2008

Assessing and Correcting the Effects of Measurement Error Among Correlated Covariates in a Proportional Hazards Setting

Tina Juliet Thandeka Dube
University of Alabama at Birmingham

ASSESSING AND CORRECTING THE EFFECTS OF MEASUREMENT ERROR AMONG CORRELATED COVARIATES IN A PROPORTIONAL HAZARDS SETTING

by

TINA JULIET THANDEKA DUBE

LESLIE A. MCCLURE, COMMITTEE CHAIR
KARLENE BALL
CHARLES KATHOLI
DAVID ROTH
JENIFER VOEKS

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2008

ASSESSING AND CORRECTING THE EFFECTS OF MEASUREMENT ERROR AMONG CORRELATED COVARIATES IN A PROPORTIONALHAZARDS SETTING

TINA JULIET THANDEKA DUBE

BIOSTATISTICS

ABSTRACT

Older drivers, when contrasted with their younger counterparts, tend to drive fewer miles annually, yet their motor vehicle crash rates adjusted for miles driven are higher than all groups except the youngest drivers. Identifying factors associated with the risk of automobile crashes among the elderly is vital. In many elder driver studies, error free covariates such as gender or age, and error prone covariates like annual mileage or cognitive performance, are used to study time to an at-fault motor vehicle crash. We performed a simulation study that mimics this situation, using both an error free covariate as well as a correlated mismeasured covariate with replicates, to explore the effects of measurement error while conducting proportional hazards modeling. The pair, correlation and measurement error are considered, since both affect parameter estimates. We used modified regression calibration and risk set calibration to account for and correct for the measurement error among correlated variables.

The focus of this research was consideration of approaches that correct for measurement error when using correlated covariates in Cox Regression modeling. Three methods, regression calibration, risk set calibration and maximum likelihood each adjusted for correlated covariates, were explored. More specifically, the former methods were used to simulate situations where two correlated covariates are used as prognostic factors to determine a relationship between time to an at-fault or fault unknown motor

vehicle crash. After construction of data conforming to the correlation/measurement error structure, the data were then analyzed via the aforementioned methods. The resulting parameter estimates were then assessed to determine the ability of each method to correct for the effects of correlation and measurement error. The amount of both absolute and relative bias, along with coverage, mean square error and relative efficiency was examined at each correlation/measurement error design point. As a result of this simulation study, it was concluded that regression calibration corrected for correlation method tended to outperform risk set calibration corrected for correlation in the estimation of the true parameter estimates.

DEDICATION

This dissertation is dedicated to my parents, Dr. Thomas and Ruth Dube, who taught me how to dream.

ACKNOWLEDGMENTS

I would like to thank my dissertation advisor, Dr. Leslie McClure who has been a tremendous help and support to me. Your constant encouragement through the dissertation process has been invaluable. Also, I would like to express my gratitude to my dissertation committee, Drs. Karlene Ball, Charles Katholi, David Roth and Jenifer Voeks. Thank you for your unfailing patience during the dissertation process. A special note of thanks goes to Dr. Charles Katholi for serving as a mentor throughout my doctoral studies. I have learned more from him than he will ever realize.

I would be remiss if I didn't thank two women of distinction, Candace Porter and Kiya Hamilton. Their friendship, love and compassion helped me to stay the course. I would also like to thank the members of Fullness Christian Fellowship. Your prayers and constant encouragement helped me to continue the race set before me.

Thanks to Ian and Allick Dube, my brother and uncle and Jolene Osei. Your support means more to me than you will ever realize. There are so many people that I owe a debt of thanks. I can not mention you all by name because I will miss some. To the remainder of my family, thank you for believing in me and always saying "keep pressing on, the finish line is in sight". To my mother Ruth Dube, thank you for your encouragement and words of life and to my father Thomas Dube thank you for your constant reminder to be patient. To Gumindenga Mabvuta, thank you for never doubting my ability. I love you all. You all made achievement possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABS(BIAS)	Absolute Bias
GRIMPS	Gross Impairment Screening Battery
ML	Maximum likelihood
MMVA	Maryland Motor Vehicle Administration
MSE	Mean Square Error
MVC	At-fault or fault unknown motor vehicle crash
MVPT	Motor-Free Visual Perception test
RCCORR	Regression Calibration Corrected for Correlation
RCNO	Regression Calibration not Corrected for Correlation
REL(BIAS)	Relative Bias
RSCADJ	Risk Set Calibration Adjusted for small risk sets
RSCCORR	Risk Set Calibration Corrected for Correlation
TRAILS	Trail-making test, Part B
UABMVS	University of Alabama at Birmingham Maryland Motor Vehicle Study
UFOV®	Useful Filed of View®, Subtest 2

CHAPTER 1

INTRODUCTION

The literature on elder driver crash risk is extensive, yet of this body of work, few studies describe risk factors associated with predicting time to an at-fault or fault unknown motor vehicle crash (MVC). The *UAB Maryland Motor Vehicle Study* (UABMVS), a prospective design, evaluates the likelihood of experiencing motor vehicle crashes among elder drivers. The objective of the study is to assess the relationship between time to MVC and a battery of performance based screening measures and demographic risk factors related to elder driver competence.

The null hypothesis is as follows:

H_o: There is no relationship between time to MVC and performance based cognitive measures related to elder driver competence.

For the UABMVS, to assess the hazard associated with specific risk factors, Cox proportional hazard survival models will be used. The Cox model is an important statistical tool as it allows for proportionality among any two sets of covariates when examining differences in hazards.[1, 2] When all the covariates are fixed at baseline, the hazard rate for two subjects with distinct values of one specific covariate is proportional. Survival time is based on time to event occurrence or time to censoring.[1] In the UABMVS, baseline risk factors are annual mileage, cognitive test results such as Trails Making A and B, Useful Field of View®, Motor Free Visual Perception Test and

demographic information; survival time is measured as the time to censoring or MVC. It is thought that a subset of the covariates may be measured with error, including annual mileage, Trails Making A and B, walk time and tap time (in seconds).

For the covariates measured precisely or without error, the planned analyses will produce accurate parameter estimates and tests.[3] However, analyses including those variables that are error contaminated or both error contaminated and correlated with other important risk factors will produce biased parameter estimates.[4] It is the latter situation which motivates the problem addressed in this dissertation.

Statement of the Problem

Measurement error problems arise frequently in statistical analyses of failure time data. Complexities in the data, which make analysis difficult, arise due to the data measurement mechanism, the environment, biological variability, data of questionable quality, laboratory analysis error or reliance on self report data. As a result, precise measurements of the variables of interest are not directly observable. Thus, these variables are measured with error. Further, in addition to containing considerable amounts of error, the variables may be correlated with error free variables. Statistical analyses based on data of this type may result in biased parameter estimates and subsequently inaccurate inferences for both the mismeasured variable and the variable free of error.

Some scenarios to consider include:

1. A single mismeasured, continuous covariate;
2. One error prone, continuous covariate, uncorrelated with others free of error; or

3. One mismeasured, continuous covariate, correlated with covariates that are free of error.

In each of scenarios 1 and 2, the parameter estimate for the mismeasured variable will be biased towards the null.[5] Yet in the third, the bias may follow any direction based on the amount of error and the strength and direction of the correlation, as well as the other covariates in the regression model.[6] Moreover, estimates for variables that are not measured with error may also be biased.[5]

The difficulties resulting from measurement error also affect inference based on biased estimates. There is a loss of statistical power to detect important differences and relationships among variables of interest. Furthermore, measurement error may mask features of the data, making graphical model analysis difficult.[7] This is compounded when correlation among variables of interest exists.

In this work, the effect of measurement error identified in scenario three is investigated. Specifically, the resulting bias when observing two correlated variables, one error free and the other measured with error, is evaluated and methods to correct for the resulting bias are explored. The choice of two continuous variables is common in social science research settings. In the UABMVS, the error prone covariate of interest is annual mileage, while either age or the resulting cognitive exam score represents the error free variable.

The Cox Proportional Hazards Regression Model

Assume that a random sample of n observations from the population of interest is observed. Throughout the text $\mathbf{X}$ represents the vector of true values of the error prone

covariate. **X** may be an exposure or confounder variable. The surrogate vector, **W**, represents the vector of observed values of the error prone variable, **X**. Vector **Z** contains predictors which are measured accurately and free of error. The predictors represented by **Z** may or may not be correlated with the exposure variables found in **X**.

Let (V_i, C_i), $i=1,...,n$, represent independent failure and censoring time random variables respectively. The survival time for the i^{th} subject, T_i, is the minimum of failure and censoring times, $T_i = min(V_i, C_i)$. The event indicator, δ_i, takes the value of 1 when failure time is less than or equal to the censoring time or $\delta_i = I(V_i \le C_i)$. Consequently, observed, right censored survival data for the i^{th} subject is written as $(T_i, \delta_i, \mathbf{X}_i, \mathbf{Z}_i)$ where T_i is the length of time on the study for the i^{th} subject, $(\mathbf{X}_i, \mathbf{Z}_i)$ is the observed covariate vector for the i^{th} subject.[1, 2] The survival time is related to the covariates of interest through the hazard function

$$\lambda(t \mid \mathbf{X}_i, \mathbf{Z}_i) = \lambda_0(t) \exp\left(\boldsymbol{\beta}_X^T \mathbf{X}_i + \boldsymbol{\beta}_Z^T \mathbf{Z}_i\right), \tag{1.1}$$

where $\left(\boldsymbol{\beta}_X^T, \boldsymbol{\beta}_Z^T\right)^T$ is a $p \times 1$ vector of regression coefficients. The vector of error prone covariates, **X**, is a $k \times n$ vector and **Z** is a p-$k \times n$ vector of the covariates that are accurately measured. In the hazard function, t denotes elapsed time, $\lambda_0(t) > 0$ is an unspecified baseline hazard function for continuous time t.[1, 8, 9]

Two components comprise the Cox model: an unspecified baseline hazard and parametrically modeled relationship between the survival time t and the covariates of interest. Because of this, the Cox model is deemed semiparametric. More specifically, a parametric form is assumed for only the covariate effect. Since the model is

semiparametric, special methods are used to conduct inference on parameters of interest. Partial likelihood methodology is used to determine the parameters of interest. [1, 10, 11]

Measurement Error Modeling

There are three major defining characteristics for measurement error models: the properties of the unobserved values, the structure of the error model, and the type of additional data available. In this section we will describe these characteristics in some detail.

