0.080	0.172	0.643
67	-0.679	0.472	0.150	-0.609	0.486	0.210	-0.199	0.411	0.628	0.162	0.276	0.557
68	0.429	0.356	0.228	0.288	0.319	0.367	0.145	0.262	0.581	0.102	0.154	0.510
69	-0.211	0.281	0.453	-0.132	0.277	0.634	-0.043	0.242	0.858	0.041	0.169	0.806
70	0.057	0.297	0.847	0.080	0.314	0.799	0.040	0.274	0.885	-0.105	0.178	0.554
71	0.257	0.230	0.264	0.280	0.239	0.243	0.283	0.219	0.195	0.210	0.154	0.172
72	-0.043	0.349	0.903	-0.249	0.366	0.496	-0.335	0.310	0.280	-0.244	0.186	0.190
73	0.009	0.443	0.983	0.037	0.451	0.935	-0.033	0.376	0.929	-0.062	0.234	0.792
74	0.311	0.372	0.403	0.348	0.368	0.344	0.224	0.327	0.494	0.050	0.229	0.828
75	-0.340	0.264	0.198	-0.422	0.282	0.134	-0.405	0.248	0.102	-0.307	0.161	0.057
76	-0.132	0.253	0.602	-0.218	0.271	0.420	-0.203	0.254	0.423	-0.070	0.175	0.688
77	0.172	0.205	0.401	0.230	0.211	0.276	0.190	0.184	0.300	0.133	0.124	0.281
78	0.111	0.197	0.572	0.096	0.210	0.646	0.026	0.196	0.893	0.005	0.138	0.973
79	0.065	0.254	0.799	0.078	0.272	0.775	0.114	0.252	0.652	0.198	0.186	0.286
80	-0.164	0.246	0.504	-0.124	0.268	0.643	-0.128	0.250	0.608	-0.033	0.189	0.862
81	0.405	0.405	0.317	0.445	0.419	0.288	0.187	0.361	0.605	-0.118	0.233	0.612
82	-0.632	0.445	0.155	-0.734	0.481	0.127	-0.660	0.426	0.122	-0.460	0.298	0.122
83	0.659	0.304	0.030	0.676	0.318	0.033	0.615	0.293	0.036	0.639	0.216	0.003
84	-0.042	0.443	0.924	0.047	0.461	0.919	0.228	0.413	0.581	0.146	0.267	0.586
85	0.596	0.644	0.354	0.549	0.643	0.393	0.139	0.558	0.804	-0.023	0.387	0.952
86	0.224	0.237	0.345	0.209	0.246	0.397	0.132	0.218	0.544	0.095	0.142	0.501
87	-0.047	0.284	0.868	-0.065	0.298	0.828	-0.105	0.263	0.689	-0.116	0.190	0.542
88	-0.545	0.345	0.114	-0.691	0.351	0.049	-0.413	0.311	0.184	-0.265	0.224	0.236
89	0.114	0.320	0.721	0.047	0.342	0.891	-0.084	0.309	0.785	-0.106	0.215	0.621
90	0.161	0.254	0.527	0.068	0.265	0.797	0.057	0.236	0.809	0.204	0.161	0.206
91	0.451	0.474	0.341	0.550	0.461	0.233	0.293	0.400	0.463	0.016	0.249	0.949
92	0.343	0.373	0.358	0.515	0.388	0.184	0.227	0.327	0.488	-0.072	0.190	0.705
93	-0.015	0.304	0.962	-0.042	0.310	0.891	0.084	0.262	0.749	0.150	0.172	0.381
94	-0.391	0.347	0.259	-0.353	0.352	0.316	-0.154	0.307	0.616	0.032	0.212	0.881
95	-0.387	0.528	0.463	-0.377	0.548	0.492	-0.380	0.481	0.430	-0.455	0.340	0.181
96	-0.210	0.329	0.524	-0.348	0.339	0.305	-0.588	0.318	0.064	-0.662	0.244	0.007
97	-0.187	0.399	0.639	-0.321	0.427	0.451	-0.203	0.368	0.581	0.086	0.233	0.713
98	-0.815	0.462	0.078	-0.845	0.425	0.047	-0.338	0.358	0.344	0.072	0.256	0.777
99	-0.497	0.369	0.178	-0.381	0.365	0.296	-0.357	0.318	0.260	-0.291	0.216	0.178
100	0.417	0.404	0.302	0.456	0.405	0.260	0.253	0.343	0.460	0.140	0.228	0.537

Table 2.14: Ranks of the genes based on level of significance from the sensitivity analysis on the DLBCL Data (1)

Rank	DLBCL	NAIVE Method				SIMEX Method			
		$10\%\sigma_x$	$25\%\sigma_x$	$30\%\sigma_x$	$50\%\sigma_x$	$10\%\sigma_x$	$25\%\sigma_x$	$30\%\sigma_x$	$50\%\sigma_x$
1	63	92	92	20	75	47	47	47	83
2	61	47	20	92	92	63	63	83	96
3	47	91	19	19	27	61	61	44	47
4	99	81	44	75	20	46	46	65	44
5	57	99	91	27	29	83	40	23	30
6	24	30	89	44	3	27	22	96	41
7	30	24	75	12	19	65	83	40	24
8	68	63	12	3	91	98	65	46	57
9	29	29	81	91	12	26	98	22	65
10	83	76	27	29	44	29	44	75	40
11	7	89	3	89	96	40	88	27	75
12	81	49	29	81	54	22	23	82	27
13	35	18	96	96	72	31	2	30	60
14	15	19	21	21	74	1	31	9	34
15	96	55	86	86	83	88	27	36	16
16	56	37	30	72	80	23	1	88	10
17	27	23	72	83	89	56	36	24	15
18	65	41	83	37	76	36	82	35	53
19	46	57	37	30	57	67	9	60	82
20	37	44	47	50	37	82	75	71	23
21	67	83	50	8	100	44	33	2	71
22	9	12	76	54	59	99	32	61	35
23	36	4	57	76	8	9	60	34	99
24	95	48	8	57	21	2	92	54	95
25	71	3	70	74	32	24	67	41	50
26	19	20	18	59	50	75	35	63	72
27	53	74	54	73	70	35	29	57	54
28	48	93	73	70	65	57	19	99	7
29	26	36	4	18	73	33	91	16	59
30	75	53	60	80	81	15	26	31	90
31	44	7	7	7	45	32	71	72	88
32	94	80	59	100	68	60	100	19	49
33	41	64	74	60	30	68	77	48	36
34	31	73	41	47	86	7	81	50	56
35	97	98	100	4	18	30	99	77	77
36	69	22	9	9	88	94	96	10	79
37	92	70	65	65	94	49	16	98	37
38	80	26	80	55	39	71	4	37	32
39	40	59	99	68	7	100	94	15	20
40	4	68	55	28	67	81	56	76	93
41	100	96	28	45	40	42	30	95	22
42	98	85	68	41	9	43	58	53	4
43	5	86	63	32	23	91	74	7	5
44	88	5	35	23	90	86	49	100	19
45	11	9	23	94	55	85	48	91	39
46	58	87	24	42	69	48	68	11	58
47	78	65	49	35	60	92	24	92	26
48	20	34	58	58	64	64	11	62	8
49	76	69	45	99	72	77	54	74	86
50	38	56	32	85	58	74	7	33	68

Table 2.15: Ranks of the genes based on level of significance from the sensitivity analysis on the DLBCL Data (2)

Rank	DLBCL	NAIVE Method				SIMEX Method			
		$10\%\sigma_x$	$25\%\sigma_x$	$30\%\sigma_x$	$50\%\sigma_x$	$10\%\sigma_x$	$25\%\sigma_x$	$30\%\sigma_x$	$50\%\sigma_x$
51	62	72	94	16	66	4	85	13	100
52	82	60	85	6	38	17	15	1	87
53	18	50	42	63	36	11	50	86	38
54	10	97	5	24	6	69	86	20	55
55	2	46	22	67	16	95	64	97	70
56	1	21	6	97	4	58	3	68	67
57	60	28	36	84	98	20	34	84	9
58	52	71	16	52	35	80	76	5	84
59	13	75	2	2	41	12	57	81	64
60	42	66	10	49	28	3	20	80	48
61	84	84	48	36	2	96	97	94	81
62	73	10	71	79	95	90	17	67	89
63	16	94	34	38	26	37	42	79	46
64	55	43	79	40	11	50	95	4	66
65	14	8	38	5	15	41	72	29	18
66	90	45	97	17	97	19	62	32	29
67	49	82	53	39	82	78	45	87	1
68	43	25	17	71	51	45	43	3	76
69	21	31	64	10	52	76	37	93	92
70	91	58	67	13	34	53	41	58	97
71	86	15	77	88	99	97	13	39	14
72	17	14	88	34	49	59	59	66	33
73	70	52	40	77	13	8	53	55	21
74	74	62	87	48	17	62	69	89	28
75	89	61	33	52	43	14	80	42	98
76	28	2	13	87	62	51	78	49	11
77	64	11	31	69	47	13	18	85	73
78	3	54	93	33	77	89	8	90	69
79	77	100	84	90	10	38	12	18	31
80	12	40	15	98	24	16	5	38	25
81	6	27	39	82	79	55	10	52	74
82	66	1	52	43	14	34	79	56	32
83	33	95	82	53	61	79	52	69	52
84	23	90	1	66	22	54	90	25	3
85	34	17	78	51	48	5	70	28	17
86	50	33	43	31	63	28	25	8	45
87	8	88	61	95	93	18	55	17	43
88	22	79	62	84	31	70	38	70	80
89	25	16	25	11	56	39	87	64	61
90	32	32	51	26	78	21	28	78	94
91	39	39	56	93	85	87	39	21	63
92	45	38	90	1	53	72	14	6	2
93	51	6	98	61	87	84	66	43	51
94	54	78	66	62	33	66	89	51	12
95	59	42	11	46	25	93	93	73	91
96	72	13	46	78	5	10	84	14	85
97	79	51	14	25	71	6	51	26	6
98	85	77	26	15	1	73	73	59	78
99	87	35	69	56	46	25	21	12	62
100	93	67	95	14	84	52	6	45	13

2.5 Summary

The impact of measurement error in covariates has been extensively studied in the literature for survival data. To our knowledge, no investigation has been done on the impact of measurement error in survival relevant gene selection in microarray data analysis. We investigate the effect of measurement error on gene selection when measurement error is accounted and use the SIMEX method to adjust this effect. Our simulation studies and real data analysis demonstrate that the SIMEX approach does outperform the naive method. With a certain amount of adjustment for the bias induced by error prone covariates, the SIMEX method can always select more accurate model than the naive method. While in the naive method, as the measurement error becomes substantial, the biases of the estimates increase and the standard error estimates decrease. This causes the corresponding p-values to be smaller than the nominal level, leading to incorrect hypothesis test results. The SIMEX method, by contrast, reduces the estimate biases.

Chapter 3

Prediction of Survival Time by Combining Mismeasured Gene Expression Data from Different Platforms

3.1 Introduction

Knowledge of the human genome and its gene expressions might greatly enhance our understanding of cancer (Brown and Botstein, 1999). For example, the gene expressions of cancer tissue are very useful in helping develop predictions of the patient's survival time. Microarray technology allows for the measurement of the expression levels of thousands of genes simultaneously, thereby leading great power to gene expression based research such as those aforementioned. However, measurement error might be produced from various sources during the microarray experiment process. It is well known that ignoring measurement error could lead to substantially biased estimates of covariate coefficients, invalid hypothesis tests, or significantly masks the feature of the data (Carroll et al., 2006).

In survival analysis with microarray data, one of the main goals is to predict the survival time of future patients based on high dimensional gene expressions and patient specific covariates. When all covariates are observed accurately or error prone covariates have homoscedastic measurement error, i.e., the measurement error variance is assumed the same for all subjects, then there is no need to adjust the effect of measurement error in the prediction model. The survival time of future observation can be predicted as no measurement error case (Carroll et al., 2006). In practice, however, data may be collected under different conditions, making the observations associated with heteroscedastic measurement errors. Hence, a naive prediction model that ignores measurement error may not be appropriate (Carroll et al., 2009).

Microarray experiments are often performed over a long period of time on samples that are prepared and collected under different conditions. Moreover, different protocols or methodologies may be applied in the experiment, such as microarray print batches, array hybridization procedures, etc. Hence, all these factors contribute to a possibility of heteroscedastic measurement error associated with microarray data set. In practice, it is also important to combine microarray data from different labs or different platforms, which represents a natural way to increase sample size so that reliable statistical analysis may be conducted. In many statistical analyses, prediction is one of the ultimate goals. Thus, prediction of survival time under heteroscedastic measurement error is of primary importance.

Heteroscedastic measurement error models have been applied in epidemiology (Kulathinal et al., 2002) and analytical chemistry (Cheng and Riu, 2006) to avoid bias in parameter estimation. Carroll and Stefanski (1990) considered the heteroscedastic measurement error issue for generalized linear models. Devanarayan and Stefanski (2002) proposed an empirical simulation extrapolation (SIMEX) method for measurement error models with replicate measurements in the case of heteroscedastic measurement error. Recently, some work has been done on nonparametric regression estimation in the presence of heteroscedastic measurement error (Delaigle and Meister, 2007, 2008; Staudenmayer et al., 2008). However, survival analysis with heteroscedastic measurement error has not yet received much attention except for some sporadic studies, for example Carroll et al. (2009) investigated a nonparametric method to predict survival time in heteroscedastic measurement error models, and Augustin et al. (2008) considered regression calibration for the Cox proportional hazards (PH) model under heteroscedastic measurement error. Here we consider prediction of survival time for future observation under the accelerated failure time (AFT) model with covariates subject to heteroscedastic measurement error.

3.2 Methodology

3.2.1 Notation and Assumptions

Assume we have two types of covariates: $\mathbf{Z}_i$ consisting of the covariates that can be observed accurately and $\mathbf{X}_i$ consisting of those subject to measurement error. The AFT model specified by (1.1) can be rewritten as

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta_x} + \mathbf{Z}_i'\boldsymbol{\beta_z} + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where Y_i is the logarithm transformed survival time that may be subject to right censoring and $\boldsymbol{\beta} = (\boldsymbol{\beta}_x, \boldsymbol{\beta}_z)'$ is the vector of regression parameters. The intercept coefficient is incorporated with $\boldsymbol{\beta}_z$. Instead of observing $\mathbf{X}_i$, we observe its contaminated version $\mathbf{W}_i$. The relationship between $\mathbf{X}_i$ and $\mathbf{W}_i$ could be assumed through the classic additive measurement error model given by (1.2) with measurement error $\mathbf{U}_i \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{u_i})$. The components of $\mathbf{U}_i$ are independent (i.e., each $\boldsymbol{\Sigma}_{u_i}$ is a diagonal matrix); however, $\boldsymbol{\Sigma}_{u_i}$ might be different for each subject. For $i = 1, \ldots, n$, $\{\mathbf{X}_i, \mathbf{U}_i, \epsilon_i\}$ are mutually independent.

Assume that the future error prone covariates are contaminated by $\mathbf{W}_{F_i} = \mathbf{X}_{F_i} + \mathbf{U}_{F_i}$, $i = 1, \ldots, n_F$, where $\mathbf{U}_{F_i} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{u_{F_i}})$. The components of $\mathbf{U}_{F_i}$ are independent and $\{\mathbf{X}_{F_i}, \mathbf{U}_{F_i}\}$ are mutually independent. Also, $\boldsymbol{\Sigma}_{u_{F_i}}$ might be different for each subject.

