JACK N. AVERITT COLLEGE of GRADUATE STUDIES GEORGIA SOUTHERN UNIVERSITY

Georgia Southern University
Georgia Southern Commons

Electronic Theses and Dissertations

Jack N. Averitt College of Graduate Studies

Fall 2017

Application of the Misclassification Simulation Extrapolation (Mc-Simex) Procedure to Log-Logistic Accelerated Failure Time (Aft) Models In Survival Analysis

Varadan Sevilimedu

Follow this and additional works at: https://digitalcommons.georgiasouthern.edu/etd

 Part of the Epidemiology Commons, Health Services Research Commons, and the Other Public Health Commons

Recommended Citation
Sevilimedu, Varadan, "Application of the Misclassification Simulation Extrapolation (Mc-Simex) Procedure to Log-Logistic Accelerated Failure Time (Aft) Models In Survival Analysis" (2017). *Electronic Theses and Dissertations*. 1659.
https://digitalcommons.georgiasouthern.edu/etd/1659

This dissertation (open access) is brought to you for free and open access by the Jack N. Averitt College of Graduate Studies at Georgia Southern Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Georgia Southern Commons. For more information, please contact digitalcommons@georgiasouthern.edu.

APPLICATION OF THE MISCLASSIFICATION SIMULATION EXTRAPOLATION (MC-SIMEX) PROCEDURE TO LOG-LOGISTIC ACCELERATED FAILURE TIME (AFT) MODELS IN SURVIVAL ANALYSIS

by

Varadan Sevilimedu

Under the Direction of Lili Yu

ABSTRACT

Survival analysis is the study of time to event outcomes. Accelerated Failure Time models (AFT) serve as a useful tool in survival analysis to study the time of occurrence of an event and its relation to the covariates of interest. The accuracy of estimation of parameters in AFT models is dependent upon the correct classification of binary covariates. Considering that perfect classification is highly unlikely, it is imperative that the performance of the existing bias-correction methods be analyzed in AFT models. However, certain areas of bias-correction in AFT models still remain unexplored. One of these unexplored areas, is a situation where the survival times follow a log-logistic distribution. In this dissertation, we evaluate the performance of the Misclassification simulation extrapolation (MC-SIMEX) procedure, a well known procedure for bias-correction due to misclassification, in AFT models where the survival times follow a standard log-logistic distribution. In addition, a modified version of the MC-SIMEX procedure is also proposed, that provides an advantage in situations where the sensitivity and specificity of classification are unknown. Lastly, the performance of the original MC-SIMEX procedure in lung cancer data provided by the North Central Cancer Treatment Group (NCCTG), is also evaluated.

Key Words: AFT models, Survival analysis, Log-logistic distribution, MC-SIMEX, Cancer

APPLICATION OF THE MISCLASSIFICATION SIMULATION EXTRAPOLATION (MC-SIMEX) PROCEDURE TO LOG-LOGISTIC ACCELERATED FAILURE TIME (AFT) MODELS IN SURVIVAL ANALYSIS

by

VARADAN SEVILIMEDU

MBBS., Gandhi Medical College, India, 2004

MPH., Georgia Southern University, 2010

A Dissertation Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PUBLIC HEALTH

STATESBORO, GEORGIA

© 2017

VARADAN SEVILIMEDU

All Rights Reserved

APPLICATION OF THE MISCLASSIFICATION SIMULATION EXTRAPOLATION (MC-SIMEX) PROCEDURE TO LOG-LOGISTIC ACCELERATED FAILURE TIME (AFT) MODELS IN SURVIVAL ANALYSIS

by

VARADAN SEVILIMEDU

Major Professor:	Lili Yu
Committee:	Hani M. Samawi
	Haresh Rochani

Electronic Version Approved:

December 2017

ACKNOWLEDGMENTS

I would like to thank Drs. Lili Yu, Hani Samawi and Haresh Rochani of the Department of Biostatistics at the Jiann Ping Hsu College of Public Health, for providing their time and effort in making this dissertation a success.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

INTRODUCTION

Survival analysis is the study of time-to-event outcomes, and is commonly used in morbidity and mortality analyses[1]. The accuracy of estimation of parameters in survival models depends upon the correct specification of binary covariates in the model. However, correct specification of binary covariates seldom occurs, thus resulting in biased parameter estimates. For example, misclassification of immunization status resulted in biased hazard estimates of preterm birth, in a study by Ahrens et al. in 2012[2]. Misclassification of radiation exposure status resulted in biased hazard estimates in a study by Prentice in 1982[3, 4]. Misclassification of socioeconomic status resulted in biased estimates of risk of chronic disease, in a study by Kauhanen et al. in 2006[5, 6]. Despite the common occurrence of misclassification error, more research in this area is still needed.

Misclassification error can be classified into non-differential and differential misclassification error. Non-differential misclassification error occurs when the information provided by W (misclassified or naive covariate), about Y (response) is irrelevant as long as its corresponding true covariate X and the other confounding covariate Z are available. In this case, W is called the surrogate for X. For example, classifying an individual as hypertensive based on his/her systolic blood pressure measurement on a single day (W), versus systolic blood pressure measurement over a prolonged period of time (X) can result in non-differential misclassification error[7]. On the other hand, if W provides additional information about Y, even when X and Z are already available, then a differential misclassification error ensues. For example, assigning an individual to a category of high risk for coronary heart disease, based upon total cholesterol measurements as opposed to low density lipoprotein (LDL) measurements, can result in differential misclassification error [8]. In this dissertation, we focus mainly on non-differential misclassification error.

Survival analysis broadly employs two models: the cox regression model and the accelerated failure time model (AFT). The Cox regression model regresses the risk/hazard of a certain event, at a certain time, on the risk/hazard at baseline and on the covariates included in the model. The AFT model, on the other hand regresses the log of the time of occurrence of an event on the covariates of interest[9, 10]. The effect of misclassification has been well studied in Cox regression models[11]. Ahrens et al.[2] studied the effect of misclassification using the probabilistic bias analysis in a Cox regression model. Cole et al.[12] used the regression calibration and multiple imputation approach to correct for the bias caused by misclassification in a Cox proportional hazards model. Zucker et al.[13] used the weighted least squares methods for correction of misclassification in a Cox proportional hazards model, followed by the pseudo-partial likelihood[14] and the corrected score function approach[15, 16]. Bang et al.[17] apply the pooled estimation technique proposed by Spiegelman in 2001[18], to the Cox proportional hazards model. Zhou and Pepe[19] used an estimated partial likelihood function to correct for misclassification-bias in Cox regression models. However, the effect of misclassification has not been studied extensively in AFT models, despite the transparent interpretation provided by them[20, 21]. Bang et al.[17] studied the effect of misclassification in survival data where the survival times follow the Weibull distribution. Slate et al.[22] studied the effect of misclassification in a log-normal AFT model. The Weibull and the log normal distributions can only model survival data where the hazard rate is monotonic. However, survival data in which the hazard rate does not follow a monotonic pattern, is also common. For example, breast cancer and lung cancer [23, 24]. The log-logistic distribution is a very popular distribution to model such non-monotonic patterns[23]. Despite the importance of log-logistic distribution in survival studies[23], and its flexibility in accommodating non-monotonic hazards[23, 24], the effect of misclassification in log-logistic AFT models has not been explored yet. Therefore, in this dissertation, we study the effect of non-differential misclassification of binary covariates in a log-logistic AFT model.

There are several methods to handle misclassified data. One such method, the MC-

SIMEX (Misclassification Simulation Extrapolation), is a simulation-based method that makes efficient use of misclassification rates (sensitivity and specificity) to produce bias-corrected estimates. MC-SIMEX is a flexible approach which only requires the presence of a consistent estimator in the absence of misclassification error. In this dissertation, we employ the MC-SIMEX method to handle non-differential misclassification of binary covariates in a log-logistic AFT model. Further details of the MC-SIMEX procedure and the log-logistic AFT model are given in the Methodology section.

Chapter 2

LITERATURE REVIEW

In this chapter, we review the methods that have been used for the correction of bias caused by misclassification error in survival analysis. Over the period of the past three decades, several methods have been proposed and applied to correct for bias in parameter estimates caused by misclassification error.

2.1 Regression calibration

Regression calibration (RC) is a well known method used for correction of bias caused by measurement error[25, 26]. In this method, the value of the true covariate X is estimated by regressing X on the naive covariate W [27, 28]. The estimate thus obtained is then used as a substitute for X in non-validation data[27]. The standard errors of these estimates are calculated by using statistical techniques such as bootstrapping or sandwich methods[27].

Even though the RC method is predominantly used for correction of measurement error in continuous covariates[29, 30], its convenience and ease of interpretation has led to its use, even in binary covariates that are prone to misclassification[17]. Cole et al.[12] studied the effect of misclassification of categorized glomerular filtration rate (GFR) on the 4 year incidence of end stage renal disease (ESRD) using regression calibration in the Cox proportional hazards model. Bang et al.[17], used the regression calibration approach to correct for bias caused by misclassification, in a simulation study, where the survival times followed Weibull distribution.

The regression calibration method assumes that the RC model offers a good fit to the data and that the censoring mechanism involved is independent of the conditional distribution of X given W. The advantage of the RC method is that it is convenient and most popular for discrete

data and non-normal data[17, 31].

2.2 Pooled estimation

Spiegelmann[18] proposed the pooled estimation method to increase the efficiency of parameter estimates obtained through regression calibration. This method involves calculating the weighted averages of coefficients obtained, both from regression calibration and from primary regression in validation data. Bang et al.[17], applied the pooled estimation method to survival data, using the Cox regression model.

While the pooled estimation technique provides the advantage of improved efficiency, its performance is also contingent upon availability of large validation datasets. However, in the context of Cox regression models, the availability of large validation datasets is not always feasible[17].

2.3 Multiple Imputation (MI)

Multiple imputation was originally developed by Rubin et al.[32, 33], to correct for bias caused in parameter estimates, due to missing values in the true covariate. MI involves fitting a logistic regression model between the true covariate X and the naive covariate W, in the validation data i.e. $logitP(X = 1|W)$. The naive covariate in the non-validation data is then replaced by the corrected value (0 or 1) by using the estimated probability from the logit function[17]. Cole et al.[17, 12], used the multiple imputation for measurement error (MIME) algorithm in the Cox proportional hazards model, assuming that data was missing at random (MAR)[33].

The advantage of using an MI procedure is that it uses the values of true covariates, whenever available[17]. In addition to this, it can handle differential measurement error better

than other methods[34, 35]. Finally, it is also very user-friendly and is easily available in any standard statistical package[17]. However, it has two disadvantages, one being that the correct specification of the model is crucial for its successful performance. The second disadvantage is that MI is harder to implement in data with censored outcomes[36, 37].

2.4 Corrected score function

Zucker et. al.[15, 16, 38], suggested a corrected score function approach, to correct for bias caused by misclassification of covariates in Cox regression. If the true score function in the absence of misclassification is represented by $\Psi_{true}(Y, Z, X, \theta)$, then the corrected score function in the presence of misclassification of X can be represented as $\Psi_{CS}(Y, Z, W, \theta)$, where the expected value of the corrected score function equals the true score function[27]. This corrected score function is then used for the estimation of the parameter vector θ and the calculation of standard errors, using procedures such as the bootstrap method or the sandwich method[17]. Augustin[38, 39] proposes an exact corrected score estimate for the proportional hazards model in the presence of heteroscedastic measurement error.

The corrected score function can accommodate situations where the validation sample is not representative of all study participants[17]. Another distinct advantage of the corrected score function method is that it allows for dependence of the censoring mechanism on the true exposure variable X. However, there is loss of efficiency in estimating the Π matrix, which is used for estimating the value of the true variable X from W[17]. In addition, when the number of individuals at risk gets smaller as time progresses, numerical problems are known to occur in calculating the corrected score function[17, 40, 41].

2.5 Estimated partial likelihood function

Based on their previous work on uncensored data[42, 43], Zhou and Pepe[19] proposed an estimated partial likelihood for inference using information from both validation (X) and non-validation data (W). For non-validation data, they calculate the empirical risk function by averaging the values of risk functions for individuals in the validation data, that have the same covariate value. The total estimated risk function is the sum of risk functions over the validation dataset and the non-validation dataset. The relative risk parameter estimate is then obtained by maximizing the estimated partial likelihood function.

The estimated partial likelihood approach does not make assumptions regarding the baseline hazard function nor the conditional distribution of X given W, which is estimated non-parametrically. However, the disadvantage with this method is that when the dimension of W is large, the sample size for each substratum of W may be small, which may result in unstable estimates. A second disadvantage with the estimated partial likelihood method is that it assumes that the validation sample comes from a simple random sample, a non-adherence to which can result in unstable estimates[19].

2.6 Misclassification Simulation extrapolation (MC-SIMEX)

The Simulation extrapolation method (SIMEX) was first proposed by Cook and Stefanski[44] in 1994 to correct for bias caused by measurement error in continuous covariates. He et al.[45] first proposed the use of SIMEX method in survival analysis when continuous covariates were subject to measurement error, using data from the Busselton Health Study[46]. Kuchenhoff et al.[47, 44] in 2006 came up with a modification of the SIMEX procedure that could be applied to a situation where there is measurement error in binary/categorical variables. Since the measurement error in categorical variables is equivalent to misclassification, they called it the misclassification SIMEX or simply MC-SIMEX. Slate et al.[22] applied the MC-SIMEX procedure to evaluate the effect of misclassification in periodontal outcomes, in a log-normal AFT model. Bang et al.[17], in their re-

view, evaluate the MC-SIMEX procedure, by using a Poisson approximation[48, 49] to the Weibull AFT model. The details of the MC-SIMEX procedure are provided in the Methodology section.

Chapter 3

METHODOLOGY

The main purpose of this dissertation is to extend the works of Bang et al.[17] and Slate et al.[22] by applying the MC-SIMEX procedure to the log-logistic distribution in AFT models. We build a model that has a dependent variable that follows log-logistic distribution and subject to right censoring, and a mis-specified binary variable X along with a correctly measured continuous confounding variable Z.