Properties of Unobserved Data

The unobserved data **X** are modeled as either structural or functional. In structural modeling, a parametric distribution may be assigned to the data such as $\mathbf{X} \sim \mathbf{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\sigma}_x^2\right)$. Alternatively, in functional modeling, the $\boldsymbol{X_i}$, $i=1,\ldots,n$, is taken as either a fixed constant or random variable. No assumptions or minimal assumptions are made regarding $\boldsymbol{X_i}$.[12]

Models for the Measurement Error Process

The classical measurement error model can be expressed as **W**=**X**+**U**. The measurement error, **U**, has a mean of **0** and variance of σ_u^2. **W** is unbiased for **X**. It follows that the expectation of **W** given both **X** and covariates free of error, **Z**, is equal to **X**. Also, **U** is independent of **X**. The corresponding error structure of **U** can be either homoscedastic or heteroscedastic.

Data Sources

The data sources necessary for modeling the error structure can be divided into two categories: internal data and external data. An internal data set is a subset of the primary data. A data set is classified as external when the measurement error process is not assessed directly, but rather from independent studies. When external data sets are used, the transportability of the model must be assessed. Transportability is defined as the event when the model and the relevant parameter estimates can be transported from one model to another without bias.[13] It is often the case that the same classical error model can be assumed to hold across different studies, hence the model is transportable. Therefore, the parameters from one study can be transported to another. Typically, only a subset of the model parameters need be transportable in order to use the information in an analysis.[7]

Each of internal and external data can be further divided into three subgroups: validation data, instrumental data and replication data. A validation data set has the true measurement **X** observed alongside the error prone value in a subset of study data; whereas a data set in which the error prone measurement is made more than once in some or all of the subjects is called a replication data set. An instrumental data set is one that has, in addition to **W**, another observable variable, the instrumental variable, **T** which satisfies the following: **T** must be independent of (**W**-**X**); and **T** must be a surrogate for **X** and **T** must be correlated with **X**.[14]

Non-differential and Differential Error

Measurement error is non-differential when the model relating the dependent

variable given (**Z**,**X**,**W**) is the same as the model relating the dependent variable given (**Z**,**X**); otherwise, measurement error is differential. Non-differential error allows for estimation of parameters relating the response to the true predictor with only minimal additional information on the error distribution. **W** contains no information about the response other than what is available in **X**. In this event, **W** is called a surrogate for **X**. Thus, the true predictor necessarily need not be observed.

Differential error generally occurs in case-control studies. Disease or response status is obtained first, and then the exposures, as well as other covariates, are measured later. Also, differential measurement error may occur when the observed value **W** is a separate variable serving as a substitute for **X**. Analysis of models with differential error can be challenging. It is necessary to observe the true value of **X** on some subjects. Problems with differential error are generally analyzed using missing data techniques.[7]

Scope of Dissertation

Several methods are useful in assessing the effects of measurement error in variables both measured with and without error. Yet, failure to consider other measured, correlated variables is a major limitation of the current literature. It is this situation which will be addressed in this dissertation. Our goal is to assess the effect of measurement error and correlation among covariates and provide guidance on choosing the most appropriate methods in the proportional hazards setting, in light of the amount of error and correlation.

The remainder of Chapter 1 contains an overview of related literature. Chapter 2 provides a close examination of regression calibration and risk set calibration each

corrected for correlation, including similarities, differences and benefits of the approaches. Chapter 3 explores the maximum likelihood method. A detailed explanation of the research methodology and simulation study is contained chapter 4. In Chapter 5, the proposed methods of analysis are applied to real data. The performance of each method related to parameter estimation will be examined. Lastly, the dissertation concludes with a discussion of the findings, conclusions and suggestions for further research in Chapter 6.

Review of Related Literature

Various approaches have been proposed to handle measurement error in time to event settings. Generally, these approaches are classified as providing either approximate estimation or consistent estimation in the presence of measurement error. In the presence of mild levels of error, approximate methods are sufficient to handle the measurement error problem. Alternatively, moderate to high levels of measurement error are best handled by consistent methods.

The ensuing discussion is a description of the literature relevant to the problem proposed in this dissertation. While not exhaustive, its intent is to acquaint the reader with various methodologies related to fitting Cox regression models with mismeasured covariates. Several of the methods are applicable to the UABMVS data; however, none address our specific hypothesis of interest.

Approximate Methods

One generally applicable approach to account for measurement error in regression

models is regression calibration. Prentice (1982) extended the Cox proportional hazards model to handle measurement error.[15] From equation (1.1), Prentice wrote the induced hazard as

$$\lambda(t \mid \mathbf{X}) = \lambda_0(t) E\left[\exp\left(\boldsymbol{\beta}_{\mathbf{X}}^T \mathbf{X} \mid \mathbf{W}, \mathrm{T} \geq t\right)\right]. \tag{1.2}$$

Assuming independent censoring, the hazard function for $\boldsymbol{T}$ is independent of **W** given **X**, the distribution of **X** given **W** is normal, and that event occurrence is rare, the regression calibration procedure includes replacing **W** with the estimate $E\left(\mathbf{X} \mid \mathbf{W}\right)$ and then applying the standard analysis.[15] Partial likelihood methodology, used in the Cox Proportional Hazards setting, may still be applied to solve for parameter estimates since the partial likelihood does not depend on the unknown hazard function. Since $E\left(\mathbf{X} \mid \mathbf{W}\right)$, the calibrated function, is a function of the observed data **W**, it can be estimated once an error model for **X** is specified.

In 1995, the regression calibration method was extended to missing data.[16] Using discrete auxiliary data, the relationship between the missing covariate and the observed mismeasured covariate is estimated nonparametrically. Zhou and Wang continued along this line by including continuous covariates and using a kernel smoother method for estimating the induced hazard.[17] Regression calibration incorporating missing covariates was also studied by Wang. Here, the distribution of **X** given **Z** is estimated using validation data then used to impute missing data. [18]

Regression calibration is commonly used due to the simplicity of implementation. It reduces bias in the parameter estimates relative to the naïve approach. However, in the presence of large amounts of error, regression calibration may yield biased results.[19]

A modification to Prentice's regression calibration work was introduced by Clayton. While an assumption of Prentice's work is that event occurrence is rare, Clayton's method is applicable in the absence of a validation data set and with or without the rare event assumption. At each event time, Clayton suggests that regression calibration be completed within each risk set.[20]

Stefanski and Cook and then later Cook and Stefanski introduced and further developed a simulation based method of estimating and reducing bias due to measurement error.[21, 22] This simulation extrapolation approach is referred to as SIMEX. SIMEX is akin to regression calibration in that this method is easily implemented. SIMEX can be used with any error such as additive or multiplicative, that can be imitated using Monte Carlo methods.

Consistent Methods

Nakamura introduced the approximately corrected score function for parameter estimation in proportional hazards regression.[23, 24] Nakamura's approach assumes that the measurement errors are additive, independently, identically and normally distributed with known covariance. The approximately corrected score function is based on first and second order Taylor expansion. This process reduces bias as compared to naïve estimation because the corrected score function satisfies regularity conditions. The first order correction is consistent and asymptotically normal.[3, 25] The correction is not affected by whether the covariates are time dependent or not.

Similar to the work of Nakamura, Buzas also provided a class of corrected score estimators where the measurement error is additive.[4] Different from Nakamura, Buzas

represented the error distribution by including the existence of a known moment generating function and considered uncorrelated mismeasured covariates and those free of error.[4] Here, the scores change based on the moment generating function of the measurement error. Kong and Gu further extended the work of Nakamura by showing that as the sample size grows, the estimator resulting from the corrected score is consistent, and that the covariance structure may be estimated using a sandwich estimator.[25] Hu and Lin also extended the work of Nakamura and Buzas by using validation data to estimate the error distribution. Each of the measurement error and the covariate distributions remain unspecified. This approach can be extended to more than one mismeasured predictor.[26]

Likelihood (ML) approaches provide another class of consistent estimators. They differ from corrected score or regression calibration methods in that stronger distributional assumptions are required. Also, the baseline hazard does appear in the likelihood function while it does not appear in the partial likelihood function which all of the other methods described here utilize. In ML, the baseline hazard at each event time, as well as the regression coefficients are estimated. Using normal additive error, Hu and colleagues examined three approaches to single parameter estimation using maximum likelihood: fully parametric, fully nonparametric and semiparametric.[27] The fully parametric method requires a specific parametric distribution be placed on the error prone covariate; while the semiparametric approach allows for a milder assumption, such as that the covariate has a density. Alternatively, the nonparametric approach allows for using a class of discrete distributions to describe the covariate. While ML provides consist estimates, these methods can be computationally intensive.

Additional related approaches include Lin and Ying, Paik and Tsai, and Chen and Little who introduce consistent approaches to estimate regression parameters in the presence of missing covariates.[28-30] Lin and Wei (1989) explore the effects of misspecification of the hazard function.[31]

CHAPTER 2

REGRESSION CALIBRATION (RCCORR) AND RISK SET CALIBRATION (RSCCORR) CORRECTED FOR CORRELATION

In this chapter, we will discuss regression calibration and risk set calibration each corrected for correlation. Additionally, the similarities, strengths and weakness will be discussed.

Regression Calibration Corrected for Correlation (RCCORR)

The regression calibration algorithm is a replacement approach to compute an estimate for mismeasured observations by employing analysis of variance techniques. Unobserved **X** is replaced with a calibrated estimate and then the standard analyses are conducted. Regression calibration is applicable in a broad number of settings where replication or validation data is available.