Using the settings by Carroll et al. (2009), we further assume a two-error model. That is, the training data, $\{Y_i, \mathbf{W}_i, \mathbf{Z}_i\}$, $i = 1, \ldots, n$, has been rearranged such that the first m ($m \leq n$) observations have type 1 measurement error with covariance matrix $\boldsymbol{\Sigma}_{u_i} = \boldsymbol{\Sigma}_{uu_1}$, $i = 1, \ldots, m$; the rest $n - m$ observations have type 2 measurement error with covariance matrix $\boldsymbol{\Sigma}_{u_i} = \boldsymbol{\Sigma}_{uu_2}$, $i = m + 1, \ldots, n$; and all future observations have type 2 measurement error, $\boldsymbol{\Sigma}_{u_{F_i}} = \boldsymbol{\Sigma}_{uu_2}$, $i = 1, \ldots, n_F$.

The covariance matrix of the measurement error may be assumed known or replicated measurements of $\mathbf{W}_i$ are available.

3.2.2 The Effect of Measurement Error and Adjustment

The naive estimate $(\widehat{\boldsymbol{\beta}}_w, \widehat{\boldsymbol{\beta}}_z)'$, which is known to be inconsistent and asymptotically biased, can be obtained by solving the AFT model (3.1) with training data samples $\{Y_i, \mathbf{W}_i, \mathbf{Z}_i\}$ without adjusting the measurement error. The naive prediction model of the future observation with covariate $(\mathbf{W}_{F_i}, \mathbf{Z}_{F_i})$ is given by

$$\widehat{Y}_{F_i} = \mathbf{W}'_{F_i}\widehat{\boldsymbol{\beta}}_w + \mathbf{Z}'_{F_i}\widehat{\boldsymbol{\beta}}_z, \quad i = 1, \ldots, n_F.$$

We propose to use two variations of the SIMEX method applied to the samples $\{Y_i, \mathbf{W}_i, \mathbf{Z}_i\}$, to adjust the effect of the measurement error and obtain the estimates of the coefficients, $\widehat{\boldsymbol{\beta}}_x$ and $\widehat{\boldsymbol{\beta}}_z$. Then, we use the surrogate $(\mathbf{W}_i, \mathbf{W}_{F_i})$ and error-free variable $(\mathbf{Z}_i, \mathbf{Z}_{F_i})$ together to predict the corresponding unobserved future error prone covariate $\widehat{\mathbf{X}}_{F_i}$. Using the coefficient estimates computed from the training data, we can predict the survival time of future observation, $\widehat{Y}_{F_i}$ with covariate $(\mathbf{W}_{F_i}, \mathbf{Z}_{F_i})$ by

$$\widehat{Y}_{F_i} = \widehat{\boldsymbol{X}}'_{F_i}\widehat{\boldsymbol{\beta}}_x + \mathbf{Z}'_{F_i}\widehat{\boldsymbol{\beta}}_z, \quad i = 1, \ldots, n_F.$$

3.2.3 Two Variation of the SIMEX Algorithm

The SIMEX algorithm is a popular tool to adjust the effect of measurement error. See section 2.2.3 for detailed description. In this section, we briefly describe two variations of the SIMEX algorithm to calculate the SIMEX coefficient estimate $\widehat{\boldsymbol{\beta}}_{simex} = (\widehat{\boldsymbol{\beta}}_x, \widehat{\boldsymbol{\beta}}_z)'$.

3.2.3.1 Generalized SIMEX Method for Known Heteroscedastic Measurement Error

The generalized SIMEX algorithm is proposed to dealing with known heteroscedastic measurement error (Yi, 2010). Given an integer B and a grid of values $\mathbf{\Lambda} = \{\lambda_1, \ldots, \lambda_M\}$ with $\lambda_1 = 0,\ \lambda_i >= 0, i = 1, \ldots, M$. For each $\lambda \in \mathbf{\Lambda}$ and b from $1, \ldots, B$: Generate ξ_{bi} from a binomial distribution with a success probability $\frac{m}{n}$. If $\xi_{bi} = 1$, we generate pseudo errors from

$$\mathbf{U}_{bi}(\lambda) \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}_{uu_1});$$

else if $\xi_{bi} = 0$, the pseudo errors are generated from

$$\mathbf{U}_{bi}(\lambda) \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}_{uu_2}).$$

The pseudo data with increasing amount of measurement error are

$$\mathbf{W}_i(b, \lambda) = \mathbf{W}_i + \lambda^{\frac{1}{2}} \mathbf{U}_{bi}(\lambda).$$

We estimate the corresponding $\widehat{\boldsymbol{\beta}}(b, \lambda)$ by replacing $\mathbf{X}_i$ in AFT model (3.1) with $\mathbf{W}_i(b, \lambda)$ for each b, and then, average over b to obtain the SIMEX estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ for each fixed contamination level λ. Modelling the $\widehat{\boldsymbol{\beta}}(\lambda)$ as a function of λ and extrapolating back to the case $\lambda = -1$ results in the SIMEX estimate $\widehat{\boldsymbol{\beta}}_{simex}$.

3.2.3.2 Empirical SIMEX Method for Unknown Heteroscedstic Measurement Error

Consider the case where the covariance matrices for the measurement errors are not known but replicated measurements of $\mathbf{W}_i$ are available. Assume we have k replicated

measurements $\{\mathbf{W}_{i1}, \ldots, \mathbf{W}_{ik}\}$ for every subject i, such that

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k,$$

where $\mathbf{U}_{ij}$ are mutually independent from each other and are independent of $\{Y_i, \mathbf{X}_i, \mathbf{Z}_i\}$ for the training data. For fixed i, $\mathbf{U}_{ij}$ are independent and identically distributed measurement errors. The empirical SIMEX algorithm introducted in section 2.2.3 can be utilized to estimate $\widehat{\boldsymbol{\beta}}_{simex} = (\widehat{\boldsymbol{\beta}}_x, \widehat{\boldsymbol{\beta}}_z)'$.

3.2.4 Best Linear Prediction and Regression

Next, we briefly introduce the best linear prediction method to predict the unobserved error prone covariate. See Carroll et al. (2006) for technical details. Let X and Y be any two correlated random variables. The best linear predictor of Y based on X is

$$\widehat{Y} = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(X - \mu_x)$$

where μ_y is the mean of Y, σ_{xy} is the covariance between X and Y, μ_x is the mean of X and σ_x^2 is the variance of X.

For the case of the multiple linear regression model, the best linear predictor of Y based on vector of covariates $\mathbf{X}$ is

$$\widehat{Y} = \mu_y + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_x)$$

where $\boldsymbol{\Sigma}_{xy} = E\{[Y - E(Y)][\mathbf{X} - E(\boldsymbol{X})]^t\}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}_x$.

Consider the classic measurement error model (1.2), where $\mathbf{X}$ and $\mathbf{U}$ are uncorrelated and $E(\mathbf{U}) = \mathbf{0}$. If $\mathbf{X}$ and $\mathbf{Z}$ are independent, then the best linear predictor of $\mathbf{X}$ based on $\mathbf{W}$ is

$$\widehat{\mathbf{X}} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_x(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_u)^{-1}(\mathbf{W} - \boldsymbol{\mu}_x).$$

If the distribution of $\mathbf{X}$ depends on $\mathbf{Z}$, then the best linear predictor of $\mathbf{X}$ based on $\mathbf{W}$ is

$$\widehat{\mathbf{X}} = \boldsymbol{\mu}_x + \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xz} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz} & \boldsymbol{\Sigma}_z \end{pmatrix}^{-1} \left\{ \begin{pmatrix} \mathbf{W} \\ \mathbf{Z} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{pmatrix} \right\}$$

where $\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{\mu}_z$, $\mathrm{Var}(\boldsymbol{Z}) = \boldsymbol{\Sigma}_z$ and $\boldsymbol{\Sigma}_{xz}$ is the covariance matrix between $\mathbf{X}$ and $\mathbf{Z}$.

3.2.4.1 Best Linear Prediction for Known Heteroscedastic Measurement Error

When the variances of the heteroscedastic measurement errors are known, the unbiased estimate for $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$ can be calculated by

$$\widehat{\boldsymbol{\mu}}_x = \bar{\mathbf{W}} = \frac{\sum_{i=1}^{n} \mathbf{W}_i + \sum_{i=1}^{n_F} \mathbf{W}_{F_i}}{n + n_F},$$

$$\widehat{\boldsymbol{\mu}}_z = \frac{\sum_{i=1}^{n} \mathbf{Z}_i + \sum_{i=1}^{n_F} \mathbf{Z}_{F_i}}{n + n_F},$$

and

$$\widehat{\boldsymbol{\Sigma}}_z = \frac{\sum_{i=1}^{n} (\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)' + \sum_{i=1}^{n_F} (\mathbf{Z}_{F_i} - \widehat{\boldsymbol{\mu}}_z)(\mathbf{Z}_{F_i} - \widehat{\boldsymbol{\mu}}_z)'}{n + n_F - 1}.$$

Using the observations in the training data set with type 1 measurement error, $\boldsymbol{\Sigma}_{uu_1}$, we have the estimates of the $\boldsymbol{\Sigma}_{xz}$ and $\boldsymbol{\Sigma}_x$ given by

$$\widehat{\boldsymbol{\Sigma}}_{xz_1} = \frac{\sum_{i=1}^{m} (\mathbf{W}_i - \bar{\mathbf{W}})(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)}{m - 1}',$$

and

$$\widehat{\mathbf{\Sigma}}_{x_1} = \widehat{\mathbf{\Sigma}}_{w_1} - \mathbf{\Sigma}_{uu_1}.$$

where

$$\widehat{\mathbf{\Sigma}}_{w_1} = \frac{\sum_{i=1}^{m}(\mathbf{W}_i - \bar{\mathbf{W}})(\mathbf{W}_i - \bar{\mathbf{W}})'}{m-1}.$$

By merging the observations in the training data set with type 2 measurement error, $\mathbf{\Sigma}_{uu_2}$, and future observations, we have another set of estimates of the $\mathbf{\Sigma}_{xz}$ and $\mathbf{\Sigma}_x$ given by

$$\widehat{\mathbf{\Sigma}}_{xz_2} = \frac{\sum_{i=m+1}^{n}(\mathbf{W}_i - \bar{\mathbf{W}})(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)' + \sum_{i=1}^{n_F}(\mathbf{W}_{F_i} - \bar{\mathbf{W}})(\mathbf{Z}_{F_i} - \widehat{\boldsymbol{\mu}}_z)'}{n - m + n_F - 1},$$

and

$$\widehat{\mathbf{\Sigma}}_{x_2} = \widehat{\mathbf{\Sigma}}_{w_2} - \mathbf{\Sigma}_{uu_2}.$$

where

$$\widehat{\mathbf{\Sigma}}_{w_2} = \frac{\sum_{i=m+1}^{n}(\mathbf{W}_i - \bar{\mathbf{W}})(\mathbf{W}_i - \bar{\mathbf{W}})' + \sum_{i=1}^{n_F}(\mathbf{W}_{F_i} - \bar{\mathbf{W}})(\mathbf{W}_{F_i} - \bar{\mathbf{W}})'}{n - m + n_F - 1}.$$

Then, the pooled estimates of $\mathbf{\Sigma}_{xz}$ and $\mathbf{\Sigma}_x$ that will be used to predict $\mathbf{X}_F$ are given by

$$\widehat{\mathbf{\Sigma}}_{xz} = \frac{(m-1)\widehat{\mathbf{\Sigma}}_{xz_1} + (n - m + n_F - 1)\widehat{\mathbf{\Sigma}}_{xz_2}}{n + n_F - 2},$$

and

$$\widehat{\mathbf{\Sigma}}_{x} = \frac{(m-1)\widehat{\mathbf{\Sigma}}_{x_1} + (n - m + n_F - 1)\widehat{\mathbf{\Sigma}}_{x_2}}{n + n_F - 2}.$$

The estimate of $\mathbf{X}_F$ is $\widehat{\mathbf{X}}_F$ given by

$$\widehat{\mathbf{X}}_F = \widehat{\boldsymbol{\mu}}_x + \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_x & \widehat{\boldsymbol{\Sigma}}_{xz} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_x + \boldsymbol{\Sigma}_{uu_2} & \widehat{\boldsymbol{\Sigma}}_{xz} \\ \widehat{\boldsymbol{\Sigma}}_{xz} & \widehat{\boldsymbol{\Sigma}}_z \end{pmatrix}^{-1} \left\{ \begin{pmatrix} \mathbf{W}_F \\ \mathbf{Z}_F \end{pmatrix} - \begin{pmatrix} \widehat{\boldsymbol{\mu}}_x \\ \widehat{\boldsymbol{\mu}}_z \end{pmatrix} \right\}.$$

3.2.4.2 Best Linear Prediction for Unknown Heteroscedastic Measurement Error

When $\boldsymbol{\Sigma}_{uu_i}$ for $i = 1, 2$ are unknown but replicated measurements are available, we modify the best linear approximation method derived by Carroll and Stefanski (1990). Suppose for each $\mathbf{W}_i$ we have replicated measurements $\mathbf{W}_{i1}, \ldots, \mathbf{W}_{ik}$, $k > 1$ where

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, k;$$

and

$$\mathbf{W}_{F_{ij}} = \mathbf{X}_{F_i} + \mathbf{U}_{F_{ij}}, \quad i = 1, \ldots, n_F; \quad j = 1, \ldots, k.$$

The unbiased estimates for $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_{zz}$ are computed in the same way as the known measurement error variances case given by

$$\widehat{\boldsymbol{\mu}}_z = \frac{\sum_{i=1}^{n} \mathbf{Z}_i + \sum_{i=1}^{n_F} \mathbf{Z}_{F_i}}{n + n_F},$$

and

$$\widehat{\boldsymbol{\Sigma}}_z = \frac{\sum_{i=1}^{n} (\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)' + \sum_{i=1}^{n_F} (\mathbf{Z}_{F_i} - \widehat{\boldsymbol{\mu}}_z)(\mathbf{Z}_{F_i} - \widehat{\boldsymbol{\mu}}_z)'}{n + n_F - 1}.$$

For every object, the individual averages given by

$$\bar{\mathbf{W}}_i = \frac{1}{k} \sum_{j=1}^{k} \mathbf{W}_{ij} \text{ and } \bar{\mathbf{W}}_{F_i} = \frac{1}{k} \sum_{j=1}^{k} \mathbf{W}_{F_{ij}}.$$

would function as the surrogate for $\mathbf{X}_i$. The unbiased estimate of $\boldsymbol{\mu}_x$ can be calculated by

$$\widehat{\boldsymbol{\mu}}_x = \bar{\mathbf{W}} = \frac{\sum_{i=1}^{n} \bar{\mathbf{W}}_i + \sum_{i=1}^{n_F} \bar{\mathbf{W}}_{F_i}}{n + n_F}.$$

Similarly, we can obtain two sets of estimates of the $\boldsymbol{\Sigma}_{xz}$ and $\boldsymbol{\Sigma}_x$. Using the observations in the training data set with type 1 measurement error, we have

$$\widehat{\boldsymbol{\Sigma}}_{uu_1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)'}{m(k-1)},$$

$$\widehat{\boldsymbol{\Sigma}}_{x_1} = \frac{\sum_{i=1}^{m}(\bar{\mathbf{W}}_i - \bar{\mathbf{W}})(\bar{\mathbf{W}}_i - \bar{\mathbf{W}})'}{m-1} - \frac{\widehat{\boldsymbol{\Sigma}}_{uu_1}}{k},$$

$$\widehat{\boldsymbol{\Sigma}}_{xz_1} = \frac{\sum_{i=1}^{m}(\bar{\mathbf{W}}_i - \bar{\mathbf{W}})(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)'}{m-1},$$