3.1 Accelerated failure time models (AFT)

3.1.1 Basic notation and formula

Let $f(t)$ be the probability distribution function of the continuous time variable. Then the probability that an event occurs within a given time interval, say, (0,t) is the cumulative distribution function of the random variable T [1].

$$F(t) = Pr(T \leq t) = \int_0^t f(u)du \tag{3.1}$$

The survival function $S(t)$ is the complement of the cumulative density function [1]. In other words, it is the probability that the individual will survive beyond a time t[1].

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t). \tag{3.2}$$

So, given $t \to \infty$, $S(0) = 1$ and $S(\infty) = 0$ [1]. The probability density function $f(t)$ can also be written in terms of the survival function as

$$f(t) = -\frac{dS(t)}{dt}. \tag{3.3}$$

The hazard function $h(t)$ is the instantaneous rate of failure at time t [1].

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr[T \in (t, t + \Delta t) | T \geq t]}{\Delta t}, \tag{3.4}$$

or equivalently,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)} \frac{dS(t)}{dt} = \frac{-dlogS(t)}{dt}. \tag{3.5}$$

The above equations show that the three functions, namely $f(t)$, $S(t)$ and $h(t)$ are intimately related to each other. If one of these functions is available, the other two can be easily calculated. For example, $S(t)$ can be written as an inverse function of equation (3.5) as:

$$S(t) = \exp(-\int_0^t h(u)du) = \exp[-H(t)], \tag{3.6}$$

where $H(t)$ is the integration of all hazard rates upto time t and is known as the cumulative hazard function at time t [1]. Alternatively, $H(t)$ can also be written in terms of $S(t)$ as:

$$H(t) = -logS(t). \tag{3.7}$$

Furthermore, the probability density function can also be written in the following form, from equations (3.5) and (3.6):

$$f(t) = h(t) \exp(-\int_0^t h(u)du). \tag{3.8}$$

3.1.2 Specifications of AFT

The AFT model is written as the regression model of the log of time over covariates [1]. Suppose that $Y = \log(T)$ is linearly associated with the covariate vector x. Then

$$Y = \mu^* + x'\beta^* + \tilde{\sigma}\epsilon, \tag{3.9}$$

with location parameter $x'\beta^*$ and the scale parameter $\tilde{\sigma}$. The term ϵ represents the random error whose distribution is determined by the form of the survival function of time $S(t)$, its cumulative

distribution function $F(t)$ and its probability density function $f(t)$[1].

From equation (3.9), it can be deduced that the survival function for individual i at time t can be written as

$$\begin{aligned} S_i(t) &= P[(\mu^* + x_i'\beta^* + \tilde{\sigma}\epsilon_i) \geq \log t], \\ &= P(\epsilon_i \geq \frac{\log t - \mu^* - x_i'\beta^*}{\tilde{\sigma}}). \end{aligned} \tag{3.10}$$

The survival function $S(t)$ can be modeled with respect to $\log t$ as a function of a fixed component $x'\beta$ and a random component ϵ[1].

$$S(t|x) = S_0(\frac{\log t - \mu^* - x'\beta^*}{\tilde{\sigma}}), \qquad -\infty < \log t < \infty. \tag{3.11}$$

Similarly, considering that $H(t) = -logS(t)$, the cumulative hazard function can be expressed in terms of equation (3.11) as

$$\begin{aligned} H(t|x) &= -\log S_0(\frac{\log t - \mu^* - x'\beta^*}{\tilde{\sigma}}), \\ &= H_0(\frac{\log t - \mu^* - x'\beta^*}{\tilde{\sigma}}). \end{aligned} \tag{3.12}$$

where $-\infty < \log t < \infty$. Similarly, differentiating equation (3.12) gives the following hazard function:

$$h(t|x) = \frac{1}{\tilde{\sigma}t} \quad h_0(\frac{\log t - \mu^* - x'\beta^*}{\tilde{\sigma}}), \qquad -\infty < logt < \infty. \tag{3.13}$$

In AFT models, the effect of covariates is such that if $\exp(x'\beta) > 1$, then a deceleration of the survival (time) process ensues and if $\exp(x'\beta) < 1$, then an acceleration of the survival (time) process ensues[1, 20].

3.1.3 Likelihood function of AFT models

Statistical inference in survival analysis is unique in the sense that censoring plays an important role in determination of likelihood functions [1]. Censoring is usually assumed to be random in the sense that conditional upon the model parameters, the censoring times are independent of each other and also of the survival times[50]. Specifically to an individual, and given the parameter vector θ, survival processes are dependent on three random variables, namely observed t_i, δ_i and x_i. The value t_i is defined as the minimum of event time T_i and censoring time C_i, x_i is the covariate vector and δ_i is given by

$$\begin{aligned} \delta_i = 0 \quad if \quad T_i > t_i, \\ \delta_i = 1 \quad if \quad T_i = t_i. \end{aligned} \tag{3.14}$$

Given the covariate vector x_i and parameter vector θ, the likelihood function for a group of n individuals is given by[51]

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} f(t_i; \theta, x_i)^{\delta_i} S(t_i; \theta, x_i)^{1-\delta_i}. \tag{3.15}$$

As can be inferred from equation (3.15), when $\delta_i = 1$ the likelihood function takes on the value of the probability density function for the occurence of an event. When $\delta_i = 0$, the likelihood function takes on the value of the probability of survival beyond censoring time t. In other words, we can see that the likelihood function takes on a value for both censored and uncensored observations [1]. The same likelihood function can be written in terms of a parametric regression model with a baseline hazard function and a vector of coefficients β.

$$L(\theta) = \prod_{i=1}^{n} [h_0(t) \exp(x_i'\beta)]^{\delta_i} \exp[-\int_0^t h_o(u) \exp(x_i'\beta) du]. \tag{3.16}$$

Taking the log values on both sides of equation (3.16), a log likelihood function can be derived as

$$\log L(\theta) = \sum_{i=1}^{n} \left\{ \delta_i [\log h_0(t) + x_i'\beta] - \int_0^t h_0(u)du \exp(x_i'\beta) \right\}. \tag{3.17}$$

The same likelihood function can be easily re-parametrized for applicability in the AFT model as follows[1]:

$$L(\theta) = \prod_{i=1}^{n} \left\{ h_0(t)[t \exp(-x_i'\beta^*)] \exp(-x'\beta^*) \right\}^{\delta_i} \exp \left\{ - H_0[t \exp(-x'\beta^*)] \right\}. \tag{3.18}$$

Finally, the log likelihood function of the AFT regression model can be obtained as follows:

$$logL(\theta) = \sum_{i=1}^{n} \left\{ \delta_i [\log h_0(t) + \log t - (x_i'\beta)^2] - H_0[t \exp(-x'\beta^*)] \right\}. \tag{3.19}$$

3.2 Log-logistic AFT regression models

3.2.1 An overview of the log-logistic distribution

A log-logistic distribution is a non-monotonic distribution and is most suitable for analysis of certain kinds of cancer data [23]. The log-logistic model is especially useful in situations where the hazard rates of different groups of individuals converge over time [23, 24]. A random variable T is said to have a log-logistic distribution if the $log(T)$ has a logistic distribution [52]. The cumulative density function of a log logistic distribution is given by

$$F(T, \alpha, \beta) = \frac{1}{1 + (\frac{t}{\alpha})^{-\beta}}. \tag{3.20}$$

where $t > 0$, $\alpha > 0, \beta > 0$ [52]. The probability density function $f(t)$ can be easily derived from the first derivative of the cumulative density function with respect to T.

$$f(t, \alpha, \beta) = \frac{\frac{\beta}{\alpha}(\frac{t}{\alpha})^{\beta-1}}{(1 + (\frac{t}{\alpha})^{\beta})^2}. \tag{3.21}$$

3.2.2 Specifications of the log-logistic AFT regression models

The log-logistic AFT model can be conveniently specified in the form of equation (3.9) when the random error term ϵ follows the standard logistic distribution. To put it more simply, event/censoring time T follows log-logistic distribution if the $\log$ of T follows standard logistic distribution[1].

The cumulative density function of ϵ in equation (3.9) can be written as

$$F(\epsilon) = P[\varepsilon < \epsilon] = \frac{\exp(\epsilon)}{\left[1 + \exp(\epsilon)\right]}, \quad -\infty < \epsilon < \infty. \tag{3.22}$$

where $\epsilon = \frac{y - x'\beta^*}{\tilde{\sigma}}$ and $y = \log t$. Note that the intercept parameter μ^* is embedded in the vector of coefficients β^*. The survival function $S(\epsilon)$ can be derived from the cumulative hazard function given above, by taking its complement. This gives

$$S(\epsilon) = [1 + \exp(\epsilon)]^{-1}, \quad -\infty < \epsilon < \infty. \tag{3.23}$$

The hazard function $h(\epsilon)$ can simply be derived by using equation (3.5). This gives

$$h(\epsilon) = \frac{\exp(\epsilon)}{1 + \exp(\epsilon)}, \quad -\infty < \epsilon < \infty. \tag{3.24}$$

and $f(\epsilon)$ is derived by multiplying the hazard function with the survival function which gives

$$f(\epsilon) = \frac{\exp(\epsilon)}{(1 + \exp(\epsilon))^2}, \quad -\infty < \epsilon < \infty. \tag{3.25}$$

Given the above three AFT regression functions, the likelihood function for a sample of n individuals can be written as

$$L(\epsilon) = \prod_{i=1}^{n} \left[\frac{exp(\epsilon_i)}{1 + \exp(\epsilon_i)}\right]^{\delta_i} \left[\frac{1}{1 + \exp \epsilon_i}\right], \qquad -\infty < \epsilon < \infty. \tag{3.26}$$

Finally, the log-likelihood can be derived by taking the log of the above likelihood function

$$LogL(\epsilon) = \sum_{i=1}^{n} \Big[\delta_i \epsilon_i - (1+\delta_i) log(1+\exp(\epsilon_i))\Big], \quad -\infty < \epsilon < \infty. \tag{3.27}$$

where $\epsilon_i = \left[\frac{logt_i - x_i'\beta}{\tilde{\sigma}}\right]$.The parameter estimates are then obtained by maximizing the above log-likelihood function, as in any other standard inference procedure. The inverse of the information matrix then gives the variance covariance matrix of the parameter estimates.

3.3 An overview of the MC-SIMEX procedure

An overview of the original MC-SIMEX procedure is provided in section 3.3.1 followed by an overview of our modified MC-SIMEX procedure in section 3.3.2.

3.3.1 The original MC-SIMEX procedure

The probabilities of mis-classifcation can be denoted in the form of a misclassification matrix which is given by:

$$\Pi = \begin{bmatrix} \pi_{00} & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{11} \end{bmatrix}. \tag{3.28}$$

as described in Kuchenhoff et al[47, 53], where π_{11} is the sensitivity and π_{00} is the specificity of classification.

The parameter of interest is β^* (in equation 3.9) with the limit of the naive estimator denoted by $\hat{\beta}^*$. The proof for the existence of $\hat{\beta}^*$ and its estimation is given in the works of White et al., 1982[54]. Since the estimate of $\hat{\beta}^*$ depends on the misclassification matrix, we denote it by $\hat{\beta}^*(\Pi)$, where Π is a k x k matrix with k being the number of categorical outcomes of X. For SIMEX, the function is defined by:

$$\lambda \rightarrow \hat{\beta}^*(\Pi^\lambda), \tag{3.29}$$

indicating that $\hat{\beta}^*(\Pi^\lambda)$ (the value of $\hat{\beta}^*$ at a particular level of misclassification Π^λ) is a function

of λ. Assuming that the misclassification matrix Π^{λ} is at least positive semidefinite, Π^{λ} can be decomposed spectrally as $\Pi^{\lambda} := E\Lambda^{\lambda}E$, where Λ is the diagonal matrix of eigenvalues and E is the corresponding matrix of eigenvectors. Taking equation (3.29) into consideration, it can be stated that if W_1 is related to X through the misclassification matrix Π and W_2 is related to W_1 through the misclassification matrix Π^{λ}, then W_2 is related to X by the misclassification matrix $\Pi^{1+\lambda}$, given the two misclassification mechanisms are independent. If it is assumed that the conditions $\pi_{00} > 0.5$ and $\pi_{11} > 0.5$ are satisfied, then the existence of Π^{λ} is ensured[13, 47].

3.3.1.1 Simulation and extrapolation:

The MC-SIMEX procedure consists of a simulation step that simulates datasets with varying degrees of misclassification of a binary covariate using the misclassification matrix Π^{λ} and the extrapolation step where the corresponding parameter estimates produced with each degree of misclassification are extrapolated using a parametric function of the form[47]:

$$\lambda \rightarrow \hat{\beta}^{*}(\Pi^{\lambda}) \approx D(1+\lambda, \Gamma). \tag{3.30}$$

where D is the quadratic extrapolation function and Γ is the vector of parameters for the quadratic extrapolation function. In other words, $D(1+\lambda, \Gamma) = \Gamma_0 + \Gamma_1(1+\lambda) + \Gamma_2(1+\lambda)^2$. Details of the simulation step and the extrapolation step follow.

Simulation step: For a fixed grid of values $(\lambda_1.....\lambda_m)$, L data sets are simulated for each value of λ. The misclassified X, i.e. W, for each of the L datasets is given by:

$$W_{l,i}(\lambda_k) = MC(\Pi^{\lambda})W, \tag{3.31}$$

where i=1,.....n; l=1,.....L; k = 1,.....m. In other words, for a particular value of λ, say λ_k, $W_l(\lambda_k)$ is obtained by inflating the misclassification in W by a factor λ_k. The naive estimator is then

obtained as[47]:

$$\hat{\beta^*}_{na} = L^{-1} \sum_{1}^{L} [\hat{\beta}_{na}(Y_i, W_{l,i}(\lambda_k), Z_i)], \tag{3.32}$$

where i=1...n and k = 1,....m. In other words, the naive estimator for a particular λ_k is obtained by averaging the values of the naive estimators over L bootstrap samples.