When both **X** and **Z** are observed, the proportional hazards failure time regression model can be written as

$$\lambda\left(\mathbf{t};\mathbf{X},\mathbf{Z}\right)=\lambda_0\left(\mathbf{t}\right)\exp\left(\boldsymbol{\beta}_{\mathbf{X}}\mathbf{X}+\boldsymbol{\beta}_{\mathbf{Z}}\mathbf{Z}\right). \tag{2.1}$$

In the absence of error and ties among event times, the partial likelihood,

$$PL\left(\boldsymbol{\beta}_{\mathbf{X}},\boldsymbol{\beta}_{\mathbf{Z}}\right)=\prod_{i=1}^{n}\left[\frac{\exp(\beta_X X_i+\beta_Z Z_i)}{\sum_{l=1}^{n}Y_l\left(x_i\right)\exp(\beta_X X_l+\beta_Z Z_l)}\right]^{\delta_i}$$

is maximized to determine $(\boldsymbol{\beta}_{\mathbf{X}}, \boldsymbol{\beta}_{\mathbf{Z}})$; where $Y_l(x_i)$ is the risk set indicator and the event indicator is δ_i. $Y_l(x_i) = 1$ if $t_l > t_i$ where t_l indicates the survival time for the *l*th subject and $\delta_i = 1$ if subject *i* experiences an event.[10] A consistent estimator $(\hat{\boldsymbol{\beta}}_{\mathbf{X}}, \hat{\boldsymbol{\beta}}_{\mathbf{Z}})$ of $(\boldsymbol{\beta}_{\mathbf{X}}, \boldsymbol{\beta}_{\mathbf{Z}})$ can be obtained by solving the partial likelihood score function $\mathbf{U}\begin{pmatrix}\boldsymbol{\beta}_{\mathbf{X}}\\ \boldsymbol{\beta}_{\mathbf{Z}}\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}$, where

$$\mathbf{U}(\beta_X) = \sum_{i=1}^{n} \delta_i \left\{ X_i - \frac{\sum_{j=1}^{n} X_l Y_l(X_i) \exp(\beta_X X_l + \beta_Z Z_l)}{\sum_{l=1}^{n} Y_l(X_i) \exp(\beta_X X_l + \beta_Z Z_l)} \right\} \tag{2.2}$$

and

$$\mathbf{U}(\beta_Z) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} Z_l Y_l(X_i) \exp(\beta_X X_l + \beta_Z Z_l)}{\sum_{l=1}^{n} Y_l(X_i) \exp(\beta_X X_l + \beta_Z Z_l)} \right\}.$$

Suppose that rather than the true covariate **X**, **W** the surrogate for **X** is observed. The hazard model, (2.1), can be rewritten as,

$$\lambda(\mathbf{t}; \mathbf{W}, \mathbf{Z}) = \lambda_0(\mathbf{t}) \exp\left(\boldsymbol{\beta}_{\mathbf{X}^*} \mathbf{W} + \boldsymbol{\beta}_{\mathbf{Z}} \mathbf{Z}\right). \tag{2.3}$$

In this setting, it is impossible to directly estimate $(\boldsymbol{\beta}_{\mathbf{X}}, \boldsymbol{\beta}_{\mathbf{Z}})$, $\lambda_0(\mathbf{t})$ or to conduct inference about the effect of **X** based on these parameters. In order to alleviate the problem caused by observing a surrogate or measurement error in the true covariate, Prentice introduced the induced hazard function.[7, 15] The induced hazard function provides a setting under which inference on the hazard $\lambda(t; \mathbf{W}, \mathbf{Z})$ can lead to inference on $\lambda(t; \mathbf{X}, \mathbf{Z})$.

In order for inference based on $\lambda(t;\mathbf{W},\mathbf{Z})$ to lead to inference on $\lambda(t;\mathbf{X},\mathbf{Z})$, nondifferentiability must be assumed. Nondifferentiability indicates that the observed covariate **W**, has no predictive value given the true covariate **X**. As a consequence of nondifferentiability

$$\lambda(t;\mathbf{W},\mathbf{X},\mathbf{Z}) \approx \lambda(t;\mathbf{X},\mathbf{Z}). \tag{2.4}$$

It follows that $\lambda(t;\mathbf{W},\mathbf{Z})$ can be written as the conditional expectation of the distribution of $\lambda(\mathbf{t};\mathbf{W},\mathbf{X},\mathbf{Z})$ given $\mathbf{T}\geq\mathbf{t},\mathbf{W},\mathbf{Z};$ that is

$$\lambda(t;\mathbf{W},\mathbf{Z}) = E\left(\lambda(t;\mathbf{W},\mathbf{X},\mathbf{Z}) \mid T\geq t,\mathbf{W},\mathbf{Z}\right). \tag{2.5}$$

Using (2.4), equation (2.5) can be rewritten as

$$\lambda(t;\mathbf{W},\mathbf{Z}) = E\left(\lambda(t;\mathbf{X},\mathbf{Z}) \mid T\geq t,\mathbf{W},\mathbf{Z}\right). \tag{2.6}$$

Combining (2.1) and (2.6) results in

$$\begin{aligned}\lambda(\mathbf{t};\mathbf{W},\mathbf{Z}) &\cong E\left(\left\{\lambda_o(\mathbf{t})\left(\exp\left(\boldsymbol{\beta}_{\mathbf{X}}\mathbf{X}+\boldsymbol{\beta}_{\mathbf{Z}}\mathbf{Z}\right)\right)\right\} \mid \mathbf{T}\geq\mathbf{t},\mathbf{W},\mathbf{Z}\right)\\ &= \lambda_o(\mathbf{t})\exp\left(\boldsymbol{\beta}_{\mathbf{Z}}\mathbf{Z}\right)E\left\{\exp\left(\boldsymbol{\beta}_{\mathbf{X}}\mathbf{X} \mid \mathbf{T}\geq\mathbf{t},\mathbf{W},\mathbf{Z}\right)\right\}.\end{aligned} \tag{2.7}$$

In the absence of measurement error, when solving for $(\boldsymbol{\beta}_{\mathbf{X}},\boldsymbol{\beta}_{\mathbf{Z}})$ using partial likelihood methodology, the partial likelihood factors into a product of the baseline hazard and the rest of the likelihood. In (2.7), $E\left\{\exp\left(\boldsymbol{\beta}_{\mathbf{X}}\mathbf{X} \mid \mathbf{T}\geq\mathbf{t},\mathbf{W},\mathbf{Z}\right)\right\}$ has some dependence on the baseline hazard due to conditioning on $T\geq t$. Assumptions must be specified before partial likelihood methodology can be applied.

Prentice suggested that the rare event assumption must follow in order to conduct inference on (2.7). Suppose the rare event assumption holds. This indicates that most

subjects will survive beyond time *t*, or $P(T \geq t) \approx 1$.[15] Applying the rare event assumption to (2.7) results in the following:

$$\lambda(\mathbf{t};\mathbf{W},\mathbf{Z}) \cong \lambda_o(\mathbf{t})\exp(\boldsymbol{\beta}_\mathbf{Z}\mathbf{Z})E\{\exp(\boldsymbol{\beta}_\mathbf{X}\mathbf{X} \mid \mathbf{W},\mathbf{Z})\}. \tag{2.8}$$

The induced hazard belongs to the class of proportional hazards models since the ratio of hazards,

$$\frac{\lambda(\mathbf{t};\mathbf{W}_1,\mathbf{Z}_1)}{\lambda(\mathbf{t};\mathbf{W}_2,\mathbf{Z}_2)},$$

for two subjects with regression vectors $(\mathbf{W}_1,\mathbf{Z}_1)$ and $(\mathbf{W}_2,\mathbf{Z}_2)$ does not vary with time.[1, 10, 11]

Suppose that $(\mathbf{X} \mid \mathbf{W},\mathbf{Z})$ is normal with specified mean and constant variance. Under this normality assumption, regression calibration can be applied. Assume there is a single *X* and a single Z. Let $\mathbf{Q}^T = (X, Z, U_1, ..., U_k)^T$ be a $(k+2)\times 1$ vector. Also, let the multivariate normal distribution of Q be described as

$$\begin{bmatrix} X \\ Z \\ U_1 \\ \vdots \\ U_k \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_X \\ \mu_Z \\ \mu_U \\ \vdots \\ \mu_U \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XZ} & 0_K^T \\ \sigma_{XZ} & \Sigma_Z & 0_K^T \\ 0_K & 0_K & \sigma_U^2 I_k \end{bmatrix}\right). \tag{2.9}$$

Under the classical error model, $W_{ij} = X_i + U_{ij}$, where $i = 1, ..., n$ and $j = 1, ..., k$; k is the number of replicates for subject *i*. The U_{ij} are independent identically distributed. Each U_{ij} has mean 0 and variance σ_U^2. The variance, σ_U^2, is estimated using the available data. The vectors (X, Z) and **U** are independent. Since $W_{ij} = X_i + U_{ij}$,

it follows that

$$(2.10)\qquad \begin{bmatrix} X_i \\ Z_i \\ W_{i1} \\ \vdots \\ W_{ik} \end{bmatrix} \sim MVN\left[\begin{pmatrix} \mu_X \\ \mu_Z \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_2^T \\ \Sigma_2 & \Sigma_* \end{pmatrix}\right],$$

where

$$\Sigma_1 = \begin{pmatrix} \sigma_X^2 & \sigma_{XZ} \\ \sigma_{XZ} & \sigma_Z^2 \end{pmatrix}_{2\times 2},\ \Sigma_2 = \begin{pmatrix} \sigma_X^2 & \sigma_{XZ} \\ \vdots & \vdots \\ \sigma_X^2 & \sigma_{XZ} \end{pmatrix}_{k\times 2} \text{ and } \Sigma_* = \begin{pmatrix} \sigma_W^2 & \sigma_X^2 & \cdots & \cdots & \sigma_X^2 \\ \sigma_X^2 & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \sigma_X^2 \\ \sigma_X^2 & \cdots & \cdots & \sigma_X^2 & \sigma_W^2 \end{pmatrix}_{k\times k}.$$

Since $\left(X \mid W, Z\right)$ follows a multi-variable normal distribution, $\left(X \mid \bar{W}, Z\right)$ does too.