By merging the observations in the training data set with type 2 measurement error and future observations, we have

$$\widehat{\boldsymbol{\Sigma}}_{uu_2} = \frac{\sum_{i=m+1}^{n}\sum_{j=1}^{k}(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)' + \sum_{i=1}^{n_F}\sum_{j=1}^{k}(\mathbf{W}_{F_{ij}} - \bar{\mathbf{W}}_{F_i})(\mathbf{W}_{F_{ij}} - \bar{\mathbf{W}}_{F_i})'}{(n-m+n_F)(k-1)},$$

$$\widehat{\boldsymbol{\Sigma}}_{x_2} = \frac{\sum_{i=m+1}^{n}(\bar{\mathbf{W}}_i - \bar{\mathbf{W}})(\bar{\mathbf{W}}_i - \bar{\mathbf{W}})' + \sum_{i=1}^{n_F}(\bar{\mathbf{W}}_{F_i} - \bar{\mathbf{W}})(\bar{\mathbf{W}}_{F_i} - \bar{\mathbf{W}})'}{n-m+n_F-1} - \frac{\widehat{\boldsymbol{\Sigma}}_{uu_2}}{k},$$

$$\widehat{\boldsymbol{\Sigma}}_{xz_2} = \frac{\sum_{i=m+1}^{n}(\bar{\mathbf{W}}_i - \bar{\mathbf{W}})(\mathbf{Z}_i - \widehat{\boldsymbol{\mu}}_z)' + \sum_{i=1}^{n_F}(\bar{\mathbf{W}}_{F_i} - \bar{\mathbf{W}})(\mathbf{Z}_{F_i} - \widehat{\boldsymbol{\mu}}_z)'}{n-m+n_F-1},$$

Then, the pooled estimates of $\boldsymbol{\Sigma}_{xz}$ and $\boldsymbol{\Sigma}_x$ that will be used to predict $\mathbf{X}_F$ are given

by

$$\widehat{\boldsymbol{\Sigma}}_x = \frac{(m-1)\widehat{\boldsymbol{\Sigma}}_{x_1} + (n-m+n_F-1)\widehat{\boldsymbol{\Sigma}}_{x_2}}{n+n_F-2},$$

$$\widehat{\boldsymbol{\Sigma}}_{xz} = \frac{(m-1)\widehat{\boldsymbol{\Sigma}}_{xz_1} + (n-m+n_F-1)\widehat{\boldsymbol{\Sigma}}_{xz_2}}{n+n_F-2}.$$

The estimate of $\mathbf{X}_F$ is $\widehat{\mathbf{X}}_F$ given by

$$\widehat{\mathbf{X}}_F = \widehat{\boldsymbol{\mu}}_x + \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_x & \widehat{\boldsymbol{\Sigma}}_{xz} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_x + \frac{1}{k}\widehat{\boldsymbol{\Sigma}}_{uu_2} & \widehat{\boldsymbol{\Sigma}}_{xz} \\ \widehat{\boldsymbol{\Sigma}}_{xz} & \widehat{\boldsymbol{\Sigma}}_z \end{pmatrix}^{-1} \left\{ \begin{pmatrix} \bar{\mathbf{W}}_F \\ \mathbf{Z}_F \end{pmatrix} - \begin{pmatrix} \widehat{\boldsymbol{\mu}}_x \\ \widehat{\boldsymbol{\mu}}_z \end{pmatrix} \right\}.$$

3.2.5 Prediction Accuracy Criteria

The performance of the proposed SIMEX adjusted prediction models and the impact of naive prediction model are evaluated by the mean squared prediction error (MSPE)

$$\text{MSPE} = E\left(\sum_{i=1}^{n_F} \left(Y_i - \widehat{Y}_i \right)^2 \right)$$

where n_F is the total number of future observations and $\widehat{Y}_i$ is the logarithm transformed survival time predicted by the AFT model. Due to censoring, some of the true survival times are not observed, such that the censored survival time is shorter than the true potential survival time. We consider the following three methods to make transformation of the censored survival time $(Y_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)$ to $(t_i^*, \mathbf{X}_i, \mathbf{Z}_i)$ according to the rules

$$t_i^* = \delta_i \phi_1(Y_i, \mathbf{X}_i, \mathbf{Z}_i) + (1-\delta_i)\phi_2(Y_i, \mathbf{X}_i, \mathbf{Z}_i).$$

A basic requirement for this transformation is to make $E(t_i^*|\mathbf{X}_i, \mathbf{Z}_i) = E(Y_i|\mathbf{X}_i, \mathbf{Z}_i)$. According to Jin and He (2010), the following three adjustment methods satisfy the above condition. The consistent estimator of the mean squared prediction error of

adjusted survival time is

$$\text{MSPE} = \frac{1}{n_F}\sum_{i=1}^{n_F}\left(t_i^* - \widehat{Y}_i\right)^2.$$

3.2.5.1 Inverse Probability Weights (IPW)

In the Inverse Probability Weights (IPW) method, let

$$\phi_1(Y, \mathbf{X}, \mathbf{Z}) = \frac{Y}{\bar{G}(Y|\mathbf{X}, \mathbf{Z})},$$

and

$$\phi_2(Y, \mathbf{X}, \mathbf{Z}) = 0.$$

Then, we have

$$t_i^* = \frac{\delta_i Y_i}{\bar{G}(Y_i|\mathbf{X}_i, \mathbf{Z}_i)},$$

where $\bar{G}_i(t)$ is the Kaplan-Meier estimator of censored time.

3.2.5.2 Integral

In the Integral method, let

$$\phi_1(Y, \mathbf{X}, \mathbf{Z}) = \phi_2(Y, \mathbf{X}, \mathbf{Z}) = \int_0^Y \frac{dt}{\bar{G}(t|\mathbf{X}, \mathbf{Z})},$$

So, we have

$$t_i^* = \int_0^{Y_i} \frac{dt}{\bar{G}(t|\mathbf{X}_i, \mathbf{Z}_i)}.$$

3.2.5.3 Buckley James (BJ)

Another transformation comes from the idea of Buckley and James (1979). Let

$$\phi_1(Y, \mathbf{X}, \mathbf{Z}) = Y,$$

and

$$\phi_2(Y, \mathbf{X}, \mathbf{Z}) = E(Y_i | Y_i > Y, \mathbf{X}, \mathbf{Z}) = \frac{\sum\limits_{i:Y_i>Y} \delta_i Y_i}{\sum\limits_{i:Y_i>Y} \delta_i}.$$

Then, we have

$$t_i^* = \delta_i Y_i + (1 - \delta_i) \frac{\sum\limits_{j:Y_j>Y_i} \delta_j Y_j}{\sum\limits_{j:Y_j>Y_i} \delta_j}.$$

3.3 Simulation Study

We conduct simulation studies to evaluate the proposed SIMEX adjusted prediction method. We generate independent observations from the AFT model under the Weibull distribution with the survival function given by

$$S(t) = \exp\left(-t^{\alpha} e^{\mathbf{X_i}\boldsymbol{\beta}_x + \mathbf{Z_i}\boldsymbol{\beta}_z}\right)$$

where $S(t)$ is generated independently from the uniform distribution Unif$[0, 1]$. The censoring times are generated from the exponential distribution with a fixed parameter to achieve 0%, 10%, 30%, 50% and 70% censoring rate. Two values of α, $\alpha = 0.5$ and $\alpha = 1.5$, which represent the decreasing and increasing hazard rates of the Weibull model, are considered. Each simulation study consists of 100 data sets of size $n = 125$ with $m = 100$ samples having type 1 measurement error, $\boldsymbol{\Sigma}_{uu_1}$, and the rest of the 25 samples having type 2 measurement error, $\boldsymbol{\Sigma}_{uu_2}$. The future covariates have a size of $n_F = 50$ with type 2 measurement error. The variances of the measurement

errors are known or replicated measurements of $\mathbf{W}_i$ are available. For all simulation scenarios, we set $\lambda \in [0, 2]$ and $B = 100$ for the SIMEX algorithm. The MSPEs of the naive prediction and the SIMEX adjusted prediction models are calculated for each simulation run. The mean and standard error (SE) of these MSPEs under each scenario are calculated to compare the performance of the proposed method and the naive method.

3.3.1 X and Z are Independent

3.3.1.1 Heteroscedastic Measurement Error with Known Measurement Error Variance

In this simulation study, the true values of the covariate coefficients are $\beta_z = 0.5$ and $\beta_x = -\log(2)$. We generate z_i from a Bernoulli distribution with 50% probability of success and x_i follows a normal distribution $\mathrm{N}(1, 1)$. The observed surrogate is $w_i = x_i + u_i$. The first 100 observations have a type 1 measurement error $u_i = \mathrm{N}(0, 0.25^2)$. The remaining 25 observations of the training data and 50 future observations have type 2 measurement error $u_i = \mathrm{N}(0, 1^2)$. Here we assume that the measurement error for the future observation is severe than the training data set.

Tables 3.1 and 3.2 list the mean and SE of the MSPE of the naive and the SIMEX adjusted prediction models. Three methods to adjust the censored survival times are applied and the mean and SE of those MSPEs are reported as well.

For both naive and SIMEX adjusted prediction models with $\alpha = 0.5$, as the censoring rate increases the actual MSPE increases and the corresponding SE becomes larger. The actual MSPE that all the survival time are observed is calculated under each censoring rate. When there is no censoring, all three censor adjustment methods give the same MSPE and the corresponding SE; under low and median censoring rates (10% to 50%), all three censored survival time adjustment methods underestimate the MSPE. Compared to the other two methods, IPW method is more consistent

Table 3.1: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are independent, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	8.249	2.149	8.249	2.149	8.249	2.149	8.249	2.149
	10%	7.983	2.256	7.986	2.178	8.045	2.162	8.240	2.177
	30%	7.696	2.712	7.527	2.213	7.192	2.187	8.249	2.201
	50%	8.302	3.584	7.562	2.497	6.562	2.447	8.336	2.306
	70%	11.444	6.204	9.522	3.491	7.418	3.544	8.590	2.549
SIMEX	0%	7.848	2.076	7.848	2.076	7.848	2.076	7.848	2.076
	10%	7.484	2.109	7.509	2.053	7.583	2.044	7.845	2.088
	30%	7.015	2.520	6.871	2.071	6.553	2.043	7.859	2.104
	50%	7.567	3.323	6.669	2.235	5.673	2.163	7.900	2.108
	70%	11.040	6.015	8.408	3.054	6.284	3.040	8.082	2.265

Table 3.2: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are independent, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.917	0.239	0.917	0.239	0.917	0.239	0.917	0.239
	10%	0.916	0.250	1.079	0.372	0.915	0.245	0.918	0.241
	30%	0.890	0.247	1.443	0.501	0.866	0.246	0.920	0.245
	50%	0.882	0.316	1.953	0.652	0.786	0.250	0.928	0.254
	70%	0.937	0.452	2.833	0.862	0.754	0.323	0.951	0.282
SIMEX	0%	0.872	0.231	0.872	0.231	0.872	0.231	0.872	0.231
	10%	0.867	0.234	1.025	0.366	0.862	0.231	0.872	0.231
	30%	0.839	0.232	1.375	0.493	0.797	0.228	0.874	0.231
	50%	0.811	0.281	1.861	0.642	0.690	0.229	0.880	0.238
	70%	0.851	0.398	2.710	0.839	0.621	0.261	0.893	0.244

with the actual MSPE at the expense of higher SE; when the censoring rate is high, IPW and Integral methods overestimate the MSPE and SE, while the BJ method underestimates the MSPE but with larger SE estimator. BJ method adjusted survival time works best under low censoring rate and Integral method is the optimal choice under heavy censoring rate.

For both naive and SIMEX adjusted prediction models with $\alpha = 1.5$, as the censoring rate increases the actual MSPE increases and the corresponding SE becomes larger. When there is no censoring, all three censored survival time adjustment methods give the same MSPE and the corresponding SE. As long as there is censored survival time, Integral method overestimates the actual MSPE and the corresponding SE while IPW and BJ methods underestimate the MSPE. Specifically, for the naive prediction method, compared to the BJ adjustment method, IPW method is more consistent with the actual MSPE at the expense of higher SE, while for the SIMEX adjusted prediction method, IPW method is more consisted in estimating the mean and SE of the MSPE.

It is clear from Tables 3.1 and 3.2 that the proposed prediction model outperforms the naive prediction model. Under each censoring rate, the proposed method has better prediction accuracy with small mean and SE of the MSPE. In order to check the performance of the proposed method, we compare four prediction models with no censored survival time. First of all, comparison is made between our proposed SIMEX adjusted prediciton model using $\hat{x}\hat{\beta}_x$ and the naive prediction model using $w\hat{\beta}_w$ in Table 3.3. Our proposed prediction model outperforms the naive model in that it gives smaller mean and SE of the MSPE. Secondly, comparing the prediction model using $x\hat{\beta}_x$ to the prediction model using $x\hat{\beta}_w$, it confirms that the naive approach gives biased estimates and the SIMEX approach corrects the bias, giving smaller MSPE and SE. Similarly, while comparing MSPE difference of $w\hat{\beta}_w$ and $x\hat{\beta}_x$ to the MSPE difference of $\hat{x}\hat{\beta}_x$ and $x\hat{\beta}_x$, it shows that $\hat{x}$ performs better than naively use w directly.

Table 3.3: Comparison of mean squared prediction errors: Scenario 3.1.1

Parameter	$w\hat{\beta}_w$		$x\hat{\beta}_w$		$\hat{x}\hat{\beta}_x$		$x\hat{\beta}_x$	
	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
$\alpha = 0.5$	8.249	2.149	6.894	1.885	7.848	2.076	6.829	1.879
$\alpha = 1.5$	0.917	0.239	0.766	0.209	0.872	0.231	0.759	0.209

3.3.1.2 Heteroscedastic Measurement Error with Replicate Measurements

When the covariance matrixes of the measurement errors are unknown, but replicates of the w_i are available. Here we assume each subject has $k = 2$ replicate measurements. All the parameter settings are kept the same as the known variances case.

Tables 3.4 and 3.5 show the mean MSPEs and the corresponding SE of the naive prediction model and the empirical SIMEX adjusted prediction model. The proposed empirical SIMEX adjusted prediction model performs better than the naive model, the difference is smaller compared to the known error case. Both naive and the emprical SIMEX adjusted prediction models perform better than the known error case. This is because with replicated measurements, the variance of the surrogate is smaller than the case of the known covariance matrix. Similarly, we compare the four prediction models with no censoring in the survival time. It confirms that the empirical SIMEX adjusted prediction model outperforms the naive prediction model.

3.3.2 X and Z are Independent but X are Correlated

In cases where gene expressions from the microarray are correlated with each other, we next simulate a scenario where $\mathbf{X}_i$ and z_i are independent, but the components of $\mathbf{X}_i$ are correlated. Let z_i be generated from a Bernoulli distribution with 50% probability of success and $\mathbf{X}_i = (x_{i_1}, x_{i_2})'$ be generated from the multivariate normal distribution with mean $(1, 1)'$. Let their coefficients be $\beta_z = 0.5$ and $\boldsymbol{\beta}_x = (-\log(2), -1)'$, respectively. Let the correlation, ρ, between the components of $\mathbf{X}_i$ be $(0.8, 0.3, -0.3, -0.8)'$ to represent moderate and heavy correlation.