Extrapolation step: The estimator $\hat{\beta}_{simex}$ is then obtained by prediction using the parametric model $D(1 + \lambda, \Gamma)$. That is, after the parameter Γ is estimated, we extrapolate $D(1 + \lambda, \Gamma)$ to a point on the y-axis where $\lambda = -1$ or equivalently, $1 + \lambda$ = 0, then

$$\hat{\beta}_{simex} = D(0, \Gamma), \tag{3.33}$$

which corresponds to $\lambda = -1$. The estimator $\hat{\beta}_{simex}$ is consistent when the $\hat{\Pi}$ is appropriately specified[47].

3.3.1.2 Calculation of the extrapolant function for a simple linear model:

Kuchenhoff et al. [47] showed that under certain situations, the quadratic function offers a suitable approximation for the exact extrapolation function. Those situations were: linear regression with misclassified X, probability estimation, logistic regression with misclassified Y, logistic regression with misclassified X and ordinal logistic regression with misclassified Y. This section highlights the first of the five situations considered by Kuchenhoff et al.[47], that being, linear regression with misclassified X. The rationale behind choosing this situation is that it is directly related to the simple AFT model that we are considering in this dissertation, where the response variable Y is the natural logarithm of the event time($Y = \log(T)$).

$$E(Y|X) = \beta_0 + \beta_1 X. \tag{3.34}$$

Considering that a random variable W_1 is related to X by a misclassification matrix Π^{λ}, we have $E(Y|W_1) = \beta_0 + \beta_1 P(X = 1|W_1)$. Denoting the marginal probability P(X=1) as π_x the following can be derived:[47],

$$E(Y|W_1) = \beta_0^* + \beta_1^* W_1. \tag{3.35}$$

$$\delta = \det(\Pi) = \pi_{00} + \pi_{11} - 1. \tag{3.36}$$

$$\beta_0^* = \beta_0 + \beta_1 \frac{(1 - \pi_{11})\pi_x}{\pi_{00} - \delta\pi_x}. \tag{3.37}$$

$$\beta_1^* = \beta_1 \frac{\delta(1 - \pi_x)(\pi_x)}{(1 - \pi_{00} + \delta\pi_x)(\pi_{00} - \delta\pi_x)}. \tag{3.38}$$

$$\Pi^{\lambda} = \frac{1}{1 - \delta} \begin{bmatrix} 1 - \pi_{11} + (1 - \pi_{00})\delta^{\lambda} & (1 - \pi_{11})(1 - \delta^{\lambda}) \\ (1 - \pi_{00})(1 - \delta^{\lambda}) & 1 - \pi_{00} + (1 - \pi_{11}\delta^{\lambda}) \end{bmatrix}. \tag{3.39}$$

The exact form of extrapolant function is then calculated by plugging in the values of the matrix Π^{λ} into the equation 3.38. It has been shown by Kuchenhoff et al.[47] that equation 3.38 provides a reasonable approximation to a quadratic function over a range of values of λ. Even though the equations (3.36-3.39) were stated by Kuchenhoff et al.[47], the proofs for the equations have not been provided. Therefore, the proofs are provided below:

3.3.1.3 Proof for 3.36

The determinant δ of the matrix

$$\Pi = \begin{bmatrix} \pi_{00} & 1 - \pi_{11} \\ 1 - \pi_{00} & \pi_{11} \end{bmatrix}.$$

is given by:

$$\begin{aligned} \delta &= \pi_{00}\pi_{11} - (1 - \pi_{00})(1 - \pi_{11}) \\ &= \pi_{00} + \pi_{11} - 1. \end{aligned}$$

3.3.1.4 Proof for 3.37: β_0:

Given that $\pi_{00} = P(W_1 = 0|X = 0); \pi_{11} = P(W_1 = 1|X = 1)$ and $P(X = 1) = \pi_x$,

$$E(Y|W_1) = \beta_0 + \beta_1 * P(X = 1|W_1) = \beta_0^* + \beta_1^* W_1.$$

When $W_1 = 0$,

$$\begin{aligned}
\beta_0 + \beta_1 * P(X = 1|W_1 = 0) &= \beta_0^*, \\
\implies \beta_0^* &= \beta_0 + \beta_1 \frac{P(X = 1, W_1 = 0)}{P(W_1 = 0)}, \\
= \beta_0 + \beta_1 \frac{P(W_1 = 0|X = 1)P(X = 1)}{P(W_1 = 0|X = 0)P(X = 0) + P(W_1 = 0|X = 1)P(X = 1)}&, \\
= \beta_0 + \beta_1 \frac{(1 - \pi_{11})\pi_x}{\pi_{00}(1 - \pi_x) + (1 - \pi_{11})\pi_x}&, \\
= \beta_0 + \beta_1 \frac{(1 - \pi_{11})\pi_x}{\pi_{00} - \delta\pi_x}&.
\end{aligned}$$

3.3.1.5 Proof for 3.38: β_1:

When W_1=1,

$$\begin{aligned}
E(Y|W_1 = 1) &= \beta_0 + \beta_1(P(X = 1|W_1 = 1),) \\
&= \beta_0 + \beta_1 \frac{P(W_1 = 1, X = 1)}{P(W_1 = 1)}, \\
= \beta_0 + \beta_1 \frac{P(W_1 = 1|X = 1)P(X = 1)}{P(W_1 = 1|X = 0)(1 - \pi_x) + P(W_1 = 1|X = 1)\pi_x}&, \\
= \beta_0 + \beta_1 \frac{\pi_{11}\pi_x}{(1 - \pi_{00})(1 - \pi_x) + \pi_{11}\pi_x}&, \\
= \beta_0^* + \beta_1^* W_1.&
\end{aligned}$$

Given that $\beta_0^* = \beta_0 + \beta_1 \frac{(1-\pi_{11})\pi_x}{\pi_{00}-\delta\pi_x}$, a simple substitution yields:

$$\beta_1^* = \beta_1 \frac{\pi_{11}\pi_x}{(1-\pi_{00})(1-\pi_x)+\pi_{11}\pi_x} - \beta_1 \frac{\pi_x(1-\pi_{11})}{\pi_{00}-\pi_x\delta},$$
$$= \beta_1 \pi_x \left[\frac{\pi_{11}\pi_{00} - \pi_x\pi_{11}\delta - (1-\pi_{11})(1-\pi_{00}+\delta\pi_x)}{(\pi_{00}-\pi_x\delta)(1-\pi_{00}+\delta\pi_x)} \right],$$
$$= \frac{\beta_1\pi_x(1-\pi_x)\delta}{(\pi_{00}-\pi_x\delta)(1-\pi_{00}+\delta\pi_x)}.$$

3.3.1.6 Proof for 3.39: Π^{λ}:

The Eigenvalues of the matrix

$$\Pi = \begin{bmatrix} \pi_{00} & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{11} \end{bmatrix}$$

is obtained by solving the equation

$$\left| \Pi - xI \right| = 0,$$

where I is the 2X2 identity matrix and x is the eigenvalue.

$$\begin{bmatrix} \pi_{00}-x & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{11}-x \end{bmatrix} = 0,$$
$$\implies (\pi_{00}-x)(\pi_{11}-x) - (1-\pi_{00})(1-\pi_{11}) = 0$$

Solving the above equation gives the following eigenvalues:

$$e_1 = \delta \text{ and } e_2 = 1.$$

The eigenvectors for the corresponding eigenvalues are obtained by solving the following equation for each eigenvalue ($e_1 = \delta$ and $e_2 = 1$).

$$\begin{bmatrix} \pi_{00}-x & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{11}-x \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = 0,$$

where $\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$ is the eigenvector. Solving the above matrix equation gives the following results for

eigenvectors:

When $e_1 = \delta$,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

When $e_2 = 1$

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1-\pi_{00}}{1-\pi_{11}} \end{bmatrix}.$$

The two eigenvectors for the corresponding eigenvalues can be combined to give a single matrix as follows:

$$E = \begin{bmatrix} 1 & 1 \\ -1 & \frac{1-\pi_{00}}{1-\pi_{11}} \end{bmatrix}.$$

The matrix Π^{λ} can be obtained by spectral decomposition which is as follows:

$$\Pi^{\lambda} = E\Delta^{\lambda}E^{-1},$$

where $E = \begin{bmatrix} 1 & 1 \\ -1 & \frac{1-\pi_{00}}{1-\pi_{11}} \end{bmatrix}$, $\Delta^{\lambda} = \begin{bmatrix} \delta^{\lambda} & 0 \\ 0 & 1 \end{bmatrix}$ and λ is a factor that denotes the degree of measurement error.

3.3.1.7 Estimation of the variance of the MC-SIMEX estimator

The variance of the MC-SIMEX estimator is obtained in the following way: For a single simulation with L replications, we calculate the sample variance of the estimator $\hat{\beta}_{sim}(\lambda_k)$ for each value of λ_k by the formula[27, 55]:

$$\hat{V}_{sim}(\lambda_k) := L^{-1}\sum_{l=1}^{L}(\hat{\beta}_{na}[Y_i, W_{l,i}(\lambda_k), Z_i] - \hat{\beta}(\lambda_k))^2, \tag{3.40}$$

with $V_{sim}(0) := 0$. The variance for each naive estimate is also calculated through the information matrix for each value of λ and denoted by $\hat{V}_{naive}(\hat{\beta}[Y_i, W_{l,i}(\lambda_k), Z_i])$ and

$$\hat{V}_{na}(\lambda_k) = L^{-1} \sum_{l=1}^{L} \hat{V}_{naive}(\hat{\beta}_{na}[(Y_i, W_{l,i}(\lambda_k), Z_i)]). \tag{3.41}$$

The variance of the simex estimator (also known as the Stefanski variance V_{ST}) is then given by the extrapolation of the difference between the sample variance and the variance obtained through the information matrix[27], i.e.

$$\hat{V}_{ST} = \lim_{\lambda \to -1} (\hat{V}_{na}(\lambda) - \hat{V}_{sim}(\lambda)). \tag{3.42}$$

3.3.2 Modified MC-SIMEX procedure

The consistency of the existing MC-SIMEX estimator depends upon the correct specification of the misclassification matrix (Π). However, in real data, the exact misclassification matrix (Π) is seldom known. Therefore, we propose a modified MC-SIMEX method in which we estimate Π. The modified MC-SIMEX procedure can be very useful in real data analysis where the true Π is unknown.

The estimation of the misclassification matrix in the modified MC-SIMEX requires four components: $\hat{\pi}_{00}$, $\hat{\pi}_{11}$, $\hat{\pi}_{10}$ and $\hat{\pi}_{01}$. $\hat{\pi}_{00}$ (specificity) is the conditional probability that the naive covariate W takes the value of 0 given that the value of the true covariate X is 0. $\hat{\pi}_{11}$ (sensitivity) is the conditional probability that W takes on the value of 1 given that the value of X is 1. $\hat{\pi}_{10}$ is the conditional probability that W takes on the value of 1 given that the value of the X is 0. $\hat{\pi}_{01}$ is the conditional probability that W takes the value of 0 given that the value of X is 1. These conditional probabilities can be estimated by calculating the number of rows in the simulated dataset where X and W take on the same value and then dividing it by the number of rows of the simulated dataset where X takes on that value. For example, $\hat{\pi}_{00}$ is obtained by dividing the number of rows in the simulated dataset where $X = 0$ and $W = 0$, by the number of rows where $X = 0$. $\hat{\pi}_{10}$ is then

estimated by subtracting the value of $\hat{\pi}_{00}$, from 1. $\hat{\pi}_{11}$ is estimated by dividing the number of rows in the simulated dataset where $X = 1$ and $W = 1$, by the number of rows where $X = 1$. $\hat{\pi}_{01}$ is then estimated by subtracting $\hat{\pi}_{11}$ from 1. The above mentioned steps can be written in the form of mathematical equations as follows:

$$\begin{aligned} \hat{\pi}_{00} &= P(W = 0|X = 0) \\ \hat{\pi}_{11} &= P(W = 1|X = 1) \\ \hat{\pi}_{10} &= 1 - \hat{\pi}_{00} \\ \hat{\pi}_{01} &= 1 - \hat{\pi}_{11} \end{aligned} \tag{3.43}$$

The estimated misclassification matrix $\hat{\Pi}_m$, for the m^{th} Monte Carlo run is then obtained as follows:

$$\hat{\Pi}_m = \begin{bmatrix} \hat{\pi}_{00} & \hat{\pi}_{01} \\ \hat{\pi}_{10} & \hat{\pi}_{11} \end{bmatrix} \tag{3.44}$$

For each Monte Carlo run, the MC-SIMEX algorithm performs 50 replications for each value of the estimated misclassification matrix ($\hat{\Pi}_m^{\lambda_k}$), where $\lambda_k > 0$. The extrapolation function $D(1+\lambda, \hat{\Gamma})$ is then estimated by plotting the $\hat{\beta}$s that are obtained at each degree of misclassification (λ_k), on the Y-axis against the $(1 + \lambda_k)$s on the X-axis. The resulting curve is then extrapolated to a point on the Y axis where λ = -1 or equivalently, 1+λ = 0, as shown below:

$$\begin{aligned} \hat{\beta}_{simex} &= \hat{D}(1 + \lambda, \hat{\Gamma}) \\ &= \hat{D}(0, \hat{\Gamma}) \end{aligned} \tag{3.45}$$

The estimation of the variance in the modified MC-SIMEX procedure is similar to the existing MC-SIMEX procedure. The addition of the estimation step in the modified MC-SIMEX procedure provides an added advantage in situations when the exact sensitivity and specificity (π_{11} and π_{00}) are unknown.