$E\left\{\exp\left(\beta_X\left(X \mid \bar{W}, Z\right)\right)\right\}$ is the moment generating function for $\left(X \mid \bar{W}, Z\right)$. Therefore

$E\left\{\exp\left(\beta_X\left(X \mid \bar{W}, Z\right)\right)\right\}$ can be rewritten as

$$(2.11)\qquad \exp\left(\frac{\beta_X \Sigma_{X\mid \bar{W},Z} \beta_X^T}{2}\right)\exp\left(\beta_X E\left(X \mid \bar{W}, Z\right)\right),$$

resulting in

$$(2.12)\qquad \begin{aligned} \lambda\left(t; W, Z\right) &\cong \lambda_o\left(t\right)\exp\left(\beta_Z Z\right)\exp\left(\frac{\beta_X \Sigma_{X\mid \bar{W},Z} \beta_X^T}{2}\right)\exp\left(\beta_X E\left(X \mid \bar{W}, Z\right)\right) \\ &= \lambda_0^*\left(t\right)\exp\left(\beta_X E\left(X \mid \bar{W}, Z\right)\right). \end{aligned}$$

At this point, the regression calibration algorithm begins with replacing unobserved X_i with its calibrated value $E(X_i \mid \bar{W}_{i\cdot}, Z_i)$. The calibrated value is found by regressing X_i on $\bar{W}_{i\cdot}$ and Z_i. Correlation between X_i and Z_i is accounted for in the calibrated value of X_i, $\hat{X}_{i,RCCorr}$. It follows that the calibrated value, $\hat{X}_{i,RCCorr}$, is written as

$$(2.13)\ \hat{X}_{i,RCCorr} = E(X_i \mid \bar{W}_{i\cdot}, Z_i) = \bar{W}_{\cdot\cdot} + \begin{pmatrix} \hat{\sigma}_X^2 & \hat{\sigma}_{XZ} \end{pmatrix} \begin{pmatrix} \hat{\sigma}_X^2 + \dfrac{\hat{\sigma}_U^2}{k} & \hat{\sigma}_{XZ} \\ \hat{\sigma}_{XZ} & \hat{\sigma}_Z^2 \end{pmatrix}^{-1} \begin{pmatrix} \bar{W}_{i\cdot} - \bar{W}_{\cdot\cdot} \\ Z_i - \bar{Z}_{\cdot} \end{pmatrix};$$

where σ_U^2 is the error variance. The error variance is calculated as

$$(2.14)\qquad \hat{\sigma}_U^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(W_{ij} - \bar{W}_{i\cdot}\right)^2}{n(k-1)}.$$

Using observed replicate data, the estimated variance for *X* and *Z*, $\hat{\sigma}_X^2$, is calculated as follows:

$$(2.15)\qquad \hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n}\left(\bar{W}_{i\cdot} - \bar{W}_{\cdot\cdot}\right)^2}{(n-1)} - \frac{\hat{\sigma}_U^2}{k}$$

and

$$(2.16)\qquad \hat{\sigma}_Z^2 = \frac{\sum_{i=1}^{n}\left(Z_i - \bar{Z}_{\cdot}\right)^2}{(n-1)}.$$

The estimate for correlation is

$$(2.17)\qquad \hat{\sigma}_{XZ} = \frac{\sum_{i=1}^{n}\left(W_{i\cdot} - \bar{W}_{\cdot\cdot}\right)\left(Z_i - \bar{Z}_{\cdot}\right)^2}{(n-1)}$$

where $\bar{W}_{\cdot\cdot}$ is the grand mean or

$$(2.18)\qquad \bar{W}_{\cdot\cdot} = n^{-1}\sum_{i=1}^{n}\bar{W}_{i\cdot};$$

and

$$(2.19)\qquad \bar{Z}_{\cdot} = n^{-1}\sum_{i=1}^{n} Z_i$$

represents the overall mean of the error free covariate Z. After obtaining the calibrated value for each X_i from(2.13), partial likelihood methodology is used to obtain estimates for β_X and β_Z.

Advantages and Limitations of RCCORR

In terms of measurement error modeling, RCCORR has one as of its main advantages that it is easily implemented and applied in standard software. The algorithm for implementation of RCCORR can be used when there is validation or replication data to estimate the error variance σ_U^2.

There are some limitations to implementation of RCCORR. One limitation to the RCCORR algorithm occurs in the presence of large amounts of error and correlation. Parameter estimates produced in this setting tend to be more biased than the naïve approach. Also, this method can not be implemented in the absence of rare events. Despite the limitations, this method is a good first order correction for error in proportional hazards modeling when the classical error model is appropriate.

Risk Set Calibration Corrected for Correlation (RSCCORR)

In survival analysis, the risk set at time *t* is the set of all those subjects alive at time *t*.[1, 10, 11] Risk Set Calibration Corrected for Correlation is another replacement approach used to compute an estimate for mismeasured observations. Similar to RCCORR, RSCCORR also uses analysis of variance techniques to provide calibrated estimates for unobserved *X*. This method differs from RCCORR in that at *each event*

time, all subjects in the risk set at time *t* receive a calibrated value. The calibrated value then contributes to the calculation of the partial likelihood at time *t*.

Similar to the methods employed in the RCCORR algorithm, suppose that $\left(X \mid \overline{W}, Z\right)$ follows a normal distribution within each risk set. As with RCCORR, under the classical additive error model, the measurement error variance is calculated using (2.14). The calibrated value, $\hat{X}\left(t\right)_{i,RSC} = E\left(X_i \mid Z_i, \overline{W}_{i.}\right)_t$, is defined as follows:

$$(2.20)\quad E\left(X_i \mid \overline{W}_{i.}, Z_i\right)_t = \hat{\mu}_{t,W} + \begin{pmatrix} \hat{\sigma}^2_{t,X} & \hat{\sigma}_{t,XZ} \end{pmatrix} \begin{pmatrix} \hat{\sigma}^2_{t,X} + \dfrac{\hat{\sigma}^2_U}{k} & \hat{\sigma}_{t,XZ} \\ \hat{\sigma}_{t,XZ} & \hat{\sigma}^2_{t,Z} \end{pmatrix}^{-1} \begin{pmatrix} \overline{W}_{i.} - \hat{\mu}_{t,W} \\ Z_i - \hat{\mu}_{t,Z} \end{pmatrix}.$$

Let $Y_{t,i}$ be the risk set indicator. The risk set indicator is defined as $Y_{t,i} = I\left(t_i > t\right)$. When the survival time for subject *i* is greater than event time *t* $Y_{t,i} = 1$. The grand mean of observed replicates at time *t*, $\hat{\mu}_{t,W}$, can be written as

$$(2.21)\qquad \hat{\mu}_{t,W} = \frac{\sum_{i=1}^{n} Y_{t,i} \overline{W}_{i.}}{\sum_{i=1}^{n} Y_{t,i}}$$

and

$$(2.22)\qquad \hat{\mu}_{t,Z} = \frac{\sum_{i=1}^{n} Y_{t,i} Z_i}{\sum_{i=1}^{n} Y_{t,i}}$$

is the grand mean for the observed error free covariate at time *t*. The variance for *X* for those subjects in the risk set at time *t* is

$$(2.23)\qquad \hat{\sigma}^2_{t,X} = \frac{\sum_{i=1}^{n} Y_{t,i} \left(\overline{W}_{i.} - \hat{\mu}_{t,W}\right)^2}{\left(n_t - 1\right)} - \frac{\hat{\sigma}^2_U}{k}$$

and the covariance is as follows:

$$\hat{\sigma}_{t,XZ} = \frac{\sum_{i=1}^{n} Y_{t,i}\left(\bar{W}_{i\cdot} - \hat{\mu}_{t,W}\right)\left(Z_i - \hat{\mu}_{t,Z}\right)^2}{\left(n_t - 1\right)}. \tag{2.24}$$

The variance of Z for those at risk at time t, $\hat{\sigma}^2_{t,Z}$, is calculated using $\hat{\mu}_{t,Z}$ and observed Z_i.

$$\hat{\sigma}^2_{t,Z} = \frac{\sum_{i=1}^{n} Y_{t,i}\left(Z_i - \bar{Z}_{\cdot}\right)^2}{\left(n_t - 1\right)}. \tag{2.25}$$

Each of the means and covariance matrices are generated at each event time. This increases computational complexity. When the risk set is small, the variance estimates tend towards 0, which makes estimation of the calibrated value difficult.

While the distribution of $\left(X, \bar{W}, Z\right)$ is normal at t=0, the distribution of $\left(X, \bar{W}, Z\right)$ may not be normal within subsequent risk sets. It follows that $\hat{X}\left(t\right)_{i,RSC}$ is an approximation to $E(X_i \mid Z_i, \bar{W}_{i\cdot})_t$. Therefore, the resulting parameter estimate, $\hat{\beta}_{RSC}$ will be asymptotically biased but the estimate should still be an improvement over the naïve estimate $\hat{\beta}_{NVE}$.

Advantages and Limitations of RSCCORR

Similar to RCCORR, RSCCORR is advantageous in that no new statistical programming software is necessary. Also, this method is applicable in both univariate and multivariate covariate settings. As with RCCORR, this method may be implemented in the presence of replicate data or validation data. As with replicate data, validation data

may be used to estimate the error variance. A limitation to implementation of RSCCORR is that the calculation of means within each risk may be cumbersome. In the presence of large amounts of correlation and/or error, the estimation of the variance components can cause the calibrated estimate to become unstable, thus causing the parameter estimates produced in this setting to be unstable.[32] While there may be limitations to this method, and situations for which RCCORR outperforms this method, it should be a reasonable approach to adjust for correlation and measurement error in the proportional hazards setting when using the classical error model.

CHAPTER 3

LIKELIHOOD METHODOLOGY

The final approach under consideration used to analyze survival data with correlated covariates subject to exposure measurement error is maximum likelihood (ML) methodology. This chapter will discuss ML methodology in detail, including construction of the likelihood and the individual components of the likelihood function. The numerical integration method, Gauss-Hermite quadrature will be discussed and its use in determining parameters of interest. Finally, the potential strengths and weaknesses of this method will be explored.