Table 3.4: Comparision of mean squared prediction errors between the NAIVE and SIMEX methods: Unknown variance, x and z are independent, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	7.494	2.448	7.494	2.448	7.494	2.448	7.494	2.448
	10%	7.351	2.476	7.316	2.418	7.365	2.423	7.488	2.446
	30%	7.230	2.836	7.081	2.514	6.753	2.495	7.569	2.548
	50%	7.916	3.732	7.262	2.692	6.145	2.587	7.636	2.553
	70%	11.151	5.644	9.493	3.881	7.326	3.602	7.896	2.736
SIMEX	0%	7.316	2.456	7.316	2.456	7.316	2.456	7.316	2.456
	10%	7.122	2.491	7.093	2.448	7.157	2.443	7.315	2.465
	30%	6.884	2.809	6.726	2.517	6.423	2.485	7.380	2.540
	50%	7.545	3.714	6.762	2.686	5.669	2.584	7.440	2.535
	70%	11.066	5.711	8.826	3.701	6.663	3.413	7.662	2.646

Table 3.5: Comparision of mean squared prediction errors between the NAIVE and SIMEX methods: Unknown variance, x and z are independent, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.833	0.272	0.833	0.272	0.833	0.272	0.833	0.272
	10%	0.834	0.280	0.987	0.348	0.830	0.273	0.833	0.273
	30%	0.832	0.281	1.384	0.480	0.793	0.250	0.842	0.283
	50%	0.839	0.304	1.928	0.603	0.733	0.247	0.849	0.280
	70%	0.906	0.455	2.790	0.770	0.719	0.309	0.868	0.293
SIMEX	0%	0.813	0.273	0.813	0.273	0.813	0.273	0.813	0.273
	10%	0.815	0.280	0.963	0.352	0.807	0.274	0.814	0.274
	30%	0.805	0.277	1.346	0.481	0.757	0.252	0.821	0.284
	50%	0.803	0.295	1.873	0.600	0.682	0.249	0.827	0.280
	70%	0.871	0.444	2.731	0.765	0.650	0.298	0.843	0.290

Table 3.6: Comparison of mean squared prediction errors: Scenario 3.1.2

Parameter	$w\hat{\beta}_w$		$x\hat{\beta}_w$		$\hat{x}\hat{\beta}_x$		$x\hat{\beta}_x$	
	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
α =0.5	7.494	2.448	6.708	2.452	7.316	2.456	6.708	2.461
$\alpha = 1.5$	0.833	0.272	0.745	0.272	0.813	0.273	0.745	0.273

$$\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

For the type 1 measurement error, we assume it comes from multivariate normal distribution specified by

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.25^2 & 0 \\ 0 & 0.25^2 \end{pmatrix}\right)$$

For the type 2 measurement error, we assume

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

3.3.2.1 Heteroscedastic Measurement Error with Known Measurement Error Variance

Consider the case that the measurement error variances are known. The mean of the MSPEs of 100 runs and the corresponding SE are reported from Tables 3.8 to 3.15 under each parameter combination. As the censoring rate increases the MSPE calculated from the real survival times increases and the corresponding SE becomes larger. When there is no censoring, all three censored survival time adjustment methods give the same MSPE and the corresponding SE. As long as there is censored survival time, three survival time adjustment methods give different MSPE and SE. No matter which method applied to adjust the censored survival time, the SIMEX adjusted

prediction model works better than the naive prediction model; the proposed prediction method has better prediction accuracy with smaller mean and SE of MSPE. The strength of the correlation between covariate $\mathbf{X}_i$ will affect the prediction accuracy; as the correlation reduces from 0.8 to -0.8, the mean and SE of the MSPE become smaller.

We compare four prediction models with no censored survival time to check the performance of the proposed method. The correlation between the component of $\mathbf{X}_i$ is 0.8. Other strength of the correlation should give similar conclusions. First of all, comparison is made between our proposed SIMEX adjusted prediction model using $\widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}}_x$ and the naive prediction model using $\mathbf{W}\widehat{\boldsymbol{\beta}}_w$ in table 3.7. Our proposed prediction model outperforms the naive prediction model in giving smaller mean MSPE with less variation. Secondly, by comparing the prediction model using $\mathbf{X}\widehat{\boldsymbol{\beta}}_x$ to the prediction model using $\mathbf{X}\widehat{\boldsymbol{\beta}}_w$, it confirms that the naive approach gives biased estimates and the SIMEX method corrects the bias, giving smaller MSPE and corresponding SE. Lastly, the results from comparing our proposed prediction model to the prediction model using $\mathbf{X}\widehat{\boldsymbol{\beta}}_x$ suggests that the performance of the proposed prediction model might be improved by better predicting on the unobservable $\mathbf{X}_F$.

Table 3.7: Comparison of mean squared pridiction errors: Scenario 3.2.1

Parameter	$w\hat{\beta}_w$		$x\hat{\beta}_w$		$\hat{x}\hat{\beta}_x$		$x\hat{\beta}_x$	
	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
α =0.5	11.352	2.589	6.860	2.059	10.313	2.405	6.779	2.049
α = 1.5	1.261	0.288	0.762	0.229	1.146	0.267	0.753	0.228

Table 3.8: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is 0.8, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	11.352	2.589	11.352	2.589	11.352	2.589	11.352	2.589
	10%	11.180	2.984	11.023	2.731	10.974	2.699	11.292	2.615
	30%	11.149	3.276	11.167	2.980	10.734	2.963	11.437	2.678
	50%	11.776	4.217	12.784	3.438	11.497	3.306	11.405	2.576
	70%	13.717	5.934	18.779	6.718	16.548	6.746	11.847	2.934
SIMEX	0%	10.313	2.405	10.313	2.405	10.313	2.405	10.313	2.405
	10%	9.841	2.687	9.711	2.457	9.715	2.452	10.328	2.426
	30%	9.425	3.067	9.008	2.558	8.614	2.554	10.410	2.410
	50%	10.209	3.916	9.900	2.958	8.652	2.871	10.442	2.389
	70%	13.253	5.966	15.078	5.800	12.833	5.819	10.793	2.442

Table 3.9: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is 0.8, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	1.261	0.288	1.261	0.288	1.261	0.288	1.261	0.288
	10%	1.255	0.312	1.415	0.352	1.287	0.310	1.260	0.287
	30%	1.237	0.320	1.852	0.500	1.274	0.318	1.268	0.293
	50%	1.237	0.391	2.512	0.678	1.318	0.370	1.282	0.298
	70%	1.323	0.547	3.809	1.021	1.651	0.600	1.316	0.333
SIMEX	0%	1.146	0.267	1.146	0.267	1.146	0.267	1.146	0.267
	10%	1.135	0.287	1.274	0.309	1.156	0.286	1.148	0.268
	30%	1.081	0.291	1.639	0.434	1.071	0.280	1.152	0.269
	50%	1.050	0.343	2.209	0.604	1.008	0.301	1.164	0.267
	70%	1.147	0.487	3.364	0.908	1.203	0.495	1.177	0.277

Table 3.10: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is 0.3, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	10.942	2.556	10.942	2.556	10.942	2.556	10.942	2.556
	10%	10.723	2.860	10.582	2.642	10.560	2.624	10.900	2.571
	30%	10.495	3.082	10.413	2.846	10.003	2.838	11.006	2.624
	50%	11.136	3.676	11.354	3.153	10.213	3.025	11.010	2.584
	70%	13.455	5.889	15.942	5.484	13.747	5.412	11.353	2.965
SIMEX	0%	9.935	2.350	9.935	2.350	9.935	2.350	9.935	2.350
	10%	9.357	2.608	9.349	2.395	9.376	2.409	9.942	2.371
	30%	8.809	2.906	8.445	2.498	8.079	2.485	10.017	2.376
	50%	9.530	3.520	8.763	2.740	7.652	2.657	10.046	2.332
	70%	12.964	5.785	12.582	4.487	10.361	4.427	10.277	2.407

Table 3.11: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is 0.3, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	1.216	0.284	1.216	0.284	1.216	0.284	1.216	0.284
	10%	1.210	0.303	1.362	0.342	1.236	0.299	1.217	0.284
	30%	1.198	0.331	1.803	0.492	1.197	0.314	1.223	0.291
	50%	1.177	0.361	2.391	0.644	1.198	0.344	1.240	0.296
	70%	1.265	0.496	3.538	0.950	1.413	0.506	1.275	0.327
SIMEX	0%	1.104	0.261	1.104	0.261	1.104	0.261	1.104	0.261
	10%	1.097	0.278	1.230	0.304	1.111	0.278	1.106	0.264
	30%	1.048	0.303	1.603	0.435	1.009	0.281	1.109	0.264
	50%	0.996	0.326	2.110	0.579	0.913	0.289	1.118	0.267
	70%	1.104	0.437	3.151	0.843	1.014	0.402	1.139	0.268

Table 3.12: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is -0.3, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	10.117	2.475	10.117	2.475	10.117	2.475	10.117	2.475
	10%	9.830	2.757	9.796	2.560	9.821	2.566	10.114	2.505
	30%	9.603	2.956	9.358	2.712	9.009	2.737	10.193	2.570
	50%	10.033	3.800	9.754	3.008	8.692	2.906	10.291	2.610
	70%	12.574	5.407	12.430	4.362	10.120	4.091	10.450	2.922
SIMEX	0%	9.097	2.254	9.097	2.254	9.097	2.254	9.097	2.254
	10%	8.51	2.459	8.622	2.304	8.692	2.342	9.116	2.280
	30%	7.983	2.818	7.681	2.423	7.355	2.434	9.166	2.317
	50%	8.490	3.563	7.557	2.624	6.534	2.559	9.211	2.287
	70%	11.907	5.239	9.801	3.551	7.476	3.365	9.337	2.351

Table 3.13: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is -0.3, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	1.124	0.275	1.124	0.275	1.124	0.275	1.124	0.275
	10%	1.124	0.289	1.273	0.334	1.131	0.290	1.126	0.274
	30%	1.090	0.319	1.647	0.462	1.080	0.298	1.135	0.283
	50%	1.086	0.345	2.229	0.609	1.023	0.320	1.151	0.297
	70%	1.147	0.477	3.155	0.819	1.074	0.398	1.177	0.308
SIMEX	0%	1.011	0.250	1.011	0.250	1.011	0.250	1.011	0.250
	10%	1.012	0.268	1.144	0.297	1.007	0.267	1.014	0.254
	30%	0.953	0.289	1.474	0.414	0.914	0.266	1.017	0.253
	50%	0.915	0.308	1.996	0.553	0.786	0.273	1.026	0.258
	70%	0.983	0.402	2.857	0.746	0.763	0.330	1.042	0.258

Table 3.14: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, $\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is -0.8, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	8.333	2.275	8.333	2.275	8.333	2.275	8.333	2.275
	10%	7.974	2.392	8.038	2.282	8.131	2.319	8.315	2.236
	30%	7.702	2.718	7.565	2.416	7.242	2.373	8.387	2.337
	50%	8.027	3.260	7.579	2.639	6.589	2.569	8.521	2.402
	70%	10.995	5.640	9.111	3.710	6.949	3.513	8.780	2.697
SIMEX	0%	7.706	2.148	7.706	2.148	7.706	2.148	7.706	2.148
	10%	7.331	2.267	7.383	2.165	7.486	2.188	7.714	2.143
	30%	6.823	2.627	6.748	2.283	6.446	2.241	7.752	2.178
	50%	7.173	3.079	6.496	2.473	5.505	2.414	7.779	2.193
	70%	10.513	5.636	7.791	3.182	5.617	3.058	7.934	2.223

Table 3.15: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known,$\mathbf{X}$ and z are independent, the correlation between $\mathbf{X}$ is -0.8, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.926	0.253	0.926	0.253	0.926	0.253	0.926	0.253
	10%	0.921	0.263	1.071	0.306	0.912	0.260	0.927	0.254
	30%	0.912	0.284	1.447	0.427	0.861	0.266	0.935	0.260
	50%	0.896	0.318	2.001	0.579	0.787	0.290	0.958	0.288
	70%	0.952	0.393	2.827	0.766	0.754	0.329	0.983	0.299
SIMEX	0%	0.856	0.239	0.856	0.239	0.856	0.239	0.856	0.239
	10%	0.852	0.247	0.996	0.280	0.840	0.243	0.858	0.240
	30%	0.831	0.263	1.353	0.401	0.769	0.245	0.859	0.240
	50%	0.786	0.280	1.871	0.542	0.658	0.253	0.863	0.241
	70%	0.841	0.365	2.667	0.708	0.595	0.290	0.876	0.249

3.3.3 Distribution of X Depends on Z

In this simulation scenario, we consider the case that x_i and z_i are correlated, $\beta_x = -\log(2)$ and $\beta_z = 0.5$. The covariates follow the multivariate normal distribution with x_i and z_i having the same mean and variance as the independent case. The correlation between x_i and z_i is set to be $\rho = (0.5, 0.3, -0.3, -0.5)'$. The type 1 measurement error has variance of 0.25^2 and type 2 measurement error has variance of 1^2.

$$\begin{pmatrix} x_i \\ z_i \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 0.25 \end{pmatrix}\right)$$

Here, the measurement error variances are known. The mean of the MSPEs of 100 runs and the corresponding SE are reported from Tables 3.16 to 3.23. Three survival time adjustment methods are applied to calculate the mean MSPE and the corresponding SE under different censoring rates. As the censoring rate increases, the mean and SE of the MSPE becomes larger. The proposed SIMEX adjusted prediction model works better than the naive prediction model with better prediction accuracy. It gives smaller MSPE and SE. The strength of the correlation between covariate x_i and z_i affects the prediction accuracy. When the correlation is ± 0.5, the mean MSPE and the corresponding SE are similar and are smaller than the ± 0.3 correlation cases.

Table 3.16: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is 0.5, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	6.872	2.382	6.872	2.382	6.872	2.382	6.872	2.382
	10%	6.777	2.424	6.747	2.382	6.816	2.404	6.896	2.379
	30%	6.539	2.692	6.403	2.301	6.124	2.394	6.927	2.347
	50%	7.185	3.708	6.466	2.331	5.557	2.311	7.097	2.331
	70%	10.688	6.303	8.020	3.011	6.019	2.964	7.453	2.563
SIMEX	0%	6.722	2.367	6.722	2.367	6.722	2.367	6.722	2.367
	10%	6.611	2.400	6.579	2.376	6.648	2.404	6.731	2.373
	30%	6.381	2.736	6.207	2.330	5.932	2.423	6.735	2.368
	50%	6.917	3.742	6.169	2.348	5.259	2.333	6.791	2.358
	70%	10.109	6.070	7.478	2.828	5.477	2.744	6.950	2.344

Table 3.17: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is 0.5, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.764	0.265	0.764	0.265	0.764	0.265	0.764	0.265
	10%	0.758	0.272	0.905	0.306	0.751	0.269	0.764	0.266
	30%	0.761	0.267	1.270	0.385	0.711	0.264	0.771	0.260
	50%	0.746	0.307	1.776	0.502	0.633	0.261	0.782	0.266
	70%	0.898	0.488	2.626	0.659	0.652	0.308	0.847	0.292
SIMEX	0%	0.747	0.263	0.747	0.263	0.747	0.263	0.747	0.263
	10%	0.743	0.267	0.889	0.308	0.735	0.267	0.748	0.264
	30%	0.743	0.269	1.247	0.392	0.689	0.268	0.750	0.264
	50%	0.718	0.304	1.743	0.496	0.604	0.257	0.754	0.260
	70%	0.813	0.467	2.552	0.644	0.575	0.285	0.774	0.261

Table 3.18: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is 0.3, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	7.860	2.520	7.860	2.520	7.860	2.520	7.860	2.520
	10%	7.708	2.589	7.680	2.521	7.739	2.528	7.879	2.522
	30%	7.580	2.871	7.281	2.413	6.989	2.453	7.907	2.464
	50%	8.038	4.001	7.198	2.430	6.225	2.428	7.938	2.429
	70%	11.515	6.694	8.820	3.239	6.828	3.173	8.159	2.514
SIMEX	0%	7.502	2.449	7.502	2.449	7.502	2.449	7.502	2.449
	10%	7.239	2.502	7.266	2.455	7.335	2.465	7.511	2.461
	30%	6.947	2.825	6.711	2.383	6.446	2.423	7.527	2.439
	50%	7.431	3.924	6.489	2.415	5.531	2.415	7.555	2.427
	70%	11.044	6.523	7.925	2.975	5.929	2.920	7.736	2.416

Table 3.19: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is 0.3, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.873	0.280	0.873	0.280	0.873	0.280	0.873	0.280
	10%	0.870	0.282	1.011	0.324	0.863	0.283	0.874	0.282
	30%	0.870	0.297	1.363	0.402	0.815	0.285	0.878	0.279
	50%	0.844	0.308	1.864	0.503	0.728	0.271	0.881	0.271
	70%	0.967	0.562	2.702	0.672	0.689	0.291	0.904	0.267
SIMEX	0%	0.834	0.272	0.834	0.272	0.834	0.272	0.834	0.272
	10%	0.830	0.271	0.967	0.310	0.819	0.273	0.835	0.273
	30%	0.818	0.282	1.306	0.393	0.755	0.276	0.836	0.272
	50%	0.787	0.297	1.793	0.497	0.655	0.267	0.841	0.269
	70%	0.891	0.558	2.612	0.667	0.596	0.292	0.861	0.268