When dealing with real data, the misclassification matrix is estimated by constructing

2 X 2 contingency tables from the validation data. The 2 X 2 contingency table is constructed as follows:

Table 3.1: *2 X 2 contingency table of binary variable subject to misclassificaiton*

$\frac{X(\text{true})\rightarrow}{W\ (\text{naive})\downarrow}$	$X=0$	$X=1$	Row totals
$W=0$	n_{00}	n_{01}	$n_{w=0}=n_{00}+n_{01}$
$W=1$	n_{10}	n_{11}	$n_{w=1}=n_{10}+n_{11}$
Column totals	$n_{00}+n_{10}$	$n_{01}+n_{11}$	Total (n) = $n_{00}+n_{01}+n_{10}+n_{11}$

where n_{00} denotes the number of observations where $W=0$ and $X=0$, n_{01} denotes the number of observations where $W=0$ and $X=1$, n_{10} denotes the number of observations where $W=1$ and $X=0$ and n_{11} denotes the number of observations where $W=1$ and $X=1$. The conditional probabilities of correct classification and misclassification are then calculated as follows:

$$\begin{aligned}\hat{\pi}_{00} &= \frac{n_{00}}{n_{00}+n_{10}}\\ \hat{\pi}_{10} &= 1-\hat{\pi}_{00}\\ \hat{\pi}_{11} &= \frac{n_{11}}{n_{01}+n_{11}}\\ \hat{\pi}_{10} &= 1-\hat{\pi}_{11}\end{aligned} \tag{3.46}$$

Details of analysis of real data from the North Central Cancer Treatment Group (NCCTG) lung cancer clinical trial are provided in chapter 5.

Chapter 4

SIMULATION STUDY

A simulation study is conducted to evaluate the performance of the MC-SIMEX method in an AFT model where the survival time follows log-logistic distribution. In Section 4.1, we propose a new method to simulate right censored survival data that is computationally less burdensome than existing methods and also saves processing time. Section 4.2 describes the methods used to estimate parameters. Section 4.3 describes the results of the performance of the original and modified MC-SIMEX methods, followed by an analysis of robustness to misspecification of distribution in section 4.4. Section 4.5 describes the conclusion of our simulation study.

4.1 An overview of simulation methods

An overview of the existing simulation method and modified simulation method is provided in sections 4.1.1 and 4.1.2 respectively.

4.1.1 Overview of the existing simulation method:

The existing simulation method generates survival times from a specific distribution and censoring times from a specific distribution (for e.g.: uniform) with an initial upper limit. The upper limit is adjusted iteratively until the censoring percentage falls within the stipulated censoring range. To be more specific, if the censoring rate from a particular iteration is lesser than the lower bound of the stipulated range, then the upper limit is decreased so that the censoring rate increases and falls within the range. In the other case where the censoring rate is higher than the upper bound of the stipulated range, the upper limit is increased so that the censoring rate decreases and falls within the stipulated range. This process is repeated until an appropriate upper limit is reached.

4.1.2 Overview of the proposed simulation method:

We propose a new method to simulate right censored survival data, that achieves the exact level of censoring when the survival times follow a log-logistic distribution (however, this method can also be used for other survival data distributions). Existing algorithms require many number of iterations to achieve the desired censoring rates. The proposed algorithm, on the other hand, requires only one iteration to achieve the exact rate of censoring. Our algorithm also eases computational burden and saves processing time, as opposed to other algorithms. The proposed algorithm follows:

Step 1: Assign a value each for β_0, β_1 and β_2.

Step 2: Generate n random values for the variable X which follows *Bernoulli* distribution with the probability 0.5.

Step 3: Generate n random values for the variable Z which follows a $N(0, 1)$ distribution.

Step 4: Generate n random values for the variable ϵ (residual) which follows a logistic distribution with location zero and scale 1.

Step 5: Generate n natural logarithms of survival times using the following formula:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon; \text{ where } Y = \log T.$$

Step 6: Generate n censoring times which follows a $U \sim (0, 10)$ distribution.

Step 7: Generate a new variable $r = Y - log(c)$ for each of the n observations, where c is the censoring time.

Step 8: Pick the value of r that represents a percentile corresponding to the event rate. For example, if a censoring rate of 30% is desired, we will pick a value of r that corresponds to the 70^{th}

percentile of its distribution.

Step 9: Create a new variable that represents the log of the new survival time which is obtained by deducting the value of r that represents the 70^{th} percentile of its distribution from the original log of the survival time. Say,

$$logT_{new} = Y - rq_{70},$$

where $logT_{new}$ represents the log of the new survival time, rq_{70} is the 70^{th} percentile of the distribution of r. This step allows us to order the new survival times in such a way that 30% of the observations are censored and the remaining uncensored.

Step 10: Obtain a new survival time t_{new} by taking the exponential of the value of $logTnew$ obtained from the previous step.

$$t_{new} = \exp(logT_{new}).$$

Step 11: Generate a new variable y_{new} which is the minimum of t_{new} and c, where c is the censoring time corresponding to t_{new}.

$$Y_{new} = \text{pmin } (t_{new}, c).$$

Step 12: Fit an AFT model using *survreg*[56] procedure in R (install MASS package[57] for survival analysis before implementing *survreg*), with y_{new} as the observed time, and δ as the indicator for censoring. If $t_{new} > c$ then $\delta = 0$, or else, δ takes on the value of 1. X and Z are the explanatory variables. By the end of this step, the $\hat{\beta}_{nmisc}$ ($nmisc$ stands for no misclassification) associated with the true variable X is obtained.

Step 13: Using the *misclass*[58] function in R, generate a naive variable W using the misclassification matrix Π. Fit an AFT model as in Step 12, but with the naive covariate W instead of X. By the end of this step, $\hat{\beta}_{naive}$ associated with the naive covariate W is obtained.

Step 14: Using the *misclass* function in R, generate additional naive covariates W_1, W_2, W_3 and W_4 from true covariate X. These naive covariates represent the misclassified form of the covariate X at $\lambda = 0.8$, 1.2, 1.6 and 2 respectively, where λ is the power of the misclassification matrix Π^{λ}.

Step 15: At each level of misclassification, an AFT model is fit, as described in step 12, using the naive covariate instead of the true covariate. That is, four different AFT models using the naive covariates W_1, W_2, W_3 and W_4 - one in each model, along with the confounding variable Z are fit.

Step 16: Using the quadratic extrapolation function described in chapter 3, the $\hat{\beta}$ estimates ($\hat{\beta}_{W_1}$, $\hat{\beta}_{W_2}$, $\hat{\beta}_{W_3}$ and $\hat{\beta}_{W_4}$) obtained at the corresponding level of misclassification are extrapolated to a point on the Y-axis where λ = -1. The value of $\hat{\beta}$ at this point on the Y-axis, is the $\hat{\beta}_{simex}$ estimate.

Step 17: 50 iterations[22] of steps 14 to 16 are run for each simulation. At the end of 50 iterations, average of the $\hat{\beta}_{simex}$ estimates is calculated to give the final $\hat{\beta}_{simex}$ estimate for that simulation. In addition, the empirical variance, estimated variance and Stefanski variance (V_{ST}) are also obtained using equations 3.40-3.42, as described in chapter 3.

Step 18: At the end of one simulation and 50 replications within the simulation, the MSE, bias, estimated variance and coverage of the true estimator ($\hat{\beta}_{nmisc}$), the naive estimator $\hat{\beta}_{naive}$ and the $\hat{\beta}_{simex}$ estimator are calculated.

Step 20: Steps 1 through 18 are repeated until a total of 500 Monte-Carlo runs are completed. The $\hat{\beta}_{nmisc}$s, $\hat{\beta}_{W}$s and $\hat{\beta}_{simex}$s along with their corresponding MSEs, biases, estimated variances , empirical variances and coverages are averaged over 500 Monte-Carlo runs to give the corresponding final estimates.

Despite the advantages provided by this proposed simulation method (as described at

the start of section 4.1.2), it must be noted that this algorithm also results in a distortion of the value of the intercept in the model. However, since the study of properties of the intercept is not our primary objective, we ignore this distortion. The properties of β_1 and β_2 remain unchanged despite this adjustment. The distribution of the newly generated survival times also remains the same, albeit a change in the expected value (mean) occurs.

4.2 Data simulation and estimation of parameters

In this study, a sample size of 200 is chosen. We consider two covariates, one binary and the other continuous. The binary covariate X, which is subject to misclassification error, is generated from a binomial distribution ($X \sim binom(n, 0.5)$). The continuous covariate Z is generated from a standard normal distribution($Z \sim N(0, 1)$), independent of X. The error term ϵ_i is generated from a standard logistic distribution with $\epsilon_i \sim logistic(0, 1)$ (0 is the value of the location parameter and 1 is the value of the scale parameter). The survival times T are then generated using the following equation:

$$Y = log(T) = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_i$$

$$T = \exp(Y),$$

where the values of β_1 and β_2 are pre-specified. The censoring times c are generated from a uniform distribution with $c \sim U(0, 10)$.

After conducting this preliminary simulation, the algorithm described in steps 7 through 13 of section 4.1.2 is performed. That is, a new variable $r = Y - log(c)$ is created followed by the selection of the value of r, which corresponds to a stipulated percentile of its distribution, say rq_{70}. This is followed by the creation of a new variable $logT_{new}$, that represents the difference between the variable Y and rq_{70}. The new survival time, t_{new} is then obtained by taking the exponential of the variable $logT_{new}$. Censoring statuses are then assigned and an AFT model is fit, as described in steps 12 and 13 of the algorithm mentioned above.

In this dissertation, two situations are considered, wherein the misclassification matrices are $\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$ and $\begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix}$. Estimates of β_1 are obtained for each Monte Carlo run using the true estimator (which is obtained from the AFT model using the true covariate X), naive estimator (which is obtained from the AFT model using the naive covariate W) and the MC- SIMEX estimator. Also, for each run, the corresponding bias and MSE are obtained. In order to obtain the MC-SIMEX estimator, a total of 50 replications were run for each simulation, as done by Slate et al[22]. A total of 500 simulations were run. The estimates of $\hat{\beta}_{nmisc}$, $\hat{\beta}_{naive}$ and $\hat{\beta}_{simex}$ and their corresponding bias and MSE are obtained as follows:

$$\hat{\beta}_{nmisc} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{nmisc_i}$$

$$\hat{\beta}_{naive} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{naive_i}$$

$$\hat{\beta}_{simex} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{simex_i}$$

$$\hat{MSE}_{nmisc} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{nmisc_i} - (\beta))^2$$

$$\hat{MSE}_{naive} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{naive_i} - (\beta))^2$$

$$\hat{MSE}_{simex} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{simex_i} - (\beta))^2$$

$$\hat{bias}_{nmisc} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{nmisc_i} - \beta)$$

$$\hat{bias}_{naive} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{naive_i} - \beta)$$

$$\hat{bias}_{simex} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{simex_i} - \beta)$$

The coverage probability of each estimator is then obtained by first estimating the variance and standard error (SE) of each estimator from the information matrix, as described in section 3.3 of chapter 3. 95% confidence intervals are then estimated by using the following formula:

$$95\%CI = \hat{\beta} \pm 1.96SE$$

The coverage probability is then calculated as the percentage of the occurrences where the 95% CI includes the value of the true parameter.

4.3 Results of the performance of the original and modified MC-SIMEX estimator

Tables 4.1 - 4.4 illustrate the results of the MC-SIMEX procedure when the survival times follow standard log-logistic distribution for 0%, 30%, 50% and 70% levels of censoring, with the true Π being $\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$. Table 4.5 and table 4.6 illustrate the results of the modified MC-SIMEX procedure, under similar specifications as table 4.1 and 4.2. Tables 4.7 - 4.10 illustrate the results of the MC-SIMEX procedure when the survival times follow standard log-logistic distribution for 0%, 30%, 50% and 70% levels of censoring, with the true Π being $\begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix}$. Table 4.11 and table 4.12 illustrate the results of the modified MC-SIMEX procedure, under similar specifications as table 4.7 and table 4.8.

4.3.1 Results for a sensitivity of 80% and specificity of 80%:

The performance of the MC-SIMEX estimator is evaluated for four different combinations of β_1 and β_2, those being: $\beta_1 = -\log 2$ and $\beta_2 = 0.5$, $\beta_1 = -\log 2$ and $\beta_2 = -0.5$,

$\beta_1 = \log 2$ and $\beta_2 = 0.5$ and finally, $\beta_1 = \log 2$ and $\beta_2 = -0.5$. For a Π of $\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$, tables 4.1 - 4.4 show that the MC-SIMEX estimator consistently performs better than the naive estimator. The magnitude of the bias associated with the MC-SIMEX estimator is always lower than that of the naive estimator. The MSE associated with the MC-SIMEX estimator is consistently lower than that of the naive estimator across all levels of censoring. With regard to the coverage probabilities, the MC-SIMEX estimator is shown to perform satisfactorily and consistently better than the naive estimator across all levels of censoring.

Table 4.5 and table 4.6 illustrate the performance of the modified SIMEX procedure using the log-logistic distribution of survival times for a specified true sensitivity of 80% and a true specificity of 80%. It can be seen from table 4.5 and table 4.6 that the bias, MSE and coverage probabilities for the modified SIMEX procedure are satisfactory and comparable to that of the true estimator. A comparison of tables 4.5 and 4.6 to tables 4.1 and 4.2 shows that the performance of the modified MC-SIMEX procedure is comparable to the performance of the original MC-SIMEX procedure and that there are no notable deviations in bias, MSE and coverage probabilities.