Likelihood

In order to implement the ML approach, contributing pieces of the likelihood must be specified. Our observed data consists of variable (T, W, Z). We begin by specifying the distribution of (T, W, Z), denoted $f_{T,W,Z}$.

$$f_{T_i,W_i,Z_i} = \int f_{T_i,W_i,X_i,Z_i}(t_i, w_i, x_i, z_i, \phi)\, dx_i$$
$$= \int \left[f_{T_i|W_i,X_i,Z_i}(t_i \mid w_i, x_i, z_i, \phi) \times f_{W_i,X_i,Z_i}(w_i, x_i, z_i) \right] dx_i. \tag{3.1}$$

Interest lies in estimating ϕ, where ϕ represents the set of parameters were are interested in making inference on. Since *W* serves as a surrogate for *X*, it follows from nondifferentiability that (3.1) can be rewritten as

$$f_{T_i,W_i,Z_i} = \int \left[f_{T_i|X_i,Z_i}(t_i \mid x_i, z_i, \phi) \times f_{W_i,X_i,Z_i}(w_i, x_i, z_i) \right] dx_i \,. \tag{3.2}$$

Thus, (3.2), the i^{th} term of the likelihood, can be expressed as

$$(3.3)\quad L_i = \int_{-\infty}^{\infty} \left[f_{T_i|X_i,Z_i}\left(t_i \mid x_i, z_i, \phi\right) \times f_{W_i|X_i,Z_i}\left(w_i \mid x_i, z_i\right) \times f_{X_i|Z_i}\left(x_i \mid z_i\right) \times f_{Z_i}\left(z_i\right) \right] dx_i .$$

The observed data likelihood is given by

$$(3.4)\quad \prod_{i=1}^{n} f_{T_i,W_i,Z_i} = \prod_{i=1}^{n} L_i \ ,$$

where the underlying model of primary interest, T given X, Z, is denoted $f_{T|X,Z}$, the error model is denoted by $f_{W|X,Z}$ and the exposure model is described as $f_{X,Z}$.

Model of Primary Interest

In order to perform likelihood analysis, every component of the data must have a specified parametric model. To this end, we begin with specification of the model of primary interest. Suppose for i=1,…,n, V_i represents independent and identically distributed (i.i.d) failure times with corresponding density $f_V(t \mid \phi)$; where ϕ is a vector of parameters associated with the failure time distribution. The corresponding survival function is defined as

$$(3.5)\quad \begin{aligned} S\left(t \mid \phi\right) &= 1 - P\left(T \le t \mid \phi\right) \\ &= P\left(T > t \mid \phi\right); \end{aligned}$$

and the hazard function is defined as

$$(3.6)\quad \lambda\left(t \mid \phi\right) = \frac{f\left(t \mid \phi\right)}{S\left(t \mid \phi\right)}.$$

Due to censoring for reasons such as: loss to follow-up or study termination, the failure

times may not be observed. Therefore, let the possible i.i.d censoring times be denoted as $C_1,...,C_n$ with corresponding density, $g_C(t \mid \varphi)$. Likewise, the survival and hazard functions for $g_C(t \mid \varphi)$ are defined similarly to equations (3.5) and (3.6). Suppose that the censoring times are governed by parameter φ. Under the assumption of random censoring, T_i and C_i are independent. Consequently, the distribution of the primary model, $f_{T|X,Z}$, is derived using both failure and censored times.[11, 33, 34]

Let the survival time for the i^{th} subject, T_i, be the minimum of failure and censoring times, $T_i = min(V_i, C_i)$. The event indicator, δ_i, takes on the value of 1 when failure time is less than or equal to the censoring time, that is, $\delta_i = I(V_i \le C_i)$. Subjects who do not experience an event or do not fail during the study are right censored. Consequently, the observed right censored survival data point for the i^{th} subject is written as $(\delta_i, T_i, \mathbf{X}_i, \mathbf{Z}_i)$ where $(\mathbf{X}_i, \mathbf{Z}_i)$ is the observed covariate vector of interest.

Using the data, the probability density function for the model of primary interest may be derived.

$$P(t \le T_i < t+h, \delta = 1 \mid \phi, x_i, z_i, \varphi) = P(t \le V_i < t+h, C_i > V_i \mid \phi, x_i, z_i, \varphi) \\ = P(t \le V_i < t+h, C_i > t \mid \phi, x_i, z_i, \varphi). \tag{3.7}$$

Because the survival and censoring times are independent, (3.7) can be rewritten as

$$P(t \le V_i < t+h \mid \phi, x_i, z_i)\ P(C_i > t \mid x_i, z_i, \varphi). \tag{3.8}$$

$P(C_i > t \mid x_i, z_i, \varphi)$ is the survival function of C, which can be written as $H_{C_i}(t \mid x_i, z_i, \varphi)$. Applying the limit definition of derivative to (3.8) results in

$$\lim_{h\to 0}\frac{P\left(t\le V_i<t+h\mid\phi,x_i,z_i\right)P\left(C_i>t\mid x_i,z_i,\varphi\right)}{h} = f_{V_i}\left(t\mid\phi,x_i,z_i\right)H_{C_i}\left(t\mid x_i,z_i,\varphi\right). \tag{3.9}$$

Similarly, when δ=0 or censoring has occurred,

$$\begin{aligned} P\left(t\le T_i<t+h,\delta=0\mid\phi,x_i,z_i,\varphi\right) &= P\left(t\le C_i<t+h,V_i>C_i\mid\phi,x_i,z_i,\varphi\right)\\ &= P\left(t\le C_i<t+h,V_i>t\mid\phi,x_i,z_i,\varphi\right)\\ &= P\left(t\le C_i<t+h\mid\varphi\right)\ P\left(V_i>t\mid\phi,x_i,z_i\right). \end{aligned} \tag{3.10}$$

The survival function of V, $P\left(V_i>t\mid\phi,x_i,z_i\right)$, can be written as $S_{V_i}\left(t\mid\phi,x_i,z_i\right)$. It follows from again applying the limit definition of the derivative to (3.10) that

$$\lim_{h\to 0}\frac{P\left(t\le C_i<t+h\mid\varphi\right)\ \ P\left(V_i>t\mid\phi,x_i,z_i\right)}{h} = g_{C_i}\left(t\mid\varphi\right)S_{V_i}\left(t\mid\phi,x_i,z_i\right). \tag{3.11}$$

Combining (3.9) and (3.11) results in the density

$$\begin{aligned} f_{T_i}\left(t\mid\phi,x_i,z_i,\delta_i\right) = &\left[\ f_{V_i}\left(t\mid\phi,x_i,z_i\right)H_{C_i}\left(t\mid\varphi\right)\right]^{\delta_i}\\ &\times\left[\ g_{C_i}\left(t\mid\varphi\right)S_{V_i}\left(t\mid\phi,x_i,z_i\right)\right]^{1-\delta_i}. \end{aligned} \tag{3.12}$$

Using hazard function notation, (3.12) can be rewritten as

$$\begin{aligned} f_{T_i\mid X_i,Z_i}\left(t\mid\phi,x_i,z_i,\delta_i\right) = &\left\{\left[\frac{f_{V_i}(t\mid\phi,x_i,z_i)}{S_{V_i}\left(t\mid\phi,x_i,z_i\right)}\right]^{\delta_i} S_{V_i}\left(t\mid\phi,x_i,z_i\right)\right\}\\ &\times\left\{\left[\frac{H_{C_i}\left(t\mid\varphi\right)}{g_{C_i}\left(t\mid\varphi\right)}\right]^{1-\delta_i} g_{C_i}\left(t\mid\varphi\right)\right\}. \end{aligned} \tag{3.13}$$

The interest of survival modeling is to characterize the distribution of T using the parameter ϕ. Secondly, assume noninformative censoring. Under this assumption, the density $g_C(t\mid\varphi)$ is unrelated to ϕ or censoring times are independent of the failure times. Therefore, $f_{T_i\mid X_i,Z_i}\left(t\mid\phi,x_i,z_i\right)$ is proportional to the following:

(3.14) $$\left[\frac{f_{V_i}(t\mid\phi,x_i,z_i)}{S_{V_i}(t\mid\phi,x_i,z_i)}\right]^{\delta_i} S_{V_i}(t\mid\phi,x_i,z_i).$$

Using the definition of hazard function, (3.14) can be rewritten as

(3.15) $$\lambda(t\mid\phi,x_i,z_i)^{\delta_i}\, S_{V_i}(t\mid\phi,x_i,z_i).$$

Using (1.1),

$$\lambda(t\mid\phi,x_i,z_i)=\lambda_0(T_j)\left(\exp(\beta_X X_i+\beta_Z Z_i)\right).$$

Substituting this into (3.15) results in

$$f_{T_i|X_i,Z_i}=\left\{\lambda_0(T_j)\left(\exp(\beta_X X_i+\beta_Z Z_i)\right)\right\}^{\delta_i}\exp\left\{-\int_0^{T_i}\left[\lambda_0(u)\exp(\beta_X X_i+\beta_Z Z_i)\right]du\right\}.$$

Under the exponential hazard model, which suggests a constant hazard, $f_{T_i|X_i,Z_i}$ can be rewritten as

(3.16) $$f_{T_i|X_i,Z_i}=\left\{\lambda\left(\exp(\beta_X X_i+\beta_Z Z_i)\right)\right\}^{\delta_i}\exp\left\{-\int_0^{T_i}\left[\lambda\exp(\beta_X X_i+\beta_Z Z_i)\right]du\right\}.$$

Error Model

In measurement error modeling, it is assumed that the distribution of the error model is known or specified. Suppose that

(3.17) $$\begin{bmatrix}X\\Z\\U\end{bmatrix}\sim MVN\left[\begin{pmatrix}\mu_X\\\mu_Z\\\mu_U\end{pmatrix},\begin{pmatrix}\sigma_X^2 & \sigma_{XZ} & 0\\\sigma_{XZ} & \sigma_Z^2 & 0\\0 & 0 & \sigma_U^2\end{pmatrix}\right].$$

Under the classical error model, $W_i=X_i+U_i$, for $i=1,...,n.$ It follows that $f_{W_i,X_i,Z_i}(w_i,x_i,z_i)$ follows a multivariate normal distribution as follows:

(3.18) $$\begin{bmatrix} X \\ Z \\ W \end{bmatrix} \sim MVN\left[\begin{pmatrix} \mu_X \\ \mu_Z \\ \mu_W \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XZ} & \sigma_X^2 \\ \sigma_{XZ} & \sigma_Z^2 & \sigma_{XZ} \\ \sigma_X^2 & \sigma_{XZ} & \sigma_X^2 + \sigma_U^2 \end{pmatrix}\right].$$

Let $\sigma_W^2 = \sigma_X^2 + \sigma_U^2$ .Consequently, the error model, $f_{W_i|X_i,Z_i}\left(w_i \mid x_i, z_i\right)$, is normal with