Table 3.20: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is -0.3, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	7.831	2.434	7.831	2.434	7.831	2.434	7.831	2.434
	10%	7.659	2.512	7.636	2.435	7.664	2.422	7.851	2.431
	30%	7.759	2.849	7.470	2.401	7.189	2.417	7.945	2.406
	50%	8.540	3.737	7.728	2.562	6.716	2.521	7.994	2.371
	70%	11.452	7.146	10.052	3.702	8.063	3.648	8.179	2.489
SIMEX	0%	7.535	2.378	7.535	2.378	7.535	2.378	7.535	2.378
	10%	7.288	2.481	7.284	2.387	7.328	2.372	7.542	2.377
	30%	7.244	2.820	6.930	2.396	6.672	2.383	7.582	2.379
	50%	7.941	3.684	7.049	2.440	6.059	2.405	7.634	2.318
	70%	11.009	6.957	9.240	3.504	7.254	3.447	7.826	2.389

Table 3.21: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is -0.3, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.870	0.270	0.870	0.270	0.870	0.270	0.870	0.270
	10%	0.869	0.277	1.005	0.322	0.873	0.272	0.872	0.270
	30%	0.873	0.289	1.369	0.409	0.841	0.283	0.881	0.270
	50%	0.868	0.309	1.901	0.505	0.776	0.266	0.887	0.264
	70%	0.982	0.531	2.806	0.685	0.795	0.301	0.912	0.264
SIMEX	0%	0.837	0.264	0.837	0.264	0.837	0.264	0.837	0.264
	10%	0.836	0.270	0.966	0.313	0.837	0.264	0.839	0.264
	30%	0.824	0.284	1.312	0.402	0.785	0.275	0.841	0.263
	50%	0.814	0.307	1.829	0.499	0.707	0.263	0.848	0.261
	70%	0.915	0.523	2.714	0.669	0.702	0.295	0.865	0.256

Table 3.22: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is -0.5, $\alpha = 0.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	6.879	2.363	6.879	2.363	6.879	2.363	6.879	2.363
	10%	6.732	2.481	6.731	2.334	6.762	2.322	6.886	2.340
	30%	6.948	2.832	6.694	2.386	6.394	2.356	6.965	2.355
	50%	7.963	3.838	7.374	2.562	6.341	2.448	7.146	2.314
	70%	10.849	7.077	10.534	4.140	8.460	4.083	7.612	2.323
SIMEX	0%	6.762	2.344	6.762	2.344	6.762	2.344	6.762	2.344
	10%	6.596	2.459	6.586	2.331	6.614	2.321	6.737	2.341
	30%	6.786	2.847	6.517	2.359	6.211	2.339	6.772	2.338
	50%	7.653	3.819	7.084	2.506	6.047	2.408	6.838	2.314
	70%	10.304	7.156	10.019	4.014	7.956	3.954	7.010	2.345

Table 3.23: Comparison of mean squared prediction errors between the NAIVE and SIMEX methods: Variance known, x and z are correlated, the correlation is -0.5, $\alpha = 1.5$.

Method	Censoring	IPW		Integral		BJ		Real survival time	
		E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)	E(MSPE)	SE(MSPE)
NAIVE	0%	0.764	0.263	0.764	0.263	0.764	0.263	0.764	0.263
	10%	0.759	0.266	0.900	0.311	0.774	0.261	0.765	0.261
	30%	0.757	0.274	1.271	0.399	0.754	0.271	0.771	0.258
	50%	0.770	0.301	1.842	0.503	0.720	0.268	0.786	0.258
	70%	0.917	0.543	2.801	0.684	0.802	0.338	0.840	0.278
SIMEX	0%	0.747	0.260	0.747	0.260	0.747	0.260	0.747	0.260
	10%	0.744	0.264	0.884	0.313	0.757	0.261	0.749	0.260
	30%	0.741	0.274	1.251	0.404	0.737	0.270	0.751	0.259
	50%	0.741	0.297	1.807	0.499	0.689	0.257	0.755	0.255
	70%	0.836	0.518	2.732	0.659	0.733	0.306	0.769	0.250

3.4 Conclusion

Prediction is one of the goals for many statistical analyses. When we have accurately measured covariates only or homogenetic measurement errors in the error prone covariates in the AFT model, prediction can be done the same as the model with no measurement error because no adjustments are needed. However, when the covariates are subject to heteroscedastic measurement error, the naive prediction model without adjustments to the effect of measurement error is not appropriate. In microarray studies, the gene expressions are often subject to measurement error, which varies with different labs or platforms used. In practice, it is necessary to combine these various microarray data in the analysis. Thus, prediction with heteroscedastic measurement error has to be addressed.

We propose two variations of the SIMEX methods to adjust the effect of the measurement error and obtain the estimates of the coefficients, β_0, $\widehat{\boldsymbol{\beta}}_X$ and $\widehat{\boldsymbol{\beta}}_Z$. Then, we use the surrogate $(\mathbf{W}_i, \mathbf{W}_{F_i})$ and error-free variable $(\mathbf{Z}_i, \mathbf{Z}_{F_i})$ together to predict the corresponding unobserved $\mathbf{X}_F$. Using the coefficient estimates computed from the training data, we obtain the prediction of the future survival time by replacing $\mathbf{X}_F$ with $\widehat{\mathbf{X}}_F$.

The performance of the proposed SIMEX adjusted prediction and naive prediction methods are evaluated by the mean squared prediction error (MSPE). Due to the fact that some of the survival times might be right censored, we propose the following three methods to adjust the censored survival times for calculating MSPE: inverse probability of weights method, integral method and Buckley James method. No matter which method is applied to adjust the censored survival time, the SIMEX adjusted prediction model outperforms the naive prediction model with higher prediction accuracy. The MSPE of SIMEX adjusted prediction model is smaller than the naive prediction model and is less variable.

We run simulation studies to evaluate the proposed methods. When measurement error variance is known, we first apply the general SIMEX method to obtain the

coefficient estimates. When the variance is not known but replicated measurements of the surrogate are available, we use the empirical SIMEX method to estimate the coefficients. For both the SIMEX adjusted prediction and the naive prediction methods, as the censoring rate increases, the mean and corresponding standard error of the MSPE become larger. This is due to the fact that more information about the true survival times is unavailable.

Chapter 4

simexaft: R Package for Accelerated Failure Time Models with Covariates Subject to Measurement Error

4.1 Introduction

For survival data with covariates subject to measurement error, standard inferential procedures may produce biased estimation if measurement error is not properly taken into account (Carroll et al., 2006). There has been extensive discussion in the literature to correct the bias induced by measurement error in the Cox proportional hazards (PH) model (Prentice, 1982; Li and Lin, 2003; Yi and Lawless, 2007). Although the impact of measurement error is well understood for the Cox PH models, there is little discussion on its impact on accelerated failure time (AFT) models. The AFT model is an attractive alternative to the Cox PH model since it may provide more accurate or concise summarization of the data than the Cox PH model in certain applications (Zeng and Lin, 2007).

With general AFT models, He et al. (2007) discussed inference procedures to account for effects of covariate measurement error using a simulation extrapolation (SIMEX) approach. The main advantage of the developed SIMEX method for AFT models is its simplicity and flexibility to implement. Moreover, it is robust to the distribution of error prone and error free covariates. This method is quite appealing for practitioners to accommodate covariate measurement error when analyzing survival data with AFT models. Despite great advances in the methodology of addressing covariate measurement error for survival analysis, the methods developed in current literature have not been widely used in practice. The reluctance to adopt these methods may be partly due to the lack of available software to implement

these methods. To address this practical issue, we developed an R (R Development Core Team, 2010) package `simexaft` to make the SIMEX method discussed by He et al. (2007) accessible for general users on the Comprehensive R Archive Network (CRAN) at *http://cran.r-project.org/*. The source code for this R package is attached in Appendix A.

4.2 Notation and Framework

Assume we have two types of covariates, let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ be the $p \times 1$ vector of covariates subject to possible measurement error and $\mathbf{Z}_i$ be the vector of error free covariates. The response variable $Y_i = \log(T_i)$ is characterized by the AFT model (3.1). The parameters for the AFT model (3.1) is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})'$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_x, \boldsymbol{\beta}_z)'$ is the vector of regression parameters ($\boldsymbol{\beta}_z$ may include the intercept coefficient). Interest primarily focuses on estimating parameters $\boldsymbol{\beta}$ in order to study the relationship between the response Y_i and covariates $(\mathbf{X}_i, \mathbf{Z}_i)'$. Using (3.1), we define the likelihood contributed from subject i as

$$L_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = [g(Y_i - \mathbf{X}_i'\boldsymbol{\beta}_x - \mathbf{Z}_i'\boldsymbol{\beta}_z; \boldsymbol{\alpha})]^{\delta_i}[1 - G(Y_i - \mathbf{X}_i'\boldsymbol{\beta}_x - \mathbf{Z}_i'\boldsymbol{\beta}_z; \boldsymbol{\alpha})]^{1-\delta_i},$$

where $g(\cdot)$ is the density function corresponding to the distribution function, $G(\cdot)$, of ϵ_i. Then, the log likelihood is given by

$$l(\boldsymbol{\theta}; Y, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i),$$

where $l_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = \log L_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$. If there is no measurement error present in covariates, then the maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}$, is obtained by solving

$$\frac{\partial l(\boldsymbol{\theta}; Y, \mathbf{X}, \mathbf{Z})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{4.1}$$

This estimator is consistent for $\boldsymbol{\theta}$ and has an asymptotic normal distribution. However, when error is present in covariates, the resulting estimator can be substantially biased (Li and Lin, 2003; Yi and He, 2006).

Let $\mathbf{W}_i$ be the observed version of covariate $\mathbf{X}_i$. Conditional on $\mathbf{X}_i$ and $\mathbf{Z}_i$, we assume $\mathbf{W}_i$ and $\mathbf{X}_i$ follow a classical additive measurement error model given by equation(1.2), where measurement error $\mathbf{U}_i$ follows a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_u = [\sigma_{jk}]_{p\times p}$. The parameters in $\boldsymbol{\Sigma}_u$ can be estimated in certain situations (e.g., repeated measurements for $\mathbf{W}_i$ are available). In other situations, the parameters in $\boldsymbol{\Sigma}_u$ may be assumed known because of prior knowledge about the measurement process or other similar studies. When conducting sensitivity analysis to assess the impact of different degree of measurement error on estimation of the response parameters, the parameters in $\boldsymbol{\Sigma}_u$ are typically specified to be known because of the background information about the measurement process.

4.3 Simulation Extrapolation Method

To conduct valid inference for $\boldsymbol{\theta}$ in the presence of covariate measurement error, He et al. (2007) developed a SIMEX method for the AFT model. See section 2.2.3 for a brief background of the SIMEX algorithm. This technique was initially proposed by Cook and Stefanski (1994) for measurement error correction. Later, it was generalized to account for heteroscedastic error model by Devanarayan and Stefanski (2002). Recently, it has been adapted for semiparametric measurement error model (Apanasovich, 2009). The SIMEX method is widely used in applications that involve survival analysis (Li and Lin, 2003; He et al., 2007); misclassification in regression (Kuchenhoff et al., 2006); smoothing parameter choice (Delaigle and Hall, 2008) and estimation of the variance function (Carroll and Wang, 2008; Wang et al., 2009).

The main idea of the SIMEX method is to generate additional data sets with increasingly larger measurement error, estimate the trend of the effect of the measure-

ment error on the estimation of the parameter of interest, and extrapolate the trend back to the case of no measurement error. This method is robust to the distribution of $\mathbf{X}_i$, even when it is unspecified. We consider two practical cases for the parameters in $\mathbf{\Sigma}_u$: (i) the parameters in $\mathbf{\Sigma}_u$ are given as fixed values; and (ii) the parameters in $\mathbf{\Sigma}_u$ are not known, but repeated measurements of $\mathbf{W}_i$ are available. The procedures of the SIMEX method apply to both cases except for the data simulation procedure.

The SIMEX method was generalized to handle survival data for which censoring is a typical feature by He et al. (2007). The SIMEX approach is very appealing because of its simplicity to implement and no requirement of modeling the true covariates $\mathbf{X}_i$ (often not observable). To implement this method, we need to address a few issues. The specification of B or $\mathbf{\Lambda}$ is not unique. Technically speaking, a larger value of B leads to a better SIMEX estimator in the sense that Monte Carlo sampling error in the simulation step can be reduced. For practical use, however, choosing $B = 50, 200$ or 500, and taking $\mathbf{\Lambda}$ to be the equal cut points of interval $[0, 1]$ or $[0, 2]$ with $M = 5, 10$ or 20, can often lead to fairly reasonable SIMEX estimates (Carroll et al., 2006). Another source of variation in obtaining SIMEX estimators lies in the choice of an extrapolation function. The exact extrapolation function is usually not known. Instead, a user-specified approximation is employed, hence SIMEX estimators are usually approximately consistent. Linear regression or quadratic regression function tends to be the most widely used replacement of the exact extrapolation function. Although SIMEX estimators are often not exactly consistent, they greatly outperform naive estimators for which measurement error is not accounted for. The performance of the SIMEX method has been shown superior in some highly nonlinear models (Carroll et al., 1996; Wang et al., 1998).

4.3.1 Implementation in R

An R function, entitled `simexaft`, is developed to implement the SIMEX procedures described above. Function `simexaft` produces the SIMEX estimates for interesting

parameter $\boldsymbol{\beta}$ and other parameters along with their associated SIMEX standard errors and p-values. The form of calling function `simexaft` is given by

```
simexaft(formula = formula(data), data = parent.frame(),
         SIMEXvariable = indicator, repeated = FALSE,
         repind = list(), err.mat = Sigma,  B = B,
         lambda = lambda, extrapolation = quadratic, dist = "Weibull")
```

with the arguments being described as follows

- `formula`: specifies the model to be fitted, with the variables coming with `data`. This argument has the same format as the `formula` argument in the existing R function `survreg`.
- `SIMEXvariable`: the index of the covariate variables that are subject to measurement error.
- `repeated`: set to `TRUE` or `FALSE` to indicate if there are repeated measurements for the mis-measured variables, i.e., corresponding to case (i) or (ii) in Section 4.3.
- `repind`: the index of the repeated measurement variables for each mis-measured variable. It has an R `list` form. If `repeated = TRUE`, `repind` must be specified.
- `err.mat`: specifies the covariance matrix of the measurement error. If `repeated = FALSE`, `err.mat` must be specified.
- `B`: the number of simulated samples for the simulation step. The default is set to be 50.
- `lambda`: the set of $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_M\}$ with $\lambda_1 = 0$ that is used as the grids for the extrapolation step.
- `extrapolation`: specifies the function form for the extrapolation step. The options are `linear`, `quadratic` and `both`. The default is set to be `quadratic`.

- `dist`: specifies a parametric distribution that is assumed in AFT model (3.1). This argument is the same as the `dist` option in the existing R function `survreg`, and it can take distribution such as `Weibull`, `exponential`, `Gaussian`, `logistic`, `lognormal` and `loglogistic`.

4.4 Examples

To illustrate the usage of the developed R package `simexaft`, in this section we apply the package to two real data sets, corresponding to cases with or without repeated measurements for error prone covariates.

The first example is based on a subset from real data set arising from the Busselton Health Study (Knuiman et al., 1994). The whole data set was analyzed in He et al. (2007). The data set analyzed here includes survival information for a randomly selected subset of 100 females. The survival time is taken as the age at the death, as in He et al. (2007). Systolic blood pressure (x_{i1}), cholesterol level (x_{i2}), age at registration (z_{i1}), body mass index (z_{i2}) and smoking status are risk factors related to mortality. Following Carroll et al. (2006), we rescale systolic blood pressure as $\log(x_{i1} - 50)$. Smoking status is classified by two dummy indicators, denoted by z_{i3} and z_{i4}, where $z_{i3} = 1$ indicates an individual is an ex-smoker, and 0 otherwise; $z_{i4} = 1$ represents that an individual is a current smoker, and 0 otherwise. It is known that measurements of risk factors x_{i1} and x_{i2} are subject to substantial error due to the nature of these covariates.