Table 4.1: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 80%, *Specificity* = 80%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.063	0.062
Empirical variance	0.064	0.055	0.043
Bias	-0.006	0.266	0.055
MSE	0.064	0.126	0.046
Coverage	0.944	0.828	0.956
Censoring rate = 30%			
Estimated Variance	0.067	0.069	0.068
Empirical variance	0.060	0.070	0.047
Bias	0.011	0.288	0.189
MSE	0.060	0.153	0.083
Coverage	0.960	0.784	0.930
Censoring rate = 50%			
Estimated Variance	0.078	0.079	0.078
Empirical variance	0.074	0.086	0.062
Bias	0.027	0.297	0.125
MSE	0.074	0.174	0.077
Coverage	0.948	0.808	0.948
Censoring rate = 70%			
Estimated Variance	0.108	0.106	0.107
Empirical variance	0.109	0.107	0.098
Bias	0.008	0.277	0.080
MSE	0.109	0.183	0.104
Coverage	0.946	0.846	0.942

Table 4.2: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 80%, *Specificity* = 80%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.063	0.062
Empirical variance	0.057	0.065	0.066
Bias	0.015	0.297	0.093
MSE	0.058	0.153	0.075
Coverage	0.950	0.772	0.938
Censoring rate = 30%			
Estimated Variance	0.067	0.069	0.068
Empirical variance	0.066	0.060	0.062
Bias	0.011	0.286	0.069
MSE	0.066	0.142	0.067
Coverage	0.958	0.818	0.944
Censoring rate = 50%			
Estimated Variance	0.077	0.078	0.078
Empirical variance	0.074	0.082	0.060
Bias	-0.016	0.276	0.167
MSE	0.074	0.158	0.088
Coverage	0.950	0.802	0.936
Censoring rate = 70%			
Estimated Variance	0.110	0.109	0.108
Empirical variance	0.111	0.105	0.072
Bias	-0.014	0.270	0.140
MSE	0.111	0.177	0.091
Coverage	0.952	0.854	0.952

Table 4.3: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 80%, *Specificity* = 80%. *True* β *values are* $\beta_1 = \log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.060	0.062	0.062
Empirical variance	0.062	0.065	0.049
Bias	0.012	-0.252	-0.040
MSE	0.062	0.129	0.050
Coverage	0.950	0.814	0.946
Censoring rate = 30%			
Estimated Variance	0.066	0.067	0.067
Empirical variance	0.065	0.065	0.062
Bias	-0.007	-0.287	-0.047
MSE	0.065	0.147	0.064
Coverage	0.942	0.796	0.936
Censoring rate = 50%			
Estimated Variance	0.077	0.078	0.078
Empirical variance	0.070	0.075	0.061
Bias	-0.025	-0.279	-0.130
MSE	0.07	0.153	0.078
Coverage	0.956	0.820	0.940
Censoring rate = 70%			
Estimated Variance	0.109	0.108	0.108
Empirical variance	0.107	0.110	0.084
Bias	0.012	-0.260	-0.072
MSE	0.107	0.177	0.089
Coverage	0.942	0.864	0.948

Table 4.4: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 80%, *Specificity* = 80%. *True* β *values are* $\beta_1 = \log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.063	0.062
Empirical variance	0.059	0.064	0.052
Bias	0.014	-0.253	-0.091
MSE	0.059	0.128	0.060
Coverage	0.948	0.806	0.946
Censoring rate = 30%			
Estimated Variance	0.066	0.068	0.067
Empirical variance	0.067	0.072	0.052
Bias	-0.011	-0.296	-0.091
MSE	0.067	0.159	0.060
Coverage	0.946	0.784	0.946
Censoring rate = 50%			
Estimated Variance	0.078	0.079	0.078
Empirical variance	0.072	0.078	0.075
Bias	0.012	-0.267	-0.120
MSE	0.072	0.148	0.090
Coverage	0.960	0.826	0.922
Censoring rate = 70%			
Estimated Variance	0.109	0.108	0.108
Empirical variance	0.099	0.114	0.090
Bias	0.034	-0.264	-0.001
MSE	0.100	0.183	0.090
Coverage	0.964	0.844	0.930

Table 4.5: *Results of the modified MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 80%, *Specificity* = 80%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.062
Empirical variance	0.064	0.054
Bias	-0.014	0.193
MSE	0.064	0.091
Coverage	0.946	0.920
Censoring rate = 30%		
Estimated Variance	0.066	0.067
Empirical variance	0.074	0.052
Bias	-0.015	0.172
MSE	0.074	0.081
Coverage	0.944	0.948
Censoring rate = 50%		
Estimated Variance	0.078	0.078
Empirical variance	0.078	0.065
Bias	-0.014	0.034
MSE	0.078	0.066
Coverage	0.958	0.940
Censoring rate = 70%		
Estimated Variance	0.108	0.108
Empirical variance	0.115	0.097
Bias	-0.010	0.009
MSE	0.114	0.097
Coverage	0.936	0.932

Table 4.6: *Results of the modified MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 80%, *Specificity* = 80%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{simex}$
Estimated Variance	0.062	0.062
Empirical variance	0.055	0.062
Bias	-0.001	0.036
MSE	0.055	0.063
Coverage	0.956	0.936
Censoring rate = 30%		
Estimated Variance	0.067	0.067
Empirical variance	0.072	0.045
Bias	0.000	0.073
MSE	0.072	0.051
Coverage	0.934	0.942
Censoring rate = 50%		
Estimated Variance	0.077	0.079
Empirical variance	0.079	0.058
Bias	0.001	0.076
MSE	0.079	0.064
Coverage	0.964	0.946
Censoring rate = 70%		
Estimated Variance	0.110	0.107
Empirical variance	0.115	0.075
Bias	-0.006	0.067
MSE	0.115	0.079
Coverage	0.938	0.944

4.3.2 Results for a sensitivity of 90% and specificity of 70%:

Tables 4.7-4.10 illustrate the performance of the MC-SIMEX estimator in comparison to the naive estimator, when Π is $\begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix}$. The magnitude of the bias associated with the MC-SIMEX estimator is consistently lower than that of the naive estimator at all the levels of censoring. The MSE associated with the MC-SIMEX etimator is also consistently lower than that of the naive estimator. The coverage probability associated with the MC-SIMEX estimator is satisfactory and consistently better than that of the naive estimator at all levels of censoring.

Table 4.11 and table 4.12 illustrate the performance of the modified SIMEX procedure using the log-logistic distribution of survival times for a specified true sensitivity of 90% and a true specificity of 70%. It can be seen from table 4.11 and table 4.12 that bias, MSE and coverage probabilities associated with the modified MC-SIMEX procedure are satisfactory and comparable to that of the true estimator. A comparison of tables 4.11 and 4.12 to tables 4.7 and 4.8 shows that the performance of the modified MC-SIMEX procedure is comparable to that of the original MC-SIMEX procedure and that there are no notable deviations.

4.4 Robustness

In this dissertation, the analysis of robustness is done in by mis-specifying a log-logistic distribution (of survival time), as a Weibull distribution. The results of this misspecification of survival time distribution are described in this section.

Table 4.7: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 90%, *Specificity* = 70%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.065	0.058
Empirical Variance	0.061	0.067	0.047
Bias	0.006	0.274	0.010
MSE	0.061	0.143	0.047
Coverage	0.942	0.786	0.93
Censoring rate = 30%			
Estimated Variance	0.066	0.071	0.062
Empirical Variance	0.066	0.072	0.053
Bias	-0.010	0.279	0.106
MSE	0.066	0.151	0.064
Coverage	0.946	0.804	0.938
Censoring rate = 50%			
Estimated Variance	0.077	0.083	0.075
Empirical Variance	0.077	0.075	0.089
Bias	-0.008	0.220	0.052
MSE	0.077	0.123	0.092
Coverage	0.944	0.88	0.908
Censoring rate = 70%			
Estimated Variance	0.109	0.116	0.105
Empirical Variance	0.108	0.111	0.096
Bias	0.005	0.264	0.010
MSE	0.108	0.181	0.096
Coverage	0.962	0.884	0.934

Table 4.8: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 90%, *Specificity* = 70%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.065	0.059
Empirical Variance	0.065	0.068	0.061
Bias	-0.004	0.250	0.090
MSE	0.065	0.130	0.069
Coverage	0.938	0.81	0.926
Censoring rate = 30%			
Estimated Variance	0.066	0.071	0.064
Empirical Variance	0.072	0.073	0.067
Bias	0.004	0.261	0.071
MSE	0.071	0.141	0.072
Coverage	0.946	0.824	0.944
Censoring rate = 50%			
Estimated Variance	0.077	0.080	0.073
Empirical Variance	0.076	0.079	0.066
Bias	0.031	0.272	0.095
MSE	0.077	0.153	0.075
Coverage	0.944	0.834	0.920
Censoring rate = 70%			
Estimated Variance	0.108	0.108	0.099
Empirical Variance	0.095	0.105	0.089
Bias	-0.007	0.287	0.078
MSE	0.095	0.187	0.095
Coverage	0.970	0.868	0.926

Table 4.9: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 90%, *Specificity* = 70%. *True* β *values are* $\beta_1 = \log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.065	0.059
Empirical variance	0.066	0.069	0.055
Bias	-0.006	-0.258	-0.088
MSE	0.066	0.136	0.062
Coverage	0.934	0.808	0.928
Censoring rate = 30%			
Estimated Variance	0.067	0.072	0.064
Empirical variance	0.064	0.078	0.056
Bias	-0.012	-0.250	-0.042
MSE	0.064	0.140	0.058
Coverage	0.964	0.824	0.942
Censoring rate = 50%			
Estimated Variance	0.077	0.083	0.077
Empirical variance	0.080	0.084	0.077
Bias	0.006	-0.272	-0.086
MSE	0.080	0.158	0.084
Coverage	0.950	0.830	0.926
Censoring rate = 70%			
Estimated Variance	0.110	0.118	0.105
Empirical variance	0.113	0.106	0.101
Bias	0.021	-0.220	0.027
MSE	0.113	0.154	0.102
Coverage	0.948	0.890	0.928

Table 4.10: *Results of the MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 90%, *Specificity* = 70%. *True* β *values are* $\beta_1 = \log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.065	0.059
Empirical variance	0.061	0.066	0.053
Bias	-0.003	-0.262	-0.085
MSE	0.061	0.134	0.060
Coverage	0.932	0.810	0.912
Censoring rate = 30%			
Estimated Variance	0.066	0.071	0.063
Empirical variance	0.061	0.074	0.070
Bias	0.016	-0.243	-0.187
MSE	0.061	0.133	0.104
Coverage	0.950	0.830	0.904
Censoring rate = 50%			
Estimated Variance	0.078	0.084	0.075
Empirical variance	0.082	0.085	0.080
Bias	0.010	-0.247	-0.070
MSE	0.082	0.146	0.085
Coverage	0.944	0.854	0.920
Censoring rate = 70%			
Estimated Variance	0.110	0.117	0.110
Empirical variance	0.091	0.104	0.114
Bias	0.018	-0.243	-0.012
MSE	0.091	0.163	0.114
Coverage	0.964	0.896	0.926

Table 4.11: *Results of the modified MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 90%, *Specificity* = 70%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{simex}$
Estimated Variance	0.060	0.057
Empirical variance	0.070	0.065
Bias	-0.018	0.123
MSE	0.070	0.080
Coverage	0.936	0.904
Censoring rate = 30%		
Estimated Variance	0.066	0.061
Empirical variance	0.068	0.074
Bias	-0.003	0.086
MSE	0.067	0.081
Coverage	0.946	0.916
Censoring rate = 50%		
Estimated Variance	0.077	0.072
Empirical variance	0.078	0.069
Bias	-0.015	0.044
MSE	0.078	0.071
Coverage	0.946	0.938
Censoring rate = 70%		
Estimated Variance	0.107	0.097
Empirical variance	0.111	0.107
Bias	0.023	0.067
MSE	0.111	0.112
Coverage	0.956	0.942

Table 4.12: *Results of the modified MC-SIMEX procedure using log-logistic distribution of survival times. Sensitivity* = 90%, *Specificity* = 70%. *True* β *values are* $\beta_1 = -\log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.059
Empirical variance	0.062	0.059
Bias	0.001	0.106
MSE	0.062	0.070
Coverage	0.944	0.906
Censoring rate = 30%		
Estimated Variance	0.065	0.063
Empirical variance	0.072	0.066
Bias	0.001	0.070
MSE	0.072	0.070
Coverage	0.924	0.922
Censoring rate = 50%		
Estimated Variance	0.078	0.072
Empirical variance	0.083	0.085
Bias	-0.017	0.065
MSE	0.083	0.089
Coverage	0.942	0.926
Censoring rate = 70%		
Estimated Variance	0.109	0.098
Empirical variance	0.114	0.070
Bias	-0.007	0.041
MSE	0.114	0.071
Coverage	0.952	0.948

Tables 4.13 to 4.16 illustrate the effect of misspecification of a log-logistic distribution as a Weibull distribution, for 0%, 30%, 50% and 70% levels of censoring. In tables 4.13 and 4.14, the misclassification matrix used is $\begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$, while in tables 4.15 and 4.16, the misclassification matrix used is $\begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix}$. This analysis of robustness is done for two combinations of pre-specified β_1 and β_2 values. Tables 4.13 to 4.16 show that the MC-SIMEX procedure performs consistently better than the naive estimator in terms of bias, MSE and coverage probabilities and also that the MC-SIMEX procedure is robust to mis-specification of distribution and change in parameter values.