(3.19) $$\mu_{W_i|X_i,Z_i} = \mu_w + \begin{bmatrix} \sigma_x^2 & \sigma_{xz} \end{bmatrix}\begin{bmatrix} \sigma_x^2 & \sigma_{xz} \\ \sigma_{xz} & \sigma_z^2 \end{bmatrix}^{-1}\begin{bmatrix} X_i - \mu_X \\ Z_i - \mu_Z \end{bmatrix}$$

and

(3.20) $$\sigma^2_{w_i|X_i,Z_i} = \sigma_w^2 - \begin{bmatrix} \sigma_x^2 & \sigma_{xz} \end{bmatrix}\begin{bmatrix} \sigma_x^2 & \sigma_{xz} \\ \sigma_{xz} & \sigma_z^2 \end{bmatrix}^{-1}\begin{bmatrix} \sigma_x^2 \\ \sigma_{xz} \end{bmatrix}.$$

For simplicity, we write the density of $f_{W_i|X_i,Z_i}\left(w_i \mid x_i, z_i\right)$ as follows:

(3.21) $$f_{W_i|X_i,Z_i} = \left(2\pi\sigma^2_{w_i|X_i,Z_i}\right)^{-1/2}\exp\left\{-\frac{\left(W_i - \mu_{w_i|X_i,Z_i}\right)}{\sqrt{2\sigma^2_{w_i|X_i,Z_i}}}\right\}^2.$$

Exposure Model

The distribution of the exposure model or the model containing the covariate information may be specified in terms of the parameters of interest. Suppose that

(3.22) $$\begin{bmatrix} X \\ Z \end{bmatrix} \sim BVN\left[\begin{pmatrix} \mu_X \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XZ} \\ \sigma_{XZ} & \sigma_Z^2 \end{pmatrix}\right].$$

While other distributions are plausible, for simplicity, suppose that the exposure model, $f_{X_i|Z_i}$, also follow a normal distribution, with unknown mean and variance. Suppose z_i is observed and X_i and z_i are correlated. The density function of $f_{X_i|Z_i}$ is given by

$$(3.23) \qquad f_{X_i|Z_i} = \left(2\pi\left(1-\rho^2\right)\sigma_X^2\right)^{-1/2} \exp\left\{-\frac{\left(X_i - \left(\mu_{X_i} + \rho\frac{\sigma_X^2}{\sigma_Z^2}\left(Z_i - \mu_{Z_i}\right)\right)\right)}{\sqrt{2\left(1-\rho^2\right)\sigma_X^2}}\right\}^2,$$

with unknown parameters $\left\{\sigma_X^2, \rho, \mu_X\right\}$.

Parameter Estimation Using Gauss-Hermite Quadrature

Using the parametric specification for the model of primary interest, (3.16), the error model, (3.21), and the exposure model, (3.23), the likelihood may be calculated. Since z_i is observed, the i^{th} term of the likelihood is

$$(3.24) \quad L_i = f_{Z_i}\left(z_i\right) \times \left\{\int_{-\infty}^{\infty}\left[f_{T_i|X_i,Z_i}\left(t_i \mid x_i, z_i, \phi\right) \times f_{W_i|X_i,Z_i}\left(w_i \mid x_i, z_i\right) \times f_{X_i|Z_i}\left(x_i \mid z_i\right)\right]\right\} dx_i$$

where the parameter space of the full likelihood is

$$\left\{\lambda, \beta_X, \beta_Z, \sigma_X^2, \rho, \mu_X\right\}.$$

The likelihood consists of the product of integrals where each integrand is the product of functions of the unknown, error prone covariate x_i. [35, 36] The full likelihood is defined over the real line, $(-\infty, \infty)$. Characteristically, the log-likelihood is maximized over the unknown parameters to solve for the parameters of interest.

The difficulty in solving the full likelihood is that it involves several integrals. Consequently, due to the complexity of the likelihood, numerical methods provide a better approach in parameter estimation. In order to estimate the parameters of interest, quadrature must be used. Quadrature is a form of numerical integration which allows for the determination of a numerical value of an integral. Gauss-Hermite quadrature may be used to evaluate the integral. It is often used because of its relation to Gaussian

densities.[35, 36] In order to perform numerical integration, the integral found in (3.24) must be rewritten in the following form

$$\int_{-\infty}^{\infty} g(y)\exp\left(-y^2\right)dy. \tag{3.25}$$

In general, a Gauss-Hermite quadrature method for the integration of

$$I = \int_a^b g(y)\exp\left(-y^2\right)dy \tag{3.26}$$

has weight function of the form

$$I \approx \sum_{i=1}^{m} w_i g(y_i). \tag{3.27}$$

Combining (3.26) and (3.27),

$$\int_a^b g(y)\exp\left(-y^2\right)dy \approx \sum_{i=1}^{m} w_i g(y_i),$$

where the pairs (w_i, y_i) depend on the kernel function $\exp\left(-y^2\right)$. The y_i are the zeros of the m^{th} order orthogonal polynomial with respect to the kernel function.

Rewriting L_i results in

$$\begin{aligned} L_i = q_i \quad & \times \int_{-\infty}^{\infty} f_{T_i|X_i,Z_i}(t_i \mid x_i, z_i, \phi) \times f_{W_i|X_i,Z_i}(w_i \mid x_i, z_i) \\ & \times \exp\left\{ -\frac{\left(X_i - \left(\mu_{X_i} + \rho\frac{\sigma_X^2}{\sigma_Z^2}\left(Z_i - \mu_{Z_i}\right)\right)\right)}{\sqrt{2\left(1-\rho^2\right)\sigma_X^2}} \right\}^2 dx_i, \end{aligned} \tag{3.28}$$

where $q_i = f_{Z_i}(z_i) \times \left(2\pi\left(1-\rho^2\right)\sigma_X^2\right)^{-1/2}$. To this end, let y_i, be a sampling node; then y_i and x_i are related by

(3.29) $$y_i = \left\{ \frac{\left(X_i - \left(\mu_{X_i} + \rho \frac{\sigma_X^2}{\sigma_Z^2} \left(Z_i - \mu_{Z_i} \right) \right) \right)}{\sqrt{2\left(1-\rho^2\right)\sigma_X^2}} \right\}.$$

It follows that

(3.30) $$x_i = \left(\mu_{X_i} + \rho \frac{\sigma_X^2}{\sigma_Z^2} \left(Z_i - \mu_{Z_i} \right) \right) + y_i \sqrt{2\left(1-\rho^2\right)\sigma_X^2}.$$

So

$$\partial x_i = \left(\sqrt{2\left(1-\rho^2\right)\sigma_X^2} \right) \partial y_i$$

and

$$\frac{\partial x_i}{\left(\sqrt{2\left(1-\rho^2\right)\sigma_X^2} \right)} = \partial y_i.$$

This change of variables reduces the integral from an infinite range to a finite range allowing for the expression to be put in a form suitable for Gauss-Hermite quadrature.[36] Consequently, (3.28) can be rewritten as

(3.31) $$\int_{-\infty}^{\infty} q_i g\left(\left(\mu_{X_i} + \rho \frac{\sigma_X^2}{\sigma_Z^2} \left(Z_i - \mu_{Z_i} \right) \right) + y_i \sqrt{2\left(1-\rho^2\right)\sigma_X^2}. \right) \frac{\exp(-y^2)}{\sqrt{\pi}} dy_i.$$

Under Gauss-Hermite quadrature,(3.31) is approximately equal to

(3.32) $$\frac{q_i}{\sqrt{\pi}} \sum_{i=1}^{m} w_i g\left(\left(\left(\mu_{X_i} + \rho \frac{\sigma_X^2}{\sigma_Z^2} \left(Z_i - \mu_{Z_i} \right) \right) + y_i \sqrt{2\left(1-\rho^2\right)\sigma_X^2} \right) \right);$$

where the weights are $q_i w_i \Big/ \sqrt{\pi}$. The nodes y_i and weights $q_i w_i \Big/ \sqrt{\pi}$ are uniquely determined by the likelihood space and function $g(y)$. Thus,

$$(3.33)\prod_{i=1}^{n} L_i = \sum\cdots\sum \frac{q_i}{\sqrt{\pi}} \sum_{i=1}^{m} w_i g\left(\left(\left(\mu_{X_i} + \rho\frac{\sigma_X^2}{\sigma_Z^2}\left(Z_i - \mu_{Z_i}\right)\right) + y_i\sqrt{2\left(1-\rho^2\right)\sigma_X^2}\right)\right).$$

Quadrature relies on a deterministic approximation to an integral as a weighted sum of the integrand evaluated at selected set of values. The approximation requires the integrand to be evaluated at each node and then the values are weighted and summed. The nodes and weights factors can be obtained from tabulations found in Abramowitz & Stegun (1964).[35-38] Using (3.33) to determine the parameters of interest, $\theta = \left\{\lambda, \beta_X, \beta_Z, \sigma_X^2, \rho, \mu_X\right\}$, the derivative of the loglikelihood is maximized with respect to θ_i.

Due to the computational difficulties inherent to this method, we will not explore its properties either in simulation or real data. However based on Hu, Tsiatis, Davidian (1998), we would expect that maximum likelihood estimation would provide the best performing parameter estimates.[26, 39] These resulting parameter estimates tend to have both smaller bias and variability.

CHAPTER 4

SIMULATION STUDY

In this chapter, a detailed explanation of the research methodology is presented. The objectives of the simulation study are explained followed by a description of the steps involved in the simulation procedure. Finally, measures used to compare the properties of the parameter estimates are presented. In addition, conclusions regarding the simulation results from the different methods are summarized.

Objectives of Simulation Study

The problem under investigation is the effect of correlation between continuous error prone and non-error prone covariates. More specifically, we are interested in comparing methods used to correct the effects of correlation between error free and error prone covariates in Cox regression settings. We will perform a simulation study to assess, compare and contrast the effectiveness of regression calibration and risk set calibration each corrected for correlation. Further, we will compare each of these methods to the Naïve method. The ability of each method to reduce bias and minimize variance will be assessed and we will discuss the settings in which each method may or not be preferred.

Simulation Details

To develop these methods and study the resultant parameter estimates achieved via ignoring correlation and measurement error, regression calibration without correcting

for correlation (RCNO), RCCORR and RSCCORR, we performed a simulation study. The simulation study sought to produce parameter estimates from each method and compare properties of the methods. As a result of computational difficulty, a third method emerged-risk set calibration adjusted for small risk sets (RSCADJ). Although three methods were compared, they each adhered to the same organizational structure. In this section, the specifications of the simulation study are discussed.