The logarithms of the failure times are postulated by model

$$Y_i = \beta_0 + x_{i1}\beta_{x_1} + x_{i2}\beta_{x_2} + z_{i1}\beta_{z_1} + z_{i2}\beta_{z_2} + z_{i3}\beta_{z_3} + z_{i4}\beta_{z_4} + \epsilon_i,$$

where error ϵ_i follows a specific distribution. The standard extreme value distribution is assumed for an illustration. We assume that errors in both risk factors x_{i1} and x_{i2}

can be represented by model (1.2).

To use the developed R package `simexaft`, we need to install the package from a zip file `simexaft.zip` and load it to R

```
> library(simexaft)
```

Next, load the data that are properly organized with the variable names specified. In this example, the data set "BHS" included in the package is called by

```
> data(BHS)
> dataset = BHS
> dataset$SBP = log(dataset$SBP-50)
```

For illustrative purposes, we use settings with $B = 50$, $\lambda_M = 2$ and $M = 20$. Assume the parameters in $\boldsymbol{\Sigma_u}$ are known. This is a typical case when conducting sensitivity analysis. Here we set $\sigma_{11} = \sigma_{22} = 0.75$ and $\sigma_{12} = \sigma_{21} = 0$ as an example.

The naive AFT approach without considering measurement errors in covariates gives the output,

```
> formula = Surv(SURVTIME, DTHCENS) ~ SBP + CHOL + AGE + BMI
            + SMOKE1 + SMOKE2
> out1 = survreg(formula=formula, data=dataset, dist= "weibull")
> summary(out1)
Call:
survreg(formula = formula, data = dataset, dist = "weibull")
              Value Std. Error      z        p
(Intercept) 12.5302     3.3587  3.731 0.000191
SBP         -1.2524     0.7766 -1.613 0.106807
CHOL        -0.0512     0.1096 -0.467 0.640360
AGE         -0.0603     0.0223 -2.712 0.006692
BMI          0.0337     0.0400  0.842 0.399920
SMOKE1      -0.7392     0.3993 -1.851 0.064158
SMOKE2      -0.8232     0.4178 -1.970 0.048805
Log(scale)  -0.5142     0.2079 -2.474 0.013375
Scale= 0.598
Weibull distribution
Loglik(model)= -83.5   Loglik(intercept only)= -98.5
        Chisq= 30.02 on 6 degrees of freedom, p= 3.9e-05
Number of Newton-Raphson Iterations: 9
n= 100
```

To adjust for possible effects of measurement error in variables SBP and CHOL, we call the developed function `simexaft` for the analysis.

```
> set.seed(120)
> formula = Surv(SURVTIME,DTHCENS) ~ SBP+CHOL+AGE+BMI+SMOKE1+SMOKE2
> ind = c("SBP","CHOL")
> err.mat = diag(rep(0.5625,2))
> ###fit a AFT model with quadratic extrapolation
> out2 = simexaft(formula=formula,data=dataset,SIMEXvariable=ind,
        repeated="FALSE",repind=list(),err.mat=err.mat,B=50,
        lambda=seq(0,2,0.1),extrapolation="quadratic",dist="weibull")
> summary(out2)
$coefficients
             Estimate Std. Error      P value
Intercept 16.33008771 3.91664272 3.053897e-05
SBP       -2.40116761 0.93348413 1.010358e-02
CHOL      -0.05630569 0.12982884 6.645124e-01
AGE       -0.04846142 0.02063056 1.882334e-02
BMI        0.05933523 0.04278722 1.655177e-01
SMOKE1    -0.60168913 0.36963556 1.035694e-01
SMOKE2    -0.79819843 0.39230144 4.188551e-02
$scalereg
(Intercept)
  0.5791607
$extrapolation
[1] "quad"
$SIMEXvariable
[1] "SBP"  "CHOL"
attr(,"class")
[1] "summary.simaxaft"
```

Now we demonstrate the use of `simexaft` for the case that the parameters in $\mathbf{\Sigma}_u$ is unknown, but repeated measurements for error prone covariates are available. This is illustrated by the example from a study of pulmonary exacerbations and rhDNase. Fuchs et al. (1994) reported on a double-blind randomized multicenter clinical trial designed to assess the effect of rhDNase, a recombinant deoxyribonuclease I enzyme, versus placebo on the occurrence of respiratory exacerbations among patients with cystic fibrosis. The rhDNase operates by digesting the extracellular DNA released by leukocytes that accumulate in the lung as a result of bacterial infection, and so it

was expected that aerosol adminstration of rhDNase would reduce the incidence of exacerbations (Cook and Lawless, 2007).

Six hundred and forty five patients were recruited in this trial. Each subject was randomly assigned to treatment or placebo group, and was followed up approximately 169 days for pulmonary exacerbations. Data on the occurrence and resolution of all exacerbations were recorded. The forced expiratory volume (FEV) was considered a risk factor and was measured twice at randomization. The response is defined as the logarithm of the time from randomization to the first pulmonary exacerbation.

To investigate the effect of the FEV on the time to first pulmonary exacerbation, we postulate the model

$$Y_i = \beta_0 + FEV * \beta_1 + trt * \beta_2 + \epsilon_i,$$

where "trt" is the indicator of treatment, and error ϵ_i follows a specific distribution. The standard extreme value distribution is taken again for illustrations. We assume that measurement errors in risk factors FEV can be represented by model (1.2).

First, load the data "rhDNase" into R by issuing

```
>data(rhDNase)
```

Two repeated measurements for covariate FEV, $fev1$ and $fev2$, are called in `simexaft` using the option `repeated=TRUE`, along with a list of index of the repeated measurements.

Existing R function `survreg` can provide the analysis with no measurement error effects properly taken into account, by merely taking the FEV measurements as the average of the two repeated observations.

```
> fev.ave = (rhDNase$fev + rhDNase$fev2)/2
> output1 = survreg(Surv(rhDNase$time2, rhDNase$status)~rhDNase$trt
                    +fev.ave, dist="weibull")
> summary(output1)
Call:
```

```
survreg(formula = Surv(rhDNase$time2, rhDNase$status) ~ rhDNase$trt +
    fev.ave, dist = "weibull")
              Value Std. Error     z         p
(Intercept)  4.5183    0.15470 29.21 1.61e-187
rhDNase$trt  0.3570    0.12179  2.93  3.38e-03
fev.ave      0.0193    0.00275  7.00  2.50e-12
Log(scale)  -0.0782    0.05959 -1.31  1.89e-01
Scale= 0.925
Weibull distribution
Loglik(model)= -1617.5   Loglik(intercept only)= -1652.9
        Chisq= 70.98 on 2 degrees of freedom, p= 3.3e-16
Number of Newton-Raphson Iterations: 5
n= 641
```

Similar analysis results can be obtained if using the `simexaft` function to accommodate covariate error effects. In this example, we note that variation in the two repeated measurements of FEV is too minor to suggest different results obtained from the methods of ignoring or accounting for covariate measurement error. Here we perturb the two repeated observations by adding additional noise, e.g., 15% of sample standard error, and then apply the developed R function to produce the output. This artificial procedure may not be customary when one focuses on a genuine data analysis. However, it is useful for illustration purposes. Moreover, this approach can provide some insights if conducting sensitivity analyses is of prime interest.

```
> set.seed(120)
> error.sd = 0.15*sd(rhDNase$fev)
> error.sd2 = 0.15*sd(rhDNase$fev2)
> fev.error = rhDNase$fev+rnorm(length(rhDNase$fev),mean=0,sd=error.sd)
> fev.error2 = rhDNase$fev2+rnorm(length(rhDNase$fev2),mean=0,sd=error.sd2)
> dataset2 = cbind(rhDNase$time2, rhDNase$status, rhDNase$trt,
                   fev.error, fev.error2)
> colnames(dataset2) = c("time2","status","trt","fev.error","fev.error2")
> dataset2 = as.data.frame(dataset2)
> formula = Surv(time2, status)~trt + fev.error
> ind = c("fev.error")
```

Below is the output obtained from the naive approach that ignores covariate measurement error for perturbed data.

```
> ######naive model using the average FEV value###########
> fev.error.c = (fev.error + fev.error2)/2
> output2 = survreg(Surv(rhDNase$time2, rhDNase$status) ~ rhDNase$trt
                + fev.error.c, dist="weibull")
> summary(output2)
Call:
survreg(formula = Surv(rhDNase$time2, rhDNase$status) ~ rhDNase$trt +
    fev.error.c, dist = "weibull")
              Value Std. Error     z         p
(Intercept)  4.5303    0.15413 29.39 6.66e-190
rhDNase$trt  0.3555    0.12191  2.92  3.54e-03
fev.error.c  0.0190    0.00273  6.98  3.05e-12
Log(scale)  -0.0772    0.05962 -1.30  1.95e-01
Scale= 0.926
Weibull distribution
Loglik(model)= -1617.9   Loglik(intercept only)= -1652.9
        Chisq= 70.02 on 2 degrees of freedom, p= 6.7e-16
Number of Newton-Raphson Iterations: 5
n= 641
```

Now we apply the developed function `simexaft` to adjust for the measurement error effects, with the perturbed data analyzed using the repeated measurements option.

```
> formula = Surv(rhDNase$time2, rhDNase$status)~rhDNase$trt + fev.error
> output3 = simexaft(formula=formula,data=dataset2,SIMEXvariable=ind,
            repeated="TRUE",repind=list(c("fev.error","fev.error2")),
            err.mat=NULL,B=50, lambda=seq(0,2,0.1),
            extrapolation="quadratic", dist="weibull")
> summary(output3)
$coefficients
            Estimate Std. Error      P value
Intercept 4.50991887 0.15790876 0.000000e+00
trt       0.36252461 0.12196482 2.955100e-03
fev.error 0.01935275 0.00279358 4.281020e-12
$scalereg
(Intercept)
   0.925138
$extrapolation
[1] "quad"
$SIMEXvariable
[1] "fev.error"
attr(,"class")
[1] "summary.simaxaft"
```

Compared to the previous results, it is clearly seen that when covariate measurement error is not minor, ignoring it can lead to biased results. Properly accounting for error effects is necessary, and this can be easily accomplished by applying the developed R function `simexaft`.

The estimated covariate coefficients for simulation steps are stored in the results, and the extrapolation curve can be shown through R function `plotsimex`. The `plotsimex` function plots the extrapolation of the estimate of each covariate effect with the option of `linear`, `quadratic` or `both` to view the performance of different extrapolant methods. Here we plot the variable "SBP" in the first example with both linear and quadratic extrapolants.

```
> plotsimexaft(test,"SBP","both",ylim=c(-3,1))
```

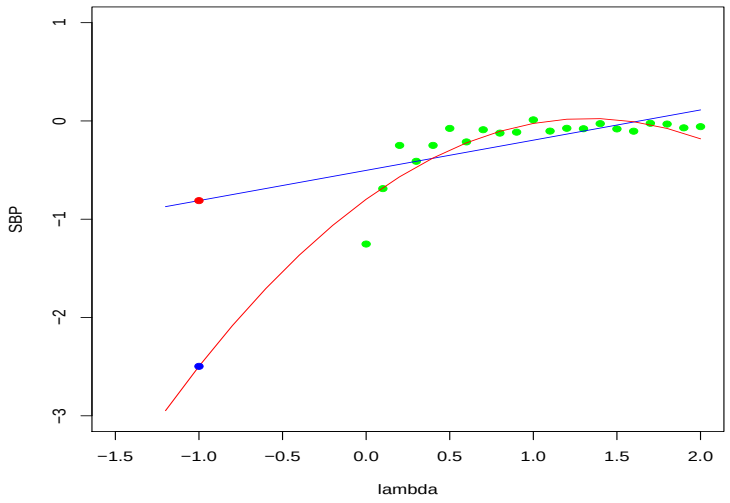

Figure 4.1: Extrapolation of the coefficient

4.5 Discussion

The impact of measurement error in covariates is well documented for survival data that are typically postulated by the Cox PH models, but there is relatively little

discussion on the AFT models. The AFT model is a useful tool for analyzing survival data and "in many ways more appealing than Cox PH models because of its quite direct physical interpretation", as noted by D. R. Cox (Reid 1994). Yi and He (2006) explored the measurement error problem for bivariate survival data under AFT models, but their discussion was mainly focused on the AFT models with normal error distributions. To address the measurement error effects on inferential procedures under AFT models with general distributional forms, He et al. (2007) describe a simulation based method. This method is appealing because it is easy to implement and does not require the specification of the distribution of the error prone true covariates that is generally unobservable. For practical interest, we developed an R package `simexaft` to adjust for biases induced by covariate measurement error under AFT models. Our demonstrations showed that this R package is simple to use. It is anticipated that such development is of great interest to data analysts when handling survival data with covariate measurement error.

Chapter 5

Conclusion and Future Work

Microarray technology is a tool for simultaneously measuring thousands of gene expression that can be used as predictors for survival outcomes. However, microarray data are often subject to measurement error. In current literature, this error is commonly ignored in the analysis of microarray data, which may cause problems in analysis of microarray data. In this thesis, we focus on using the accelerated failure time (AFT) model to investigate the survival analysis of microarray data with measurement error in gene expression being accounted for.

A typical microarray data set has a large number of genes far exceeding the sample size. Proper selection of survival relevant genes contributes to an accurate prediction model. The impact of measurement error in covariates has been extensively studied in the literature for survival data; however, no investigation has been done on the impact of measurement error in survival relevant gene selection in microarray data analysis. We study the effect of the measurement error on survival relevant gene selection under the AFT model setting by regularizing weighted least square estimator with adaptive LASSO penalty. Simulation studies and real data analysis demonstrate that ignoring measurement error will affect survival relevant gene selection. Simulation extrapolation (SIMEX) method is employed to adjust the impact of measurement error to gene selection. With a certain amount of adjustment for the bias induced by the error in covariates, the model selected by the SIMEX method after adjustment is more accurate than the model selected by naively ignoring measurement error. For the naive method, as the measurement error becomes substantial, the biases of the estimates increase while the estimates of the standard

deviations decrease; as a result, the corresponding p-values tend to be smaller than they should be, leading to incorrect hypothesis test results.

Most existing variable selection procedures are limited to directly observed predictors. Variable selection for measurement error data has not been systematically studied yet. In future research, we plan to investigate variable selection for general parametric and semiparametric measurement error models for survival data.

Prediction is an ultimate goal for many statistical analyses. When there are no error prone covariate or homoscedastic error prone covariate in the AFT model, prediction can be done as a no measurement error case. So, there is no need to adjust the effect of measurement error. However, when the covariates subject to heteroscedastic measurement error, the naive prediction model without adjusting the effect of measurement error may not be appropriate. In Chapter 3, we consider a prediction AFT model using data with heteroscedastic covariate measurement error. Two variations of the SIMEX algorithm are investigated to adjust the effect of the measurement error, and a best linear prediction is employed to predict the corresponding value of the unobserved error prone covariates of the future observation.

The performance of the proposed SIMEX adjusted prediction method and naive prediction methods are evaluated by the mean squared prediction error (MSPE). Due to that some of the survival time might be subject to censoring, we propose to use three methods to adjust the censored survival times to calculate the MSPE. Simulation studies show that the SIMEX method can achieve better prediction accuracy than the naive method since the MSPE and variability of the SIMEX adjusted prediction model are smaller than those of the naive prediction model.