4.5 Conclusion

Tables 4.1 to 4.16 show that the MC-SIMEX estimator is a reliable and valid estimator even under misspecification of distribution. The above tables also show that under varying degrees of misclassification of the binary variable X and at varying levels of censoring, the MC-SIMEX estimates are very close to the true value of βs that are assigned in the simulation. The $\hat{\beta}$ estimates obtained from our modified MC-SIMEX method also proved to be efficient and comparable to the true estimator. However, we will be remiss, if we didn't note the increased bias in the MC-SIMEX estimates, when dealing with a mis-specified distribution. We attribute this finding to chance and reiterate that such findings are not totally surprising, given that similar findings have been reported in the study by Slate et. al.[22]

Table 4.13: *Results of the MC-SIMEX procedure with the log-logistic distribution of survival times, but misspecified as a Weibull distribution. Sensitivity* = 80%*, Specificity* = 80%*. True* β *values are* $\beta_1 = -\log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.063	0.114
Empirical variance	0.065	0.067	0.122
Bias	0.006	0.277	0.124
MSE	0.065	0.143	0.137
Coverage	0.942	0.796	0.922
Censoring rate = 30%			
Estimated Variance	0.066	0.068	0.064
Empirical variance	0.065	0.068	0.070
Bias	-0.003	0.281	0.070
MSE	0.065	0.148	0.075
Coverage	0.948	0.802	0.922
Censoring rate = 50%			
Estimated Variance	0.078	0.079	0.070
Empirical variance	0.074	0.087	0.062
Bias	0.003	0.292	0.171
MSE	0.074	0.172	0.091
Coverage	0.958	0.802	0.916
Censoring rate = 70%			
Estimated Variance	0.110	0.108	0.097
Empirical variance	0.115	0.112	0.068
Bias	0.006	0.286	0.189
MSE	0.115	0.193	0.104
Coverage	0.948	0.834	0.934

Table 4.14: *Results of the MC-SIMEX procedure with the log-logistic distribution distribution of survival times, but misspecified as a Weibull distribution. Sensitivity* = 80%*, Specificity* = 80%*. True* β *values* $\beta_1 = -\log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.062	0.111
Empirical variance	0.059	0.064	0.101
Bias	-0.003	0.268	0.114
MSE	0.059	0.135	0.113
Coverage	0.954	0.818	0.924
Censoring rate = 30%			
Estimated Variance	0.066	0.067	0.064
Empirical variance	0.063	0.068	0.055
Bias	0.015	0.294	0.130
MSE	0.063	0.154	0.071
Coverage	0.954	0.782	0.944
Censoring rate = 50%			
Estimated Variance	0.078	0.079	0.070
Empirical variance	0.074	0.087	0.062
Bias	0.003	0.292	0.171
MSE	0.074	0.172	0.091
Coverage	0.958	0.802	0.916
Censoring rate = 70%			
Estimated Variance	0.108	0.106	0.098
Empirical variance	0.099	0.096	0.062
Bias	0.018	0.313	0.089
MSE	0.099	0.193	0.070
Coverage	0.956	0.846	0.950

Table 4.15: *Results of the MC-SIMEX procedure with the log-logistic distribution distribution of survival times, but misspecified as a Weibull distribution. Sensitivity* = 90%*, Specificity* = 70%*. True* β *values are* $\beta_1 = -\log 2, \beta_2 = 0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.066	0.106
Empirical variance	0.065	0.069	0.114
Bias	-0.003	0.260	0.128
MSE	0.065	0.137	0.130
Coverage	0.942	0.826	0.920
Censoring rate = 30%			
Estimated Variance	0.066	0.070	0.060
Empirical variance	0.067	0.075	0.058
Bias	-0.015	0.257	0.160
MSE	0.067	0.140	0.083
Coverage	0.940	0.818	0.912
Censoring rate = 50%			
Estimated Variance	0.077	0.080	0.063
Empirical variance	0.080	0.076	0.068
Bias	0.010	0.270	0.187
MSE	0.080	0.149	0.102
Coverage	0.932	0.838	0.910
Censoring rate = 70%			
Estimated Variance	0.107	0.106	0.090
Empirical variance	0.112	0.111	0.079
Bias	-0.008	0.253	0.155
MSE	0.112	0.174	0.103
Coverage	0.940	0.864	0.924

Table 4.16: *Results of the MC-SIMEX procedure with the log-logistic distribution distribution of survival times, but misspecified as a Weibull distribution. Sensitivity* = 90%*, Specificity* = 70%*. True* β *values are* $\beta_1 = -\log 2, \beta_2 = -0.5$

Censoring rate = 0 %	$\hat{\beta}_{nmisc}$	$\hat{\beta}_{naive}$	$\hat{\beta}_{simex}$
Estimated Variance	0.061	0.065	0.109
Empirical variance	0.054	0.062	0.136
Bias	0.002	0.259	0.049
MSE	0.054	0.129	0.138
Coverage	0.960	0.846	0.902
Censoring rate = 30%			
Estimated Variance	0.067	0.070	0.059
Empirical variance	0.074	0.083	0.052
Bias	-0.016	0.261	0.122
MSE	0.074	0.151	0.067
Coverage	0.944	0.802	0.934
Censoring rate = 50%			
Estimated Variance	0.076	0.080	0.064
Empirical variance	0.082	0.082	0.063
Bias	-0.001	0.262	0.118
MSE	0.082	0.151	0.077
Coverage	0.944	0.832	0.942
Censoring rate = 70%			
Estimated Variance	0.109	0.108	0.088
Empirical variance	0.119	0.113	0.072
Bias	-0.001	0.271	0.150
MSE	0.118	0.186	0.095
Coverage	0.924	0.842	0.944

Chapter 5

APPLICATION TO LUNG CANCER DATA

5.1 Introduction

In this chapter, an analysis of a lung cancer dataset that was developed by Loprinzi et al.[59, 60] is conducted. The Karnofsky performance scale (KPS) is a widely renowned scale, not only to measure the functional status of patients with debilitating illnesses such as lung cancer, but also to assess their medical needs[61]. It has been shown to be significantly predictive of survival outcomes in patients with lung cancer. It has also been used as an outcome indicator to compare the efficacies of different clinical trial interventions[61]. KPS is a 11-point scale that ranges from 0 - dead to 100 - normal. The detailed classification is given in table 5.1 below[62].

Table 5.1: *Karnofsky performance status scale*

100	Normal, no complaints, no evidence of disease
90	Able to carry on normal activity, minor signs or symptoms of disease
80	Normal activity with effort, some signs or symptoms of disease
70	Cares for self. Unable to carry on normal activity or to do active work
60	Requires occasional assistance, but is able to care for most of his needs
50	Requires considerable assistance and frequent medical care
40	Disabled, requires special care and assistance
30	Severely disabled, hospitalization is indicated although death not imminent
20	Hospitalization necessary, very sick, active supportive treatment necessary
10	Moribund, fatal processes progressing rapidly
0	Dead

As the performance score increases from 0 to 100, the patients functional ability also increases. KPS is of two types: one that is provided by the patient, that is referred to as the patient Karnofsky performance scale and the other that is provided by the physician, which is referred to as the physician Karnofsky performance scale. The two scales have been shown to be highly correlated[59] and hence are reliable indicators. However, room for error exists.

5.2 North Central Cancer Treatment Group (NCCTG) - Lung Cancer Data

The NCCTG lung cancer dataset[59] provides survival outcomes of 228 patients with advanced lung cancer. The variables that are provided in this dataset include (variable names are in italics): *inst*- Institution code, *time*-survival time in days, *status*-censoring status (1=censored, 2=dead), *age*-age in years, *sex*- sex of the patient (Male = 1, Female =2), *ph.ECOG*-Eastern Cooperative Oncology Group performance score (scale of 0-5, 5: dead and 0: normal), *ph.karno*-Karnofsky performance score, as rated by physician on a scale of 0 to 100, *pat.karno*-Karnofsky peformance score, as rated by the patient on a scale of 0 to 100, *meat.cal*-calories consumed at meals and *wt.loss*-weight loss in the last six months.

5.2.1 Misclassification matrix

The performance score reported (one reported by physician and the other reported by the patient) was categorized into two classes, low (0) and high (1). A recorded score of 70 and below was considered as a low score and a recorded score greater than 70, was considered as a high score. This classification was done using the ability to perform normal activity as a benchmark, which corresponds to a score of above 70. The performance category, as reported by the physician was then considered as a true covariate (X) while the performance category reported by the patient was considered as a naive covariate (W). The misclassification matrix is then estimated using the method described in section 3.3.2 of chapter 3. Using this method, the estimated misclassification matrix $\hat{\Pi}$ was $\begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$

5.2.2: Distribution of survival times

The distribution of the survival times was examined using QQ plots, probability plots, plots of cumulative distribution functions and probability density plots. The following distributions were examined: log-normal, log-logistic, Weibull and logistic distribution. The plots were constructed using the *fitdistrplus*[63] package in R 3.2.2. The distribution plots for the above mentioned distributions are provided in figures 1 through 4.

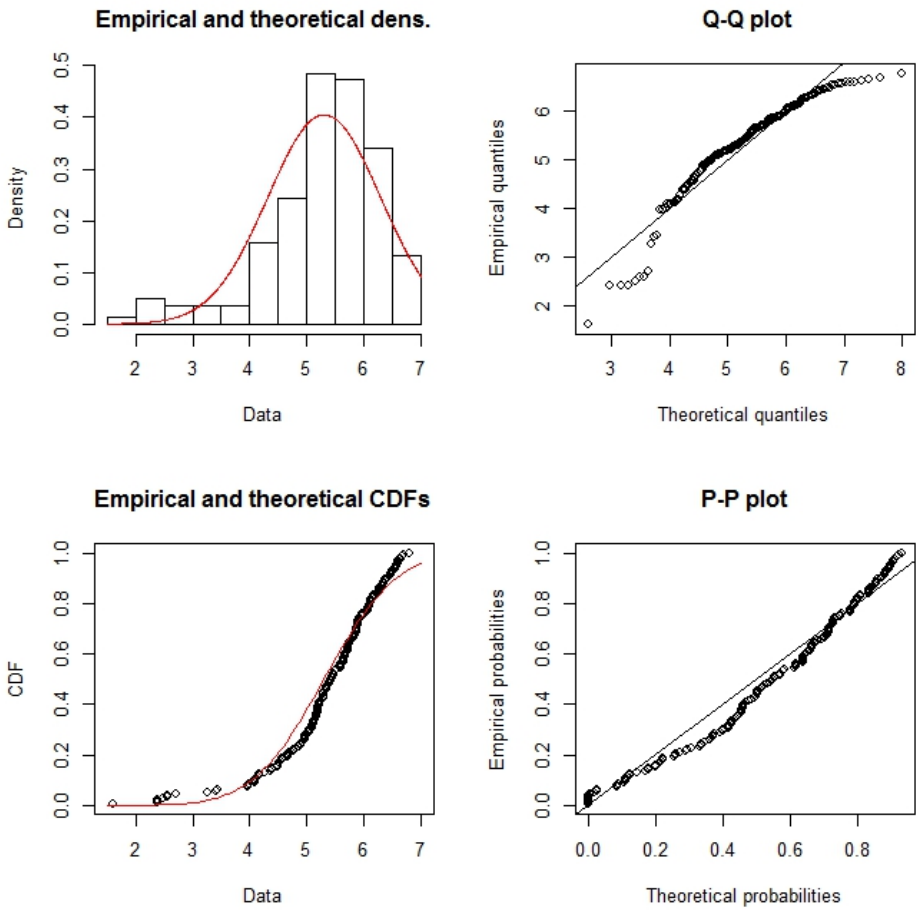

Figure 1: *PDF, QQ plot, PP plot and the CDF of empirical data compared to a log-normal distribution*

Among the distributions considered, Weibull and log-logistic distributions offered a good fit to the data. In order to assign the appropriate distribution (among Weibull and log-logistic distributions) for the survival times, the hazard functions were examined[64], stratified by performance status. Figure 5 shows the hazard functions associated with lung cancer over period of time (in days),

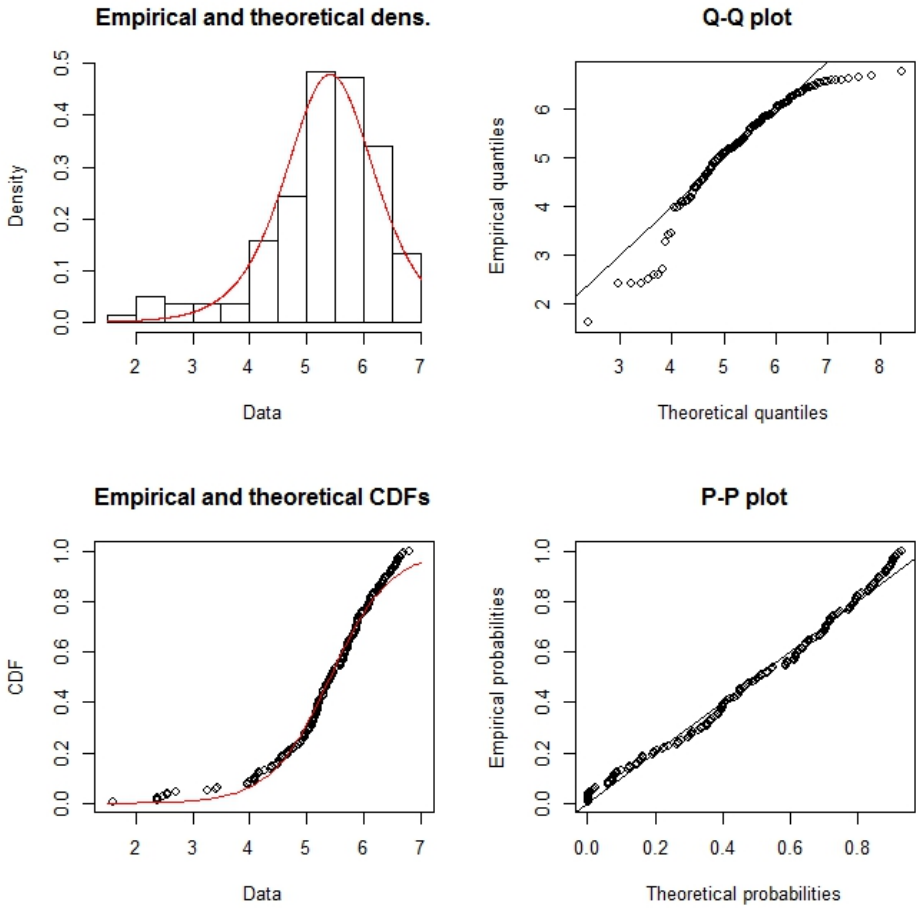

Figure 2: *PDF, QQ plot, PP plot and the CDF of empirical data compared to a log-logistic distribution*

considering performance status (low and high) as the covariate of interest. It can be seen clearly that the hazard function associated with lung cancer does not follow a monotonic distribution and hence is not an appropriate fit for the Weibull distribution. However, this non-monotonous hazard pattern provides an appropriate fit for the log-logistic distribution[23]. Hence, in this dissertation, further analysis is conducted to assess the performance and robustness of the MC-SIMEX procedure, assuming log-logistic distribution of lung cancer data.