Sample Size and Number of Required Simulations

For this simulation study, each sample size was n=150. Determination of the number of simulations is based on consideration of precision and computing expense and parameter estimate accuracy.[40] Our primary interest is in the parameter estimate of the error prone correlated covariate. Suppose that κ represents the acceptable difference from the true estimate. This is also referred to as the level of accuracy. In our setting, the variance is assumed known. The number of simulations is based on the ability to detect a κ difference from the true value with power $(1-\beta)$. The following may be used to calculate the number of Monte Carlo data sets;

$$\tilde{N} = \left[\frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta)} \right)\sigma}{\kappa} \right]^2 .$$

In order to produce an estimate within 2% of the true estimate, 1000 simulations were necessary.

After running the RSCCORR method, approximately 6% of the simulations failed to converge. Therefore, the number of simulations was increased to 1060 to maintain the ability to produce an estimate within 2% of accuracy.

Simulating Covariate Data

Many sociological studies use multivariate models to model a subject's outcome of interest. In order to replicate that via simulation study, the vector $(X_i, Z_i, U_{i1}, \ldots, U_{ik})$ was generated using a multivariate normal distribution,

$$\begin{bmatrix} X_i \\ Z_i \\ U_{i1} \\ \vdots \\ U_{ik} \end{bmatrix} \sim MVN\left[\begin{pmatrix} \mu_X \\ \mu_Z \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XZ} & 0 & 0 & 0 \\ \sigma_{XZ} & \sigma_Z^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_U^2 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \sigma_U^2 \end{pmatrix}\right] \tag{3.34}$$

where U_{ij} is independent from Z_i and X_i; also, Z_i and X_i are correlated, $i = 1, ..., n.$ Assuming the additive error model holds, the observed value W_{ij} was generated by adding X_i and U_{ij}, $W_{ij} = X_i + U_{ij}.$ It follows that the transformed vector $(X_i, Z_i, W_{i1}, \ldots, W_{ik}),$ has multivariate normal distribution and can be written as

$$\begin{bmatrix} X_i \\ Z_i \\ W_{i1} \\ \vdots \\ W_{ik} \end{bmatrix} \sim MVN\left[\begin{pmatrix} \mu_X \\ \mu_Z \\ \mu_X \\ \vdots \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}\right], \tag{3.35}$$

where

$$\Sigma_{11} = \begin{bmatrix} \sigma_X^2 & \sigma_{XZ} \\ \sigma_{XZ} & \sigma_Z^2 \end{bmatrix}; \tag{3.36}$$

$$\Sigma_{12} = \begin{bmatrix} \sigma_x^2 & \sigma_x^2 & \cdots & \sigma_x^2 \\ \sigma_{XZ} & \sigma_{XZ} & \cdots & \sigma_{XZ} \end{bmatrix}; \tag{3.37}$$

and

$$\Sigma_{22} = \begin{bmatrix} \sigma_X^2 + \sigma_U^2 & \sigma_X^2 & \cdots & \sigma_X^2 \\ \sigma_X^2 & \ddots & \sigma_X^2 & \vdots \\ \vdots & \vdots & \ddots & \sigma_X^2 \\ \sigma_X^2 & \cdots & \sigma_X^2 & \sigma_X^2 + \sigma_U^2 \end{bmatrix}. \tag{3.38}$$

The dimensions of Σ_{11}, Σ_{12} and Σ_{22} are 2×2, $2 \times k$, and $k \times k$ respectively where k is the number of replicates. Common to measurement error modeling, the distribution of $f_{W|X,Z}(W_i \mid X_i, Z_i)$ is assumed to be known. Thus, data that correspond to these distributional constraints were generated. Using the description of small, moderate and large levels of correlation presented by Cohen, we modeled correlation between X and Z at three levels 0.1, 0.3 and 0.5, with error estimates of 0.1, 0.3, 0.5, 1.0 and 2.0.[41] The combination of correlation and error resulted in 15 design points, summarized in Table 1. The correlation/measurement error combinations prove to be salient too in the real data analysis. These levels of correlation and error are similar to what is presented in the real data analysis found in Chapter 6.

Simulating Censoring and Survival Times

In Chapter 3, the importance of random and noninformative censoring was elucidated. It follows that the censoring distribution along with the survival distribution must be specified.[10, 40, 42] For simplicity, the censoring distribution was set to follow an exponential distribution with lambda equal to one,

$$g_C(t) = e^{-t}.$$

Table 1 Design Points for Simulation Study

	Error				
Correlation	0.1	0.3	0.5	1.0	2.0
0.1	(0.1, 0.1)	(0.1, 0.3)	(0.1, 0.5)	(0.1, 1.0)	(0.1, 2.0)
0.3	(0.3, 0.1)	(0.3, 0.3)	(0.3, 0.5)	(0.3, 1.0)	(0.3, 2.0)
0.5	(0.5, 0.1)	(0.5, 0.3)	(0.5, 0.5)	(0.5, 1.0)	(0.5, 2.0)

*Table entries represent all combinations of correlation and measurement error

*Error estimates were generated using 3 replicates, $W_i = (W_{i1}, W_{i2}, W_{i3})$

* $\tilde{N}$ =1060 Monte Carlo Data sets

Let T have probability density function $f(t)$ and cumulative density function $F(t)$. Using the definitions from (3.5) and (3.6), the survival and hazard functions were defined as

(3.39) $$S(t) = 1 - F(t)$$

and

(3.40) $$\lambda(t) = \frac{f(t)}{S(t)}$$

respectively. The hazard function can be used to model survival times T. The cumulative hazard, $H(t)$, is defined as

(3.41) $$H(t) = \int_0^t \lambda(u)du.$$

It follows that the cumulative hazard found in (3.41) can be rewritten as a function of the survival function.

(3.42) $$\begin{aligned} H(t) &= \int_0^t \frac{f(u)}{S(u)}\, du \\ &= \int_0^t \frac{f(u)}{1 - F(u)}\, du \\ &= -\ln S(t). \end{aligned}$$

Exponentiation applied to (3.42) results in

(3.43) $$S(t) = \exp\left(-H(t)\right).$$

This definition may be applied to the Cox proportional hazards model,

$$\lambda(t; X, Z) = \lambda_0(t)\exp\left(\beta_X X + \beta_Z Z\right).$$

It follows for the Cox proportional hazards model, the survival function is given as

(3.44) $$S(t; X, Z) = \exp\left(-H_0(t)e^{\beta_X X + \beta_Z Z}\right).$$

where

$$H_0 = \int_0^t \lambda_0(u)du$$

is the cumulative baseline hazard.

Let U be a continuous random variable with probability density function $F(t; X, Z)$. It follows that U and 1-U both follow a uniform distribution on the interval (0,1).[43] Let T be the survival time to be modeled. Since $\lambda_0(t) > 0$, we may invert H_0 and represent the survival time T for the Cox model as

(3.45) $$T = H_0^{-1}\left\{-\exp\left(-\left(\beta_X X + \beta_Z Z\right)\right)\ln U\right\}.$$

The time on study for each subject is the minimum of the resulting survival time and censoring time. Commonly used survival time distributions are exponential, Weibull and Gompertz. These distributions meet the assumption of proportional hazards needed when utilizing the Cox model. For simplicity, we used the exponential distribution in this study. This distribution specifies a constant hazard, λ with corresponding cumulative hazard function,

(3.46) $$H_0(t) = \lambda t.$$

The inverse cumulative hazard function is

(3.47) $$H_0^{-1}(t) = \lambda^{-1} t.$$

Applying the inverse cumulative hazard function to (3.45) results in

(3.48) $$T = -\frac{\ln U}{\lambda \exp\left(\beta_X X + \beta_Z Z\right)}$$

and the corresponding hazard function is $\lambda(t; X, Z) = \lambda \exp(\beta_X X + \beta_Z Z)$. Survival times for each subject can be generated using equation (3.48). In this simulation, the survival times have a constant hazard of 1, with $\beta_X = 1$ and $\beta_Z = 1$.

Evaluating the Performance of Methods

In order to evaluate the performance of the methods, the average of the parameter estimates for each Monte Carlo run, along with their corresponding sample standard error and average standard error will be computed. Also, absolute and relative bias are useful to gauge performance of each method. Absolute bias indicates how well a method approximated the true value of the parameter while relative bias provides an indication of how large an error in estimation is made in comparison to the actual parameter value. The Monte Carlo mean is defined as

$$\bar{\hat{\beta}} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \hat{\beta}_i. \tag{3.49}$$

Absolute bias is calculated as

$$\text{ABS}(\text{BIAS}) = \left| \bar{\hat{\beta}} - \beta \right|; \tag{3.50}$$

and relative bias, which is reported as a percentage is calculated as

$$\text{REL}(\text{BIAS}) = \frac{\bar{\hat{\beta}} - \beta}{\beta} \times 100. \tag{3.51}$$

The mean square error (MSE) of an estimator is the expected squared deviation between the sample mean and the parameter to be estimated. MSE is very useful because it provides a summary of accuracy and precision of the estimator. It is calculated as follows:

$$\text{MSE} = \left(\bar{\hat{\beta}} - \beta\right)^2 + \frac{1}{\tilde{N}-1}\sum_{i=1}^{\tilde{N}}\left(\hat{\beta}_i - \beta\right)^2 = BIAS^2 + \text{var}\left(\hat{\beta}\right). \tag{3.52}$$

Relative efficiency (RE) of estimator is a measure of the variability of a new estimator to another. In settings where the estimators are not unbiased, the MSE is used to compute relative efficiency. The following is used to calculate relative efficiency,

$$\text{RE} = \frac{MSE\left(\hat{\beta}_{new}\right)}{MSE\left(\hat{\beta}_{Naive}\right)}. \tag{3.53}$$

If RE < 1, then the new method is more efficient than the older method.

Finally, we use coverage as a criterion for evaluating the estimators. The proportion of times that the confidence interval contains the true parameter value is defined as the coverage of the confidence interval. The nominal coverage rate should be approximately equal to the coverage rate. Coverage should fall within 2 standard errors of the nominal coverage probability which is calculated as follows:

$$\sqrt{p(1-p)/\tilde{N}}. \tag{3.54}$$

In our setting we use a 95% coverage rate, p=0.95. This indicates that the true value should be contained in approximately 993 to 1022 confidence intervals.[40]

Summary

Using each of the Monte Carlo trials, parameter estimates at each design point were computed using each of the five methods under study, as described in Chapters 1, 2 and 4, the censoring times followed an exponential distribution with mean and variance

of one. The hazard function corresponding to survival time followed an exponential distribution with the hazard,

$$\lambda(t; X, Z) = \exp(X + Z).$$

The distributions for the censoring and survival times were selected for simplicity.