In this thesis, the SIMEX method is used to adjust for the effects of gene expression measurement error in both survival relevant gene selection and prediction model for survival. The major advantage of this method is its easy implementation and robustness to distributional assumptions for error prone covariates. The general idea of the SIMEX method is to generate additional data sets with increasingly larger

measurement error, estimate the trend of the effect of the measurement error on the estimation of the parameter of interest with respect to the magnitude of enlarged variation, and extrapolate the trend back to the case of no measurement error. For practical interest, we developed an R package `simexaft` to adjust for biases induced by covariate measurement error under AFT models. Our illustrations show that the developed package is simple to use. It is anticipated that such a development is of great interest to data analysts when handling survival data with covariate measurement error. The R package code is included in appendix, and the package is available on the Comprehensive R Archive Network website for potential users.

There are currently several methods available to correct for measurement error in the AFT model. In the future, we plan to compare the performance of the SIMEX method to other methods, such as regression calibration proposed by Yu and Nan (2009) and nonparametric method proposed by Wang (2000) to correct the effect of measurement error.

In this thesis, we investigated the AFT model with error prone covariates. It is of interest to study the effect of measurement error on other survival regression models. The Proportional Odds (PO) regression model (Cox, 1972; Bennett 1983; Yang and Prentice, 1999) relates the covariate effect on the baseline odds function. The PO model with error prone covariates is

$$\frac{S(t|\mathbf{X},\mathbf{Z})}{1-S(t|\mathbf{X},\mathbf{Z})} = \frac{S_0(t)}{1-S_0(t)} \exp\left(\mathbf{W}'\boldsymbol{\beta}_w + \mathbf{Z}'\boldsymbol{\beta}_z\right),$$

where $S(t|\mathbf{X},\mathbf{Z})$ denote the conditional survival function given covariate $\mathbf{X}$ and $\mathbf{Z}$; $S_0(t)$ denote the unspecified baseline survival function; $\mathbf{W}$ is contaminated version of covariate $\mathbf{X}$; and $\mathbf{Z}$ are the error free covariates. Extension of the SIMEX method to this model should be straightforward.

Another interesting topic is to extend the SIMEX method to a more general model which includes Cox proportional hazards (PH) model and PO model as special

cases (Yang and Prentice, 2005). Suppose we have two treatment A and B, the hazard ratio of these two treatment can be modeled as

$$\frac{\lambda_A(t)}{\lambda_B(t)} = \frac{\exp(\mathbf{X}'\boldsymbol{\beta})\exp(\mathbf{V}'\boldsymbol{\gamma})}{\exp(\mathbf{X}'\boldsymbol{\beta}) + (\exp(\mathbf{V}'\boldsymbol{\gamma}) - \exp(\mathbf{X}'\boldsymbol{\beta}))\, S_B(t)},$$

where $\lambda_A(t)$ is the hazard function of subjects in treatment A with covariate $\mathbf{X}$; $\boldsymbol{\beta}$ is the regression coefficient; $\lambda_B(t)$ is the hazard function of subjects in treatment B with covariate $\mathbf{V}$; $\boldsymbol{\gamma}$ is the regression coefficient; and $S_B(t)$ is the survival function of subjects in treatment group B. If $\boldsymbol{\gamma} = 0$, the model reduce to PO model and when $\exp(\mathbf{X}'\boldsymbol{\beta}) = \exp(\mathbf{V}'\boldsymbol{\gamma})$, the model reduce to Cox PH model.

BIBLIOGRAPHY

[1] Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, **60**, 255-265.

[2] Apanasovich, T. V., Carroll, R. J. and Maity, A. (2009). SIMEX and standard error estimation in semiparametric measurement error models. *Electronic Journal of Statistics*, **3**, 318-348.

[3] Augustin, T., Döring, A. and Rummel, D. (2008). Regression calibration for Cox regression under heteroscedastic measurement error determining risk factors of cardiovascular diseases from error-prone nutritional replication data. In *Recent Advances in Linear Models and Related Areas*, Shalabh, C. H. (Ed.) Springer: Physica-Verlag.

[4] Bennett, S. (1983). Analysi of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273-277.

[5] Breiman L. (1995). Better subset regression using the nonnegative garrote. *American Society for Quality Control and American Statistical Association* **37**, 373-384.

[6] Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, **21**, 33-37.

[7] Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429-436.

[8] Buhlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of American Statistical Association* **98**, 324-407.

[9] Buzas, J. F. (1998). Unbiased scores in proportional hazards regression with covariate measurement eror. *Journal of Statistical Planning and Inference*, **67**, 237-257.

[10] Buzas, J. S., Stefanski, L. A. and Tosteson, T. D. (2005). Measurement error. In W. Ahrens & I.Piegot (Eds.), *Handbook of Epidemiology*, London: Springer.

[11] Cai, T., Huang, J. and Tian, L. (2008). Regularized estimation for the accelerated failure time model. *Biometrics*, **65**, 394-404.

[12] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, **35**, 2313-2351.

[13] Carroll, R. J., Delaigle, A. and Hall, P. (2009). Nonparametric prediction in measurement error models. *Journal of the American Statistical Association*, **104**, 993-1014.

[14] Carroll, R. J., Küchenhoff, H., Lombard, F. and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, **91**, 242-250.

[15] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu. C. M. (2006). *Measurement Error in Nonlinear Models* (2nd edn). London: Chapman & Hall.

[16] Carroll, R. J., and Stefanski, L. A. (1990) Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, **85**, 652-663.

[17] Carroll, R. J. and Wang, Y. (2008). Nonparametric variance estimation in the analysis of microarray data: a measurement error approach. *Biometrika*, **95**, 437-449.

[18] Cheng, C. L. and Riu, J. (2006). On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics*, **48**, 511-519.

[19] Clayton, D. G. (1992). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: Dwyer, J. H., Feinlieb, M., Lippert, P. and Hoffmeister, H. (Ed.) *Statistical Models for Longitudinal Studies on Health*. New York: Oxford University Press.

[20] Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314-1328.

[21] Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer.

[22] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Association*, *Series B*, **34**, 187-202.

[23] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

[24] Datta, S., Le-Rademacher, J. and Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics*, **63**, 259-271.

[25] Debouck, C. and Goodfellow, P. N. (1999). DNA microarray in drug discovery and development. *Nature Genetics Supplement*, **21**, 48-50.

[26] Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*. **103**, 280-287.

[27] Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, **102**, 1416-1426.

[28] Delaigle, A. and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, **14**, 562-579.

[29] DellaPenna, D. (1990). Nutritional genomics: manipulating plant micronutrients to improve human health. *Science*, **285**, 375-379.

[30] DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.

[31] Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters*. **59**, 219-225.

[32] Dudoit, S., Fridlyand, J., Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors ssing gene expression data. *Journal of the American Statistical Association*. **97**, 77-87.

[33] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics Supplement*, **21**, 10-14.

[34] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. T. (2004). Least angle regression. *The Annals of Statistics*, **32**, 407-499.

[35] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

[36] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (Editor's invited paper). *Statistica Sinica*, **20**, 101-148.

[37] Faraggi, D. and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475-1485.

[38] Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, **22**, 1947-1975.

[39] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tolls (with discussion). *Technometrics* **35**, 109-148.

[40] Friedman, J. and Popescu, B. (2004). Gradient directed regularization. Technical Report, Standford University, Stanford, California.

[41] Fuchs, H. J., Borowitz, D. S., Christiansen, D. H., Morris, E. M., Nash, M. L., Ramsey, B. W., Rosenstein, B. J., Smith, A. L., and Wohl, M. E. (1994). Effect of aerosolized recombinant huamn DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. *New England Journal of Medicine*, **331**, 637-642.

[42] Fuller, W. A. (1987). *Measurement Error Models*, John Wiley & Sons.

[43] Giménez, P., Bolfarine, H. and Colosimo, E. A. (1999). Estimation in Weibull regression model with measurement error. *Communications in Statistics - Theory and Method*, **28**, 495-510.

[44] Giménez, P., Bolfarine, H. and Colosimo, E. A. (2006). Asymptotic relative efficiency of score tests in Weibull models with measurement errors. *Statistical Papers*, **47**, 461-470.

[45] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

[46] Green, W. F. and Cai, J. (2004). Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, **60**, 987-996.

[47] Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size setting, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001-3008.

[48] He, W. and Yi, G. Y. (2009). Survival prediction with gene expression profiles. *JP Journal of Biostatistics*, **3**, 17-39.

[49] He, W., Yi, G. Y. and Xiong, J. (2007). Accelerated failure time models with covariates subject to measurement error. *Statistics in Medicine*, **26**, 4817-4832.

[50] Hu, C. and Lin, D. Y. (2004). Semiparametric failure time regression with replicates of mismeasured covariates. *Journal of the American Statistical Association*, **99**, 105-118.

[51] Hu, P., Tsiatis, A. A. and Davidian, M. (1998). Estimating the parameters of the Cox model when covariate variables are measured with errors. *Biometrics*, **54**, 1407-1419.

[52] Huang, J. and Harrington, D. (2005). Iterative partial least squares with right-censored data analysis: a comparison to other dimension reduction techniques. *Biometrics*, **61**, 17-24.

[53] Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics*, **62**, 813-820.

[54] Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association*, **95**, 1209-1219.

[55] Jin, Z. and He, W. (2010). Local linear regression on clusgtered censored data. Unpublished manuscript.

[56] Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wilev & Sons.

[57] Knuiman, M. W., Cullent, K. J., Bulsara, M. K., Welborn, T. A. and Hobbs, M. S. T. (1994). Mortality trends, 1965 to 1989, in Busselton, the site of repeated health surveys and interventions. *Australian Journal of Public Health*, **18**, 129-135.

[58] Kuchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, **62**, 85-96.

[59] Kukush, A. and Schneeweiss, H. (2004). Relative efficiency of maximum likelihood and other estimators in a nonlinear regression model with small measurement errors. Collaborative Research Center 386, Discussion Paper 396.

[60] Kulathinal, S. B., Kuulasmaa, K. and Gasbarra, D. (2002). Estimation of an error-in-variables regression model when the variances of the measurement errors vary between the observations. *Statistics in Medicine*, **21**, 1089-1101.

[61] Lawless J. F. (2003). *Statistical Models and Methods for Lifetime Data*. (2nd Ed.) Wiley: New York.

[62] Li, H. (2008) Censored data regression in high dimensional and low sample size settings for genomic applications. In *Statistical advances in biomedical science: survival analysis and bioinformatics*, Biswas, A., Datta, S., Fine, J. and Segal, M. (1st Ed.) Hoboken, New Jersey: Wiley-Interscience.

[63] Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, **8**, 65-76.

[64] Li, Y. and Lin, X. (2000). Covariate measurement errors in frailty models for clustered survival data. *Biometrika*, **87**, 849-866.

[65] Li, Y. and Lin, X. (2003). Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of the American Statistical Association*, **98**, 191-203.

[66] Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement Errors. *Journal of American Statistical Association*, **104**, 234-248.

[67] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T. , Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H. and Brown, L. E. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, **14**, 1675-1680.

[68] Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli*, **16**, 274-300.

[69] Miller, R. G. (1976). Least squares regression with censored data. *Biometrika*, **63**, 449-464.

[70] Nakamura, T. (1992). Proportional hazards models with covariates subject to measurement error. *Biometrics*, **48**, 829-838.

[71] Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C. and Afshari, C. A. (1999). Microarray and toxicology: The advent of toxicogenomics. *Molecular Carcinogenesis*, **24**, 153-159.

[72] Page, G.P., Edwards, J.W., Barnes. S., Weindruch, R. and Allison, D.B. (2003). A design and statistical perspective on microarray gene expression studies in nutrition: The need for playful creativity and scientific hard-mindedness. *Nutrition*, **19**, 997-1000.

[73] Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331-342.

[74] Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, **2**, 418-427.

[75] Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, **32**, 496-501.

[76] R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[77] Reid, N. (1994). A Converstaion with Sir David Cox. *Statistical Science*, **9**, 439-455.

[78] Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*. **346**, 1937-1947.

[79] Russell, S., Meadows, L. A. and Russell, R. R. (2009). Microarray technology in practice. (Eds.) Boston : Academic Press/Elsevier.

[80] Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

[81] Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. *BMC Boinformatics*, **9**(269).

[82] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

[83] Staudenmayer, J., Ruppert, D. and Buonaccosi, J. (2008). Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*. **103**, 726-736.

[84] Stefanski, L. A. and Carroll, R. J. (1990). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Association*, *Series B*,**52**, 345-359.

[85] Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*, **23**, 461-471.

[86] Tibshirani, R. T. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Association*, *Series B* **58**, 267-288.

[87] Tibshirani, R. T. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

[88] Tseng, Y. K., Hsieh, F. and Wang, J. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, **92**, 587-603.

[89] Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, **18**, 354-372.

[90] Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Van Der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002). A gene expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, **25**, 1999-2009.

[91] Wang, N., Lin, X., Gutierrez, R. G. and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, **93**, 249-261.

[92] Wang, Q. (2000). Estimation of linear error-in-variable models with validation data under random censorship. *Journal of Multivariate Analysis*, **74**, 245-266.

[93] Wang, Y., Ma, Y. and Carroll, R. J. (2009). Variance estimation in the analysis of microarray data. *Journal of Royal Statistical Society*, *Series B*, **71**, 425-445.

[94] Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, **94**, 125-136.

[95] Yang, S. and Prentice, R. L. (2005). Semiparametric analysis of short term and long term relative risks with two sample survival data, *Biometrika*, **92**, 1-17.

[96] Yi, G. Y. (2010). Unpublished manuscript.

[97] Yi, G. Y. and He, W. (2006). Methods for bivariate survival data with mismeasured covariates under an accelerated failure time model. *Communications in Statistics - Theory and Methods*, **35**, 1539-1554.

[98] Yi, G. Y. and Lawless, J. F. (2007). A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *Journal of Statistical Planning and Inference*, **137**, 1816-1828.

[99] Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics*, **21**, 76-99.

[100] Yu, M. and Nan, B. (2009). Regression calibration in semiparametric accelerated failure time models. *Biometrics*, **66**, 405-414.

[101] Zeng, D. and Lin, D. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, **102**, 1387-1396.

[102] Zhou, H. and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, **82**, 139-149.

[103] Zhou, H. and Wang, C. Y. (2000). Failure time regression with continuous covariates measured with error. *Journal of Royal Statistical Society, Series B*, **62**, 657-665.

[104] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

[105] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of Royal Statistical Society, Series B*, **67**, 301-320.