5.3 Analysis

We categorize the physician Karnofsky performance score and the patient Karnofsky performance

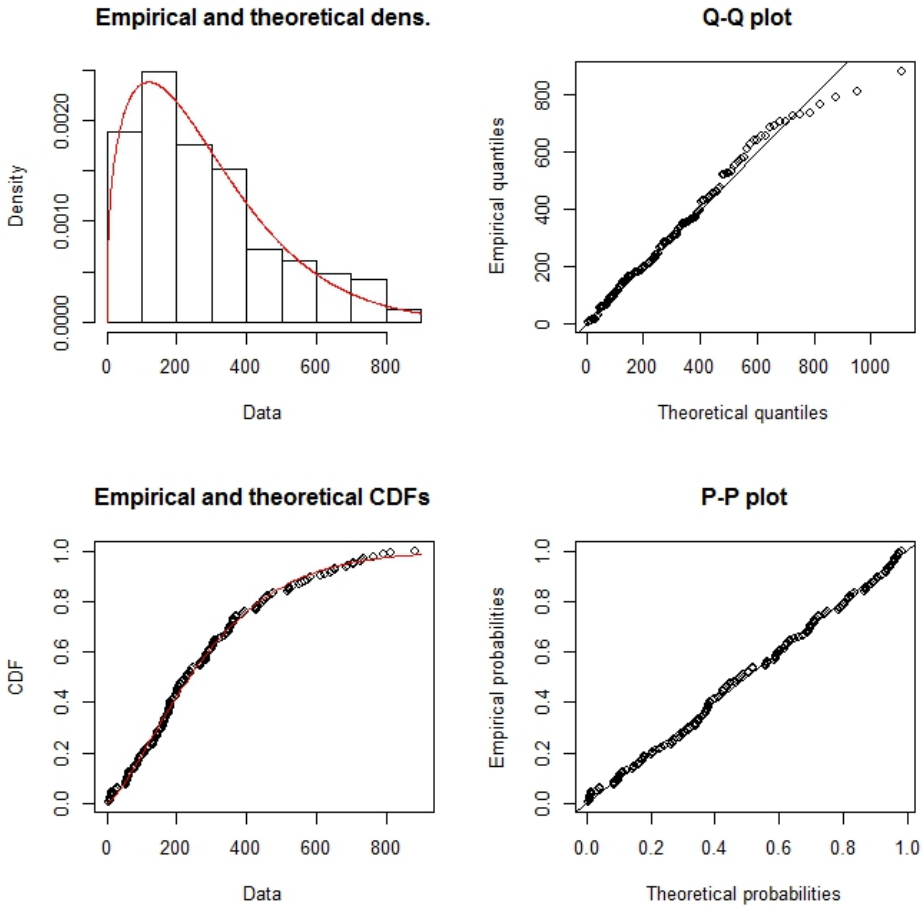

Figure 3: *PDF, QQ plot, PP plot and the CDF of empirical data compared to a Weibull distribution*

score into two categories: low performance category which includes scores of 70 or below and a high performance category which includes scores of greater than 70[23]. We treat the physician's assigned category as the true binary covariate X and the patient's self-assigned category as the naive binary covariate W.

5.4 Results

A total of 228 patients participated in this study. Among them, there were 138 males and 90 females. The mean age group among males was 63.34 years (standard deviation: 9.14, minimum=39 years, maximum=82 years) while the mean age group among females was 61.08 years (standard deviation 8.85, minimum=41 years, maximum = 77 years). Out of 228 patients, a total of

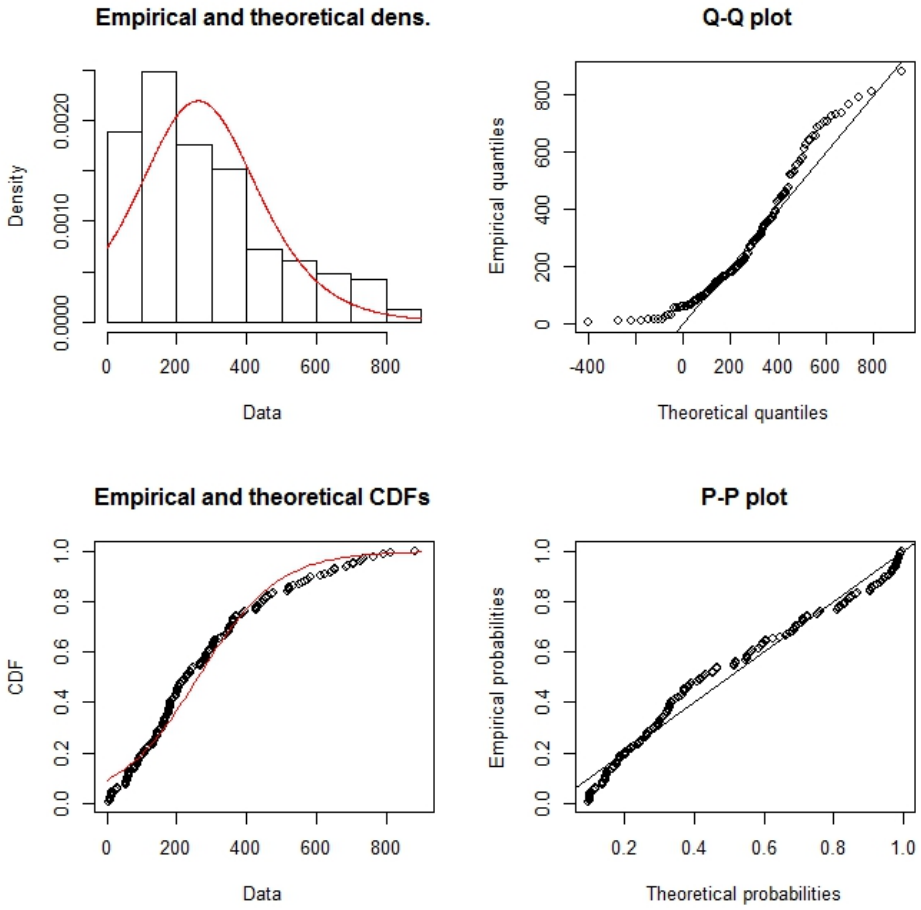

Figure 4: *PDF, QQ plot, PP plot and the CDF of empirical data compared to a logistic distribution*

63 patients were right censored while 165 patients were uncensored. The *proc lifetest*[65] function in SAS 9.2 was used to estimate the mean survival times through the Kaplan-Meir method[66]. The mean survival time among men was 321.12 days (299.52,342.72) and 439.26 days (410.86, 467.66) among women. Among men, the mean karnofsky performance score reported by the patients was 79.41 (standard deviation: 14.29, minimum: 30, maximum: 100) and the mean Karnofsky score reported among women was 80.79 (standard deviation: 15.17, minimum: 30, maximum: 100). Among men, the mean Karnosfsky performance score reported by physicians was 81.82 (standard deviation: 12.38, minimum: 50, maximum: 100) and the mean Karnofsky score reported among women was 82.11 (standard deviation: 12.32, minimum: 50, maximum: 100).

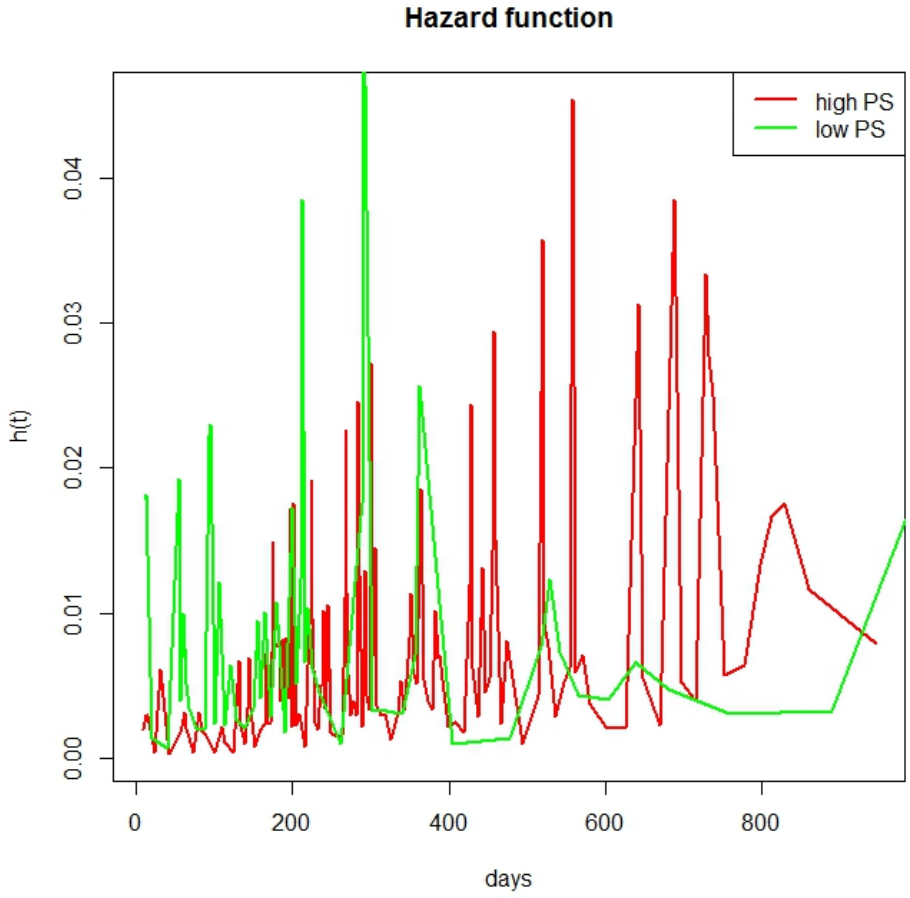

Figure 5: *Hazard functions associated with the lung cancer data considering the performance score category (PS) as the covariate.*

The results presented here are those of an AFT model, where the survival times follow log-logistic distribution. The model considered here adjusts for the performance score category (PS) and age.

$$\log(T_i) = \beta_{0i} + \beta_{1i}PS + \beta_{2i}age + \epsilon_i$$

where T_i is the survival time of an individual and ϵ_i is the error term which follows log-logistic distribution.

The coefficient $\hat{\beta}_{1simex}$ obtained using the MC-SIMEX estimator was further from zero than the naive estimator (0.74 and 0.48 respectively), as would be expected because of

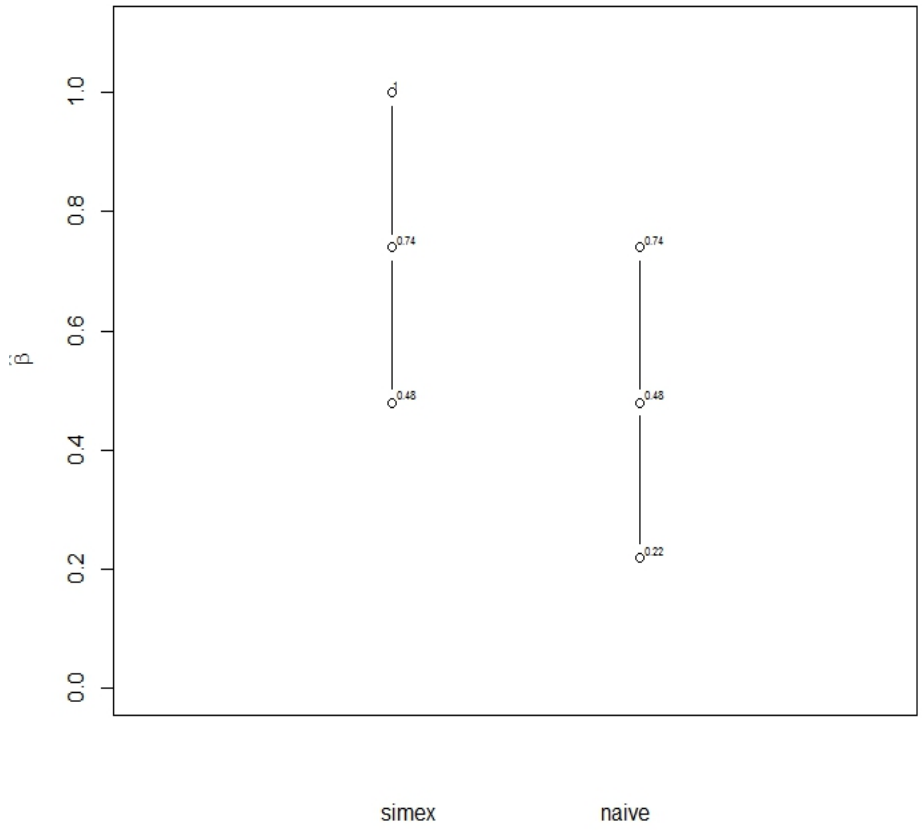

Figure 6: *Plot of the simex and naive estimators with their 95% confidence intervals. The x-axis represents the type of estimator and the y-axis represents the* $\hat{\beta}$ *values*

attenuation[22]. The 95% confidence intervals for the MC-SIMEX and the naive estimators were (0.48,1) and (0.22,0.74) respectively. The plots of 95% confidence intervals for the naive and simex estimators are shown in figure 6. As evident from the plot, there is an overlap of confidence intervals of the MC-SIMEX and the naive estimator, which shows that the two estimators are not significantly different from each other. The results can be interpreted as follows: The simex estimate indicates that after adjusting for age, a lung cancer patient in the high performance category survives 115% ($\exp^{0.77} - 1$) more (in days) than a patient in the low performance category. The naive estimate indicates that a lung cancer patient in the high performance category survives 61% ($\exp^{0.48} - 1$) more (in days) than a patient in the low performance category.

Chapter 6

CONCLUSION

Despite the vast amount of literature that existed with regard to misclassification error, we found that a few areas still remained unexplored. One of these areas is the study of misclassification in cancer survival data analysis. In this dissertation, we aim to fill this gap by studying the effect of misclassification in survival data, where the survival times follow a log-logistic distribution. Log-logistic distribution is the most common distribution encountered when dealing with lung and breast cancer data[23]. We surmise that one of the reasons for this gap in literature is a lack of a suitable mechanism within the inbuilt *mcsimex* function, to deal with survival data.

Even though MC-SIMEX procedure was put forth by Kuchenhoff et al.[47] under the assumption that the true sensitivity and specificity of misclassification are known, it is shown in this dissertation that the estimated misclassification matrix can provide a reasonable approximation in situations where the true sensitivity and specificity are not known. This is evident from the robustness analysis in Chapter 4. Also, the distribution of real data can be easily mis-specified in statistical analyses. Taking this issue into consideration, this dissertation sheds light on the robustness of the MC-SIMEX estimator, when a log-logistic distribution of data is mis-specified as a Weibull distribution (Chapter 4). The examination of robustness under the two above mentioned scenarios, offers a fresh perspective, since such an analysis of robustness has not been done before.

This dissertation provides a conduit for the application of MC-SIMEX procedure in survival analysis, considering that the existing *mcsimex* function is only amenable for generalized linear models and not for survival analysis. This dissertation also relaxes the assumption of availability of the true misclassification matrix. A new way to generate survival times and censoring rates, which is less time consuming and of lower computational burden, is also provided. This dissertation also helps in understanding the behavior of parameters with varying degrees of censoring.

This dissertation has certain limitations. First, it focuses on a simple AFT model with a confounding variable and a misclassified binary variable. However, in realistic situations, such simple statistical analysis may not suffice. Secondly, the quadratic function that is routinely used for extrapolation only provides an approximate fit to the exact extrapolant function, thereby resulting in some loss of accuracy. Third, the $\hat{\beta}_{simex}$ estimate is consistent to the true estimator only when the exact underlying distribution is known, which may not always be true. Finally, this dissertation assumes non-differential measurement error with homoscedasticity of residuals. However, this does not always hold true.[27]

This dissertation opens the door for more future research in the field of public health, pertaining to biostatistics. First, the behavior of coefficients of the confounding variable with varying degrees of misclassification of the binary covariate can be further explored for a log-logisitc distribution. Second, the efficiency of the MC-SIMEX procedure can be further evaluated in models where more than one binary variable is subject to misclassification error. Third, further analysis can be done in situations where there is differential measurement error, when considering a log-logistic distribution of survival times. Finally, the MC-SIMEX procedure can be extended to other fields of biostatistics such as mediation analysis.

REFERENCES

[1] X. Liu, *Survival analysis: models and applications*. John Wiley & Sons, 2012.

[2] K. Ahrens, T. L. Lash, C. Louik, A. A. Mitchell, and M. M. Werler, "Correcting for exposure misclassification using survival analysis with a time-varying exposure," *Annals of epidemiology*, vol. 22, no. 11, pp. 799–806, 2012.

[3] R. L. Prentice, "Covariate measurement errors and parameter estimation in a failure time regression model," *Biometrika*, vol. 69, no. 2, pp. 331–342, 1982.

[4] E. Marshall, "New a-bomb studies alter radiation estimates," *Science;(United States)*, vol. 212, 1981.

[5] B. Galobardes, J. W. Lynch, and G. D. Smith, "Childhood socioeconomic circumstances and cause-specific mortality in adulthood: systematic review and interpretation," *Epidemiologic reviews*, vol. 26, no. 1, pp. 7–21, 2004.

[6] L. Kauhanen, H.-M. Lakka, J. W. Lynch, and J. Kauhanen, "Social disadvantages in childhood and risk of all-cause death and cardiovascular disease in later life: a comparison of historical and retrospective childhood information," *International Journal of Epidemiology*, vol. 35, no. 4, pp. 962–968, 2006.

[7] W. Kannel, J. Neaton, D. f. Wentworth, H. Thomas, J. Stamler, S. Hulley, and M. Kjelsberg, "Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for the mrfit," *American heart journal*, vol. 112, no. 4, pp. 825–836, 1986.

[8] G. A. Satten and L. L. Kupper, "Inferences about exposure-disease associations using probability-of-exposure information," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 200–208, 1993.

[9] L.-J. Wei, "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis," *Statistics in medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992.

[10] S. H. Chiou, S. Kang, J. Yan, *et al.*, "Fitting accelerated failure time models in routine survival analysis with r package aftgee," *Journal of Statistical Software*, vol. 61, no. 11, pp. 1–23, 2014.

[11] D. R. Cox, "Regression models and life-tables," in *Breakthroughs in statistics*, pp. 527–541, Springer, 1992.

[12] S. R. Cole, H. Chu, and S. Greenland, "Multiple-imputation for measurement-error correction," *International journal of epidemiology*, vol. 35, no. 4, pp. 1074–1081, 2006.

[13] D. M. Zucker and D. Spiegelman, "Inference for the proportional hazards model with misclassified discrete-valued covariates," *Biometrics*, vol. 60, no. 2, pp. 324–334, 2004.

[14] D. M. Zucker, "A pseudo–partial likelihood method for semiparametric survival regression with covariate errors," *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1264–1277, 2005.

[15] D. M. Zucker and D. Spiegelman, "Corrected score estimation in the cox regression model with misclassified discrete covariates," in *Statistical Models and Methods for Biomedical and Technical Systems*, pp. 23–32, Springer, 2008.

[16] K. Akazawa, N. Kinukawa, and T. Nakamura, "A note on the corrected score function adjusting for misclassification," *Journal of the Japan Statistical Society*, vol. 28, no. 1, pp. 115–123, 1998.

[17] H. Bang, Y.-L. Chiu, J. S. Kaufman, M. D. Patel, G. Heiss, and K. M. Rose, "Bias correction methods for misclassified covariates in the cox model: comparison of five correction methods by simulation and data analysis," *Journal of statistical theory and practice*, vol. 7, no. 2, pp. 381–400, 2013.

[18] D. Spiegelman, R. J. Carroll, and V. Kipnis, "Efficient regression calibration for logistic

regression in main study/internal validation study designs with an imperfect reference instrument," *Statistics in medicine*, vol. 20, no. 1, pp. 139–160, 2001.

[19] H. Zhou and M. S. Pepe, "Auxiliary covariate data in failure time regression," *Biometrika*, vol. 82, no. 1, pp. 139–149, 1995.

[20] J. F. Lawless, *Statistical models and methods for lifetime data*, vol. 362. John Wiley & Sons, 2011.

[21] W. R. Swindell, "Accelerated failure time models provide a useful statistical framework for aging research," *Experimental gerontology*, vol. 44, no. 3, pp. 190–200, 2009.

[22] E. H. Slate and D. Bandyopadhyay, "An investigation of the mc-simex method with application to measurement error in periodontal outcomes," *Statistics in medicine*, vol. 28, no. 28, pp. 3523–3538, 2009.

[23] S. Bennett, "Log-logistic regression models for survival data," *Applied Statistics*, pp. 165–171, 1983.

[24] A. O. Langlands, S. J. Pocock, G. R. Kerr, and S. M. Gore, "Long-term survival of patients with breast cancer: a study of the curability of the disease.," *Br med J*, vol. 2, no. 6200, pp. 1247–1251, 1979.

[25] W. A. Fuller, *Measurement error models*, vol. 305. John Wiley & Sons, 2009.

[26] R. J. Carroll and L. A. Stefanski, "Approximate quasi-likelihood estimation in models with surrogate predictors," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 652–663, 1990.

[27] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

[28] B. Rosner, W. Willett, and D. Spiegelman, "Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error," *Statistics in medicine*, vol. 8, no. 9, pp. 1051–1069, 1989.

[29] L. Gleser, "Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models," *Contemp. Math*, vol. 112, pp. 99–114, 1990.

[30] B. Armstrong, "Measurement error in the generalised linear model," *Communications in Statistics-Simulation and Computation*, vol. 14, no. 3, pp. 529–544, 1985.

[31] I. Dalen, J. P. Buonaccorsi, P. Laake, A. Hjartåker, and M. Thoresen, "Regression analysis with categorized regression calibrated exposure: some interesting findings," *Emerging themes in epidemiology*, vol. 3, no. 1, p. 6, 2006.

[32] D. B. Rubin, "Inference and missing data," *Biometrika*, pp. 581–592, 1976.

[33] D. B. Rubin and R. J. Little, "Statistical analysis with missing data," *Hoboken, NJ: J Wiley & Sons*, 2002.

[34] R. J. Carroll, "Measurement error in epidemiologic studies," *Encyclopedia of biostatistics*, 1998.

[35] I. R. White, "Commentary: Dealing with measurement error: multiple imputation or regression calibration?," *International Journal of Epidemiology*, 2006.

[36] S. Van Buuren, H. C. Boshuizen, D. L. Knook, *et al.*, "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in medicine*, vol. 18, no. 6, pp. 681–694, 1999.

[37] L. Qi, Y.-F. Wang, and Y. He, "A comparison of multiple imputation and fully augmented weighted estimators for cox regression with missing covariates," *Statistics in medicine*, vol. 29, no. 25, pp. 2592–2604, 2010.

[38] T. Nakamura, "Proportional hazards model with covariates subject to measurement error," *Biometrics*, pp. 829–838, 1992.

[39] T. Augustin, "An exact corrected log-likelihood function for cox's proportional hazards model under measurement error and some extensions," *Scandinavian Journal of Statistics*, vol. 31, no. 1, pp. 43–50, 2004.

[40] H. Bang, "Medical cost analysis: application to colorectal cancer data from the seer medicare database," *Contemporary clinical trials*, vol. 26, no. 5, pp. 586–597, 2005.

[41] Y. Huang and C. Wang, "Consistent functional methods for logistic regression with errors in covariates," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1469–1482, 2001.

[42] M. S. Pepe, "Inference using surrogate outcome data and a validation sample," *Biometrika*, vol. 79, no. 2, pp. 355–365, 1992.

[43] M. S. Pepe and T. R. Fleming, "A nonparametric method for dealing with mismeasured covariate data," *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 108–113, 1991.

[44] J. R. Cook and L. A. Stefanski, "Simulation-extrapolation estimation in parametric measurement error models," *Journal of the American Statistical association*, vol. 89, no. 428, pp. 1314–1328, 1994.

[45] W. He, G. Y. Yi, and J. Xiong, "Accelerated failure time models with covariates subject to measurement error," *Statistics in Medicine*, vol. 26, no. 26, pp. 4817–4832, 2007.

[46] M. W. Knuiman, M. K. Bulsara, T. A. Welborn, M. S. Hobbs, *et al.*, "Mortality trends, 1965 to 1989, in busselton, the site of repeated health surveys and interventions," *Australian journal of public health*, vol. 18, no. 2, pp. 129–136, 1994.

[47] H. Küchenhoff, S. M. Mwalili, and E. Lesaffre, "A general method for dealing with misclassification in regression: the misclassification simex," *Biometrics*, vol. 62, no. 1, pp. 85–96, 2006.

[48] D. Loomis, D. B. Richardson, and L. Elliott, "Poisson regression analysis of ungrouped data," *Occupational and environmental medicine*, vol. 62, no. 5, pp. 325–329, 2005.

[49] J. Lindsey, "Fitting parametric counting processes by using log-linear models," *Applied Statistics*, pp. 201–212, 1995.

[50] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, vol. 360. John Wiley & Sons, 2011.

[51] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[52] M. O. Ojo and A. Olapade, "On the generalized logistic and log-logistic distributions," *Kragujevac Journal of Mathematics*, vol. 25, pp. 65–73, 2003.

[53] W. Lederer and H. Küchenhoff, "A short introduction to the simex and mcsimex," *The Newsletter of the R Project Volume 6/4, October 2006*, p. 26, 2006.

[54] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.

[55] H. Küchenhoff, W. Lederer, and E. Lesaffre, "Asymptotic variance estimation for the misclassification simex," *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 6197–6211, 2007.

[56] T. Therneau, "A package for survival analysis in s. r package version 2.37-4," *See http://CRAN. R-project. org/package= survival*, 2014.

[57] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, "Package 'mass'," *Cran R*, 2013.

[58] W. Lederer, H. Küchenhoff, M. W. Lederer, and M. by Küchenhoff, "Package 'simex'," 2009.

[59] C. L. Loprinzi, J. A. Laurie, H. S. Wieand, J. E. Krook, P. J. Novotny, J. W. Kugler, J. Bartel, M. Law, M. Bateman, and N. E. Klatt, "Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group.," *Journal of Clinical Oncology*, vol. 12, no. 3, pp. 601–607, 1994.

[60] "NCCTG Lung Cancer Data." `https://vincentarelbundock.github.io/Rdatasets/datasets.html`. Accessed: 2017-01-29.

[61] V. Mor, L. Laliberte, J. N. Morris, and M. Wiemann, "The karnofsky performance status scale: an examination of its reliability and validity in a research setting," *Cancer*, vol. 53, no. 9, pp. 2002–2007, 1984.

[62] J. W. Yates, B. Chalmer, F. P. McKegney, *et al.*, "Evaluation of patients with advanced cancer using the karnofsky performance status," *Cancer*, vol. 45, no. 8, pp. 2220–2224, 1980.

[63] M. L. Delignette-Muller, C. Dutang, *et al.*, "fitdistrplus: An r package for fitting distributions," *Journal of Statistical Software*, vol. 64, no. 4, pp. 1–34, 2015.

[64] K. Hess and R. Gentleman, "muhaz: Hazard function estimation in survival analysis, 2010," *URL http://CRAN. R-project. org/package= muhaz. R package version*, vol. 1, no. 5, p. 114.

[65] A. B. Cantor, *Extending SAS survival analysis techniques for medical research*. SAS Institute, 1997.

[66] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.