Appendix A

Appendices

A.1 R code for SIMEXAFT Package

R Code of *simexaft*

```
.packageName <- "simexaft"
‘linearextrapolation‘ <-function(A1,A2,A3,lambda){
     reg1<-numeric()
     reg2<-numeric()
     scalereg<-numeric()
     D=ncol(A1)

     for(i in 1:D)
     {
            e1=coef(lm(A1[,i]~lambda))
            a1= e1[1] - e1[2]
            reg1 = c(reg1,a1)

            e2 = coef(lm(A2[,i]~lambda))
            a2 = e2[1] - e2[2]
            reg2 = c(reg2, a2)
     }

     e3 = coef(lm(A3[,1]~lambda))
     a3 =  e3[1] - e3[2]
     scalereg= c(scalereg,a3)
     return(list("reg1"=reg1,"reg2"=reg2,"scalereg"=scalereg))
}

‘predic.simexaft‘ <-function(object,newdata,...)
{
     new.object<-object$formula
     new.object$coefficients=object$coefficients
     predict(new.object, newdata=data.frame(newdata),...)
}

‘print.simexaft‘<-function(x, digits=max(3, getOption("digits")- 3), ...)
{
    cat("\nSIMEX-Variables: ")
    cat(x$SIMEXvariable, sep = ", ")
    cat("\nNumber of Simulations: ", paste(x$B), "\n\n", sep = "")
```

```
    if (length(coef(x))) {
            cat("Coefficients:\n")
            print.default(format(coef(x), digits = digits), print.gap = 2,
            quote = FALSE)
    }
    else cat("No coefficients\n")
    cat("\n")
}

`quadraticextrapolation` <-function(A1,A2,A3,lambda){
     reg1<-numeric()
     reg2<-numeric()
     scalereg<-numeric()
     D=ncol(A1)

     for(i in 1:D)
     {
             lambda2=lambda^2
             e1=coef(lm(A1[,i]~lambda + lambda2))
             a1 = e1[1] - e1[2] + e1[3]
             reg1 = c(reg1,a1)

             e2 = coef(lm(A2[,i]~lambda + lambda2))
             a2 = e2[1] - e2[2] + e2[3]
             reg2 = c(reg2, a2)
     }

     e3=coef(lm(A3[,1]~lambda+lambda2))
     a3= e3[1]-e3[2]+e3[3]
     scalereg= c(scalereg,a3)
     return(list("reg1"=reg1,"reg2"=reg2,"scalereg"=scalereg))
}

`simexaft`<-function(formula=formula(data),data=parent.frame(),
                     SIMEXvariable=indicator,repeated="F", repind=list(),
                     err.mat=Sigma,B=100,lambda=seq(0,2,0.1),

    extrapolation="quadratic",dist="weibull")
    {  colname=colnames(data)
       SIMEXvariable=unique(SIMEXvariable)
       nSIMEXvariable=length(SIMEXvariable)

       if(!is.character(SIMEXvariable) | nSIMEXvariable>length(colname)){
               stop("Invalid SIMEXvariable object")
               }
       if(!all(SIMEXvariable %in% colname)){
               stop("SIMEXvariable must selected from the data")
               }
       if (!any(repeated == c("F", "T"))) {
```

```
        stop("Repeated indicator not implemented.")
        }

if(repeated=="F"){
         if(!is.numeric(err.mat) | any(err.mat < 0)){
              stop("Invalid err.mat object")
              }

         if (nrow(err.mat) != ncol(err.mat)) {
              stop("err.mat must be a square matrix")
              }

         if (length(SIMEXvariable) != nrow(err.mat)) {
              stop("SIMEXvariable and err.mat
                     have non-conforming size")
              }

         SSigma <- err.mat
         dimnames(SSigma) <- NULL
         if (!isTRUE(all.equal(SSigma, t(SSigma)))) {
              warning("err.mat is numerically not symmetric")
              }
         }
else if(repeated=="T"){
         if(length(SIMEXvariable) != length(repind)){
             stop("SIMEXvariable and repind
                     have non-conforming size")}
         }

if(length(B)!=1){
         stop("B must be positive integer")
         }
if(!is.numeric(B) | B<=0 ){
         stop("B must be positive integer")
         }
 else{
         B=ceiling(B)
         }
  if(!is.vector(lambda) |!is.numeric(lambda)){
             stop(":Invalid lambda object")
             }

 if (any(lambda < 0)) {
        warning("Lambda should be positive values.
        Negative values will be ignored",call. = FALSE)
        lambda <- lambda[lambda >= 0]
        }

 extrapolation = substr(extrapolation, 1, 4)
```

```
if(!is.character(extrapolation) | length(extrapolation)!=1){
        warning("Invalid extrapolation object.
        Using: quadatic\n\n",call.=FALSE)
        }
extrapolation="quad"

ndata=nrow(data)
nformula=length(attr(terms(formula),"term.labels"))+1
nlambda=length(lambda)

A1=matrix(data=NA,nlambda,nformula)
A2=matrix(data=NA,nlambda,nformula)
A3=matrix(data=NA,nlambda,nformula)
theta=matrix(data=NA,B,nformula)

colnames(theta)=c("Intercept",attr(terms(formula),"term.labels"))
p.names=colnames(theta)

theta.all=vector(mode="list",nlambda)
for(k in 1:length(lambda))
{
    w=numeric()
    v=numeric()
    omega=numeric()
    temp=data
    estivarB=matrix(data=NA,B,nformula)
    estiscaleB=matrix(data=NA,B,ncol=1)

    for(r in 1:B)
    {
        if(repeated=="F"){
        temp[SIMEXvariable]=data[SIMEXvariable]+sqrt(lambda[k])*
        rmvnorm(ndata,rep(0,length(SIMEXvariable)),err.mat)
        }
        else{
        constrast=list()
        for(i in 1:nSIMEXvariable){
              n.i=length(repind[[i]])
              z.i=rnorm(n.i, 0, 1)
              constrast[[i]]=(z.i-mean(z.i))
                           /sqrt(sum((z.i-mean(z.i))^2))
              mean.i=apply(temp[repind[[i]]],1,sum)/n.i
              temp[SIMEXvariable[i]]=mean.i + sqrt(lambda[k]/n.i)*
                                    as.matrix(temp[repind[[i]]])
                                    %*%as.vector(constrast[[i]])
              }
        }

        re = survreg(formula=formula,data=temp,dist=dist)
```

```
          scale=re$scale
          w=re$coefficients
          omega=diag(re$var)[1:nformula]
          theta[r,]=w
          estivarB[r,]=omega
          estiscaleB[r,]=scale
      }

      w=apply(theta,2,FUN=mean)
      v=apply(theta,2,FUN=var)
      omega =apply(estivarB,2,FUN=mean)
      tau=omega-v
      A1[k,] = w
      A2[k,] = tau
      A3[k,] = apply(estiscaleB,2,FUN=mean)
      theta.all[[k]]=theta
 }

theta=matrix(unlist(theta.all),nrow=B)
theta.all=list()

for (i in 1:nformula){
 theta.all[[p.names[i]]]<-
 data.frame(theta[,seq(i, nformula * nlambda, by = nformula)])
}

if(extrapolation=="line"){
     result1=linearextrapolation(A1,A2,A3,lambda)}
 else if(extrapolation=="quad"){
     result1=quadraticextrapolation(A1,A2,A3,lambda)}
 else stop("extrapolation method must be linear or quadratic")

 estimate=result1$reg1
 names(estimate)=p.names
 se=sqrt(result1$reg2)
 names(se)=p.names
 scalereg=result1$scalereg
 pvalue=2*(1-pnorm(abs(estimate/se)))

 if(repeated=="F"){
 erg=list(coefficients=estimate,se=se,scalereg=scalereg,
          pvalue=pvalue, lambda=lambda, B=B, formula=formula,
          err.mat=err.mat,extrapolation=extrapolation,
          SIMEXvariable=SIMEXvariable,theta=theta.all)
 }
 else{
 erg=list(coefficients=estimate,se=se,scalereg=scalereg,
          pvalue=pvalue,lambda=lambda, B=B, formula=formula,
          extrapolation=extrapolation, SIMEXvariable=SIMEXvariable,
```

```
                repind=repind,theta=theta.all)
        }

        class(erg)<-("simexaft")
        return(erg)
}

‘summary.simexaft‘ <-function (object, ...)
{
    p.names <- names(object$coefficients)
    est <- object$coefficients
    est.table <- list()

    se <- object$se
    pvalue <-object$pvalue
    est.table <- cbind(est, se, pvalue)
    dimnames(est.table)<-list(p.names,c("Estimate","Std. Error","P value"))

    ans <- list()
    class(ans) <- "summary.simaxaft"
    ans$coefficients <- est.table
    ans$call <- object$call
    ans$scalereg <- object$scalereg
    ans$extrapolation <- object$extrapolation
    ans$SIMEXvariable <- object$SIMEXvariable
    ans
}
```

A.2 The Impact of Ignoring Measurement Error

In this section, We summarized some of the known results about the effect of the measurement error in linear model. Fuller (1987) gave a comprehensive overview of measurement error modeling and adjusted estimators for linear models. And see Carroll *et al* (2006) for detailed coverage of nonlinear models.

The multiple linear regression model is defined as:

$$Y_i = \beta_0 + \beta^t \mathbf{X_i} + \epsilon_i \qquad i = 1, \ldots, n$$

Let $\mathbf{X_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ be the $p \times 1$ covariates subject to possible measurement error and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the coefficient parameter. And $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Now suppose that we have additive measurement error:

$$\mathbf{W_i} = \mathbf{X_i} + \mathbf{U_i}$$

Where $\mathbf{W_i}$ is unbiased for $\mathbf{X_i}$, and $\mathbf{U_i}$ is independent random variable with $E(\mathbf{U_i}|\mathbf{X_i}) = \mathbf{0}$, $Var(\mathbf{U_i}|\mathbf{X_i}) = \mathbf{\Sigma_{uu}}$. And

$$\begin{pmatrix} \mathbf{X_i} \\ \epsilon_i \\ \mathbf{U_i} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \mu_\mathbf{x} \\ 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma_{uu}} \end{pmatrix} \right)$$

Then

$$\begin{pmatrix} \mathbf{X_i} \\ \mathbf{W_i} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \mu_\mathbf{x} \\ \mu_\mathbf{x} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xx}} \\ \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}} \end{pmatrix} \right)$$

And the conditional distribution of $\mathbf{X}$ given $\mathbf{W}$ is:

$$\mathbf{X}|\mathbf{W} \sim N_p\Big((\mu_\mathbf{x} + \mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}(\mathbf{W} - \mu_\mathbf{x})), \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}\mathbf{\Sigma_{xx}}\Big)$$

Hence, the conditional mean and variance are:

$$E(\mathbf{X}|\mathbf{W}) = \mu_\mathbf{x} + \mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}(\mathbf{W} - \mu_\mathbf{x})$$

$$Var(\mathbf{X}|\mathbf{W}) = \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}\mathbf{\Sigma_{xx}}$$

With nondifferential measurement error,

$$\begin{aligned} E(Y, \mathbf{W}) &= E\{E(Y|\mathbf{X}, \mathbf{W})|\mathbf{W}\} \\ &= E\{E(Y|\mathbf{X})|\mathbf{W}\} \\ &= E(\beta_0 + {\beta_\mathbf{x}}^t\mathbf{X}|\mathbf{W}) \\ &= \beta_0 + {\beta_\mathbf{x}}^t[\mu_\mathbf{x} - \mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}\mu_\mathbf{x}] + {\beta_\mathbf{x}}^t[\mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}]\mathbf{W} \\ &= \beta_{w0} + {\beta_\mathbf{w}}^t\mathbf{W} \end{aligned}$$

The naive least squares regression of Y on **W** without adjusting for the measurement error will get a consistent estimate not of $\beta_{\mathbf{x}}$, but

$$\beta_{\mathbf{w}}{}^t = \beta_{\mathbf{x}}{}^t[\mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}]$$

The residual variance of this regression of Y on **W** is:

$$\begin{aligned} Var(Y|\mathbf{W}) &= \beta_{\mathbf{x}}{}^t Var(\mathbf{X}|\mathbf{W})\beta_{\mathbf{x}} + \sigma^2 \\ &= \beta_{\mathbf{x}}{}^t[\mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{xx}}(\mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{uu}})^{-1}\mathbf{\Sigma_{xx}}]\beta_{\mathbf{x}} + \sigma^2 \end{aligned}$$

A.2.0.1 Simple Linear Regression Model

When p = 1, we have the simple linear regression model, and:

$$E(X|W) = \frac{\sigma_u^2}{\sigma_x^2 + \sigma_u^2}\mu_x + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}W$$

$$Var(X|W) = \frac{\sigma_u^2\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

With nondifferential measurement error,

$$\begin{aligned} E(Y, W) &= \beta_0 + \frac{\beta_x\sigma_u^2}{\sigma_x^2 + \sigma_u^2}\mu_x + \frac{\beta_x\sigma_x^2}{\sigma_x^2 + \sigma_u^2}W \\ &= \beta_0^* + \beta_x^* W \end{aligned}$$

The naive least squares regression of Y on W without adjusting for the measurement error will get a consistent estimate not of β_x, but instead of $\beta_x^* = \lambda\beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1$$

The residual variance of this regression of Y on W is:

$$\begin{aligned} Var(Y|W) &= \sigma_*^2 \\ &= \beta_x^2 Var(X|W) + \sigma^2 \\ &= \lambda\beta_x^2\sigma_u^2 + \sigma^2 \end{aligned}$$

The variance of the slope estimator calculated from the true data (Y, X) would be

$$Var(\hat{\beta}_x) = \sigma^2/S_{xx} = \sigma^2/n\sigma_x^2$$

The variance of the naive slope estimator calculated from the data (Y, W) would be

$$Var(\hat{\beta}_x^*) = \sigma_*^2/S_{ww} = \sigma_*^2/n\sigma_w^2$$

The naive estimate of the slope can be asymptotically less variable than the true data estimator as long as:

$$\begin{aligned} Var(\hat{\beta}_x^*) &< Var(\hat{\beta}_x) \\ \sigma_*^2/n\sigma_w^2 &< \sigma^2/n\sigma_x^2 \\ \lambda\beta_x^2 &< \frac{\sigma^2}{\sigma_x^2} \end{aligned}$$

As pointed out by Buzas, Stefanski and Tosteson (2005) that this inequality is possible when σ^2 is large, or σ_u^2 is large , or β_x^2 is small. Note that this phenomenon cannot occur with Berkson error, for which the variance of the naive estimator is never less than the variance of the true-data estimator asymptotically. (Carroll, et al., 2006)

VITA

Name

Juan Xiong

Education

DOCTOR OF PHILOSOPHY in Statistics, 2010
The University of Western Ontario
London, Ontario, Canada
Supervisor: Wenqing He

MASTER OF SCIENCE in Biostatistics, 2006
The University of Western Ontario
London, Ontario, Canada
Supervisor: Wenqing He

BACHELOR OF SCIENCE in Information and Computer Science, 2005
East China Normal University
Shanghai, China
Supervisor: Xuecheng Pang

Employment

Teaching Assistant, September 2006 - December 2009
Department of Statistical and Actuarial Sciences, The University of Western Ontario

Research Assistant, September 2005 - November 2010
Department of Statistical and Actuarial Sciences, The University of Western Ontario

Research Interests

Survival Analysis, Measurement Error Models, Statistical Genetics.

Publications

Peer-Reviewed Journals/Proceedings

- W. He, J. Xiong and G. Y. Yi (2010). SIMEX R package for accelerated failure time models with covariates measurement error. Conditionally accepted by *Journal of Statistical Software*.

- J. Xiong, W. He (2009). Survival relevant gene selection in microarray data analysis with gene expression subject to measurement error. In *JSM Proceedings, Biostatistics Section*. Washington, DC: American Statistical Association.

- W. He, G. Y. Yi and J. Xiong (2007). Accelerated failure time models with covariates subject to measurement error, *Statistics in Medicine*, **26**, 4817 - 4832.

Submissions/In Progress

- J. Xiong and W. He (2010). Survival relevant gene selection in microarray data analysis with gene expression subject to measurement error. Submitted.

- A. Yuan, G. Chen, W. He and J. Xiong (2009). Bayesian frequentist hybrid model with application to the analysis of gene copy number changes. Submitted.

Conference Presentations

- Juan Xiong and Wenqing He (2010). Predicting of survival time by combining mismeasured gene expression data from different platforms. *The Annual General Meeting of the Statistical Society of Canada*, Quebec City, QC, Canada, May 23 May 26.

- Juan Xiong and Wenqing He (2009). Poster presentation at the *Research and Education Day of the Department of Oncology*, University of Western Ontario, The Best Western Lamplighter Inn, London, ON, Canada, June 12.

- Juan Xiong and Wenqing He (2009). Survival relevant gene selection in microarray data analysis with gene expression subject to measurement error. *The Annual General Meeting of the Statistical Society of Canada*, Vancouver, BC, Canada, May 3 June 3.

- Juan Xiong (2006). Accelerated failure time models with mismeasured covariates. *The Second Annual MSc Day of the Department of Statistics & Actuarial Sciences*, University of Western Ontario, London, ON, Canada, July 27.

Scholarships and Academic Awards

- Western Graduate Research Scholarship, The University of Western Ontario, Canada, (2005-2010)
- $3rd$-class Academic Excellence Scholarship, East China Normal University, China, (2003 academic year)
- Outstanding Student Award, East China Normal University, China, (2002 academic year)
- $1st$-class Academic Excellence Scholarship, East China Normal University, China, (2002 academic year)

Membership to Professional Society

Statistical Society of Canada