The Monte Carlo mean and variance are calculated for the Naïve, regression calibration not adjusted for correlation, RCCORR and risk set calibration approaches. The criteria for evaluating and comparing parameter estimates are absolute and relative bias, relative efficiency and coverage. The results of the Monte Carlo simulation are shown in the next section.

Simulation Results

The method for conducting the simulation was explained in detail in earlier sections of Chapter 4. For simplicity, regression parameters were selected as $\beta_X = 1$ and $\beta_Z = 1$. The censoring rates ranged from 35% to 45% for each simulation run based on the minimum time on study determined from the minimum of the censoring and failure distributions. There were three replicates of the error prone covariate. The distribution of the measurement error was

$$\begin{bmatrix} U_{i1} \\ U_{i2} \\ U_{i3} \end{bmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_U^2 & 0 & 0 \\ 0 & \sigma_U^2 & 0 \\ 0 & 0 & \sigma_U^2 \end{pmatrix} \right], \tag{3.55}$$

where the variance ranged from 0.1 to 2.0. The results based on 1,060 samples of size 150 are shown in Tables 2 through 18. In order to evaluate the methods, the average of the parameter estimates and corresponding standard errors for each Monte Carlo run,

along with their corresponding standard deviations was computed. Additionally, 95% coverage, relative bias, absolute bias and relative efficiency were calculated.

Methods that Ignore Correlation and Measurement Error

As a baseline assessment of the impact of ignoring correlation and measurement, we first ran Monte Carlo simulations of the correlation-error combinations and estimated Naïve parameter estimates. The standard error is the square root of the sample variance of the resulting estimates and the average standard error is the average of the 1060 standard error estimates.

The results from the Naïve approach are found in Table 2 and 3. The Naïve estimators provide similar estimates within each level of error. As the correlation within each level of error increases, the estimates attenuate towards 0. This approach does a poor job of estimating the mean values of both of $\hat{\beta}_X$ and $\hat{\beta}_Z$. In the error prone estimate, the value of $\hat{\beta}_X$, varies from 0.47 to 0.95 depending on the magnitude of the correlation and measurement error. This range represents a great departure from the true value of 1. Correlation and measurement error do not impact the parameter estimates of the error free covariate as much as the error prone covariate. The values of the error free parameter estimates $\hat{\beta}_Z$, are less biased towards the null. The values vary from 0.92 to 1.10. Larger amounts of correlation and error result in attenuated parameter estimates in the error prone covariate. In the covariate that is not mismeasured, the Naive method is more robust. As expected, the Naïve approach is most robust in the setting of low correlation and error, 0.1 and 0.1 respectively.

Coverage probabilities, relative and absolute bias for the Naïve approach are all found in Table 3. The coverage probability for the error free estimate at small to moderate levels of error is greater than 95% which indicates that the naive method is somewhat robust at these levels of error. Alternatively, for most design points, the coverage for $\hat{\beta}_X$ is below the nominal level indicating that failing to correct for correlation and error will provide inaccurate estimates.

Within each level of error, the relative bias indicates that the error prone covariate is underestimated as correlation increases. Simultaneously, correlation affects the error free estimate, $\hat{\beta}_Z$, differently. The magnitude of relative bias changes from negative to positive indicating at higher levels of correlation and error, the parameter estimate is overestimated as opposed to under estimated. These results indicate that failure to correct for correlation and measurement error may lead to misleading inference.

Methods that Incorporate Correlation and Measurement

Tables 4 – 18 report results of the estimates obtained from the methods which incorporate a correction for correlation and measurement error. Estimates for β_Z resulting from RCNO should be the same as the results from the Naïve approach. This occurs because RCNO only corrects for measurement error and not correlation. The correlation measurement error combination not only impacts the error prone covariate, it also impacts the error free one.

Table 4 reports the average parameter estimates, standard error and average standard error when $\sigma_U^2 = 0.1$. Notice that for design points found in this table, the range for the RCNO estimates is 0.017 for both $\hat{\beta}_X$ and $\hat{\beta}_Z$. Correspondingly, the range for

RCCORR is 0.012 and 0.009 for $\hat{\beta}_X$ and $\hat{\beta}_Z$ respectively; for RSCCORR, the range is 0.033 and 0.068 and for RSCADJ, the range is 0.029 and 0.66 for $\hat{\beta}_X$ and $\hat{\beta}_Z$ respectively. Under the method RCNO, the estimates of β_X are all less than 1. This indicates that RCNO does correct for the expected attenuation that occurs as a result for the correlation/measurement error problem. Also, as we observe the estimates of β_X resulting from RCCORR, RCSCORR and RSCADJ, it is evident that these methods also correct for the expected attenuation problem as the estimates are relatively close to 1. Yet, each method over estimates our expected value except at design point (0.5, 0.1). RCCORR provides an estimate at this design point slightly below 1. Also, this same method tends have to better estimates as correlation increases. Additionally, note that the estimates for standard error and average standard error are acceptable. The estimates of β_Z indicate that each method corrects for attenuation. Failure to correct for correlation and measurement error results in the best results. When correlation is small, RSCCORR and RSCADJ perform well followed by RCCORR. As correlation increases, we see a separation in the performance of the methods. At moderate levels of correlation, RCCORR, RSCCORR and RSCADJ perform similarly, yielding similar results. At the highest level of correlation, we see that the performance of the risk set methods is not a good and the regression calibration methods. Again, the standard error and average standard error estimates are appropriate.

The performance of estimators is also examined by the empirical coverage, relative and absolute bias. These results are displayed in Table 5. At each level of correlation, the coverage for the regression calibration methods is adequate.

Alternatively, the coverage of $\hat{\beta}_X$ resulting from RSCCORR and RSCADJ is not as robust; it is less than the nominal level. Correspondingly, the risk set calibration methods overestimate the true value of β_X by at least 11.6%. The regression calibration methods perform better. The impact of correlation/measurement error is not as great in the error free estimators. Relative bias for these same methods ranges from -7% to 0%. The results in Table 5 corroborate the results found in Table 4. The absolute bias is small for RCNO and RCCORR at each level of correlation. In the estimate of β_X, we observe that as correlation increases, the RCCORR method outperforms that of RCNO. As expected, the correlation adjustment improves the ability to estimate true β_X. The risk set calibration methods do not perform as well as the regression calibration methods at this level of error. While the amount of bias is acceptable, both of these methods are outperformed by the Naïve approach as found in Table 3. The absolute bias for the Naïve method is smaller at each level of correlation than either risk set calibration method. The same trends do not occur in methods used to estimate β_Z. At the lowest level of correlation, we observe that RSCCORR and RSCADJ perform well, yet, as correlation increases, the absolute bias increases, indicating that the methods are sensitive to increases in correlation. As correlation increases, the estimates acquired from the RCNO method are more accurate than the other methods. As expected, the largest absolute bias occurs under the RSCCORR method, which is very sensitive to risk set size. The estimates of β_X indicate that each method corrects for the attenuation effect observed in Table 2.

Table 6 records the mean square error and relative efficiency of the estimator from each method under the prescribed design points. The relative efficiency is the ratio

of the mean square error of the estimator resulting from the method that corrects for correlation and error and the mean square of the naive method. The mean square error for each method at each level of correlation is consistent. The relative efficiency is greater than 1 for estimates from the RSCCORR and RSCADJ methods. This indicates that under each design point, the Naïve method is more efficient than the methods that utilize the risk set approach. Alternatively, RCNO is more efficient than the Naïve approach because the relative efficiency of $\hat{\beta}_X$ resulting from RCNO is less than 1 at each level of correlation. As indicated earlier, the performance of RCCORR improves as correlation increases. We observe that the relative efficiency of RCCORR goes from greater than 1 to less than 1 as correlation increases. Also note, the relative efficiency for $\hat{\beta}_Z$ resulting from RCNO is 1. This will occur at every design point since the method does not adjust for correlation; the parameter estimate is the same as the Naïve approach. The relative efficiency of $\hat{\beta}_Z$ resulting from RSCCORR at the lowest level of correlation is 0.997. This indicates the RSCCORR is more efficient than the Naïve method.

Tables 7, 8 and 9 show the results from $\sigma_U^2 = 0.3$. Notice, for these design points, that we observe similar results to that of the previous design point configuration. The estimates for β_X and β_Z vary from 0.96 to 1.65 and 0.74 to 1.03 respectively. RCNO and RCCORR seem to estimate the true parameters relatively well. When correlation is small, RCNO seems to perform slightly better as the parameter estimate is very close to 1. As the amount of correlation increases, RCCORR shows better performance in estimating $\beta_X = 1$. Both RCNO and RCCORR provide estimates that are closer to $\beta_X = 1$ as compared to the Naïve approach found in Table 2. Upon closer inspection of Table 7, we

Table 2 Simulation Summary Statistics for $\hat{\beta}_X$ and $\hat{\beta}_Z$ with true $\beta_X = 1,\ \beta_Z = 1$ and sample size *n*= 150; True covariates are generated from bivariate normal distributions; Results are from 1060 replicates

Method	ρ	σ_U^2	$\hat{\beta}_X$	$SE\left(\hat{\beta}_X\right)$	Average$\left(SE\left(\hat{\beta}_X\right)\right)$	$\hat{\beta}_Z$	$SE\left(\hat{\beta}_Z\right)$	Average$\left(SE\left(\hat{\beta}_Z\right)\right)$
Naïve	0.1	0.1	0.954	0.135	0.142	0.992	0.135	0.146
	0.3		0.948	0.140	0.146	0.999	0.141	0.151
	0.5		0.938	0.152	0.156	1.009	0.155	0.162
	0.1	0.3	0.877	0.133	0.135	0.977	0.136	0.145
	0.3		0.867	0.138	0.139	0.998	0.142	0.150
	0.5		0.848	0.149	0.148	1.025	0.155	0.161
	0.1	0.5	0.812	0.131	0.130	0.965	0.137	0.144
	0.3		0.800	0.135	0.133	0.997	0.142	0.150
	0.5		0.773	0.145	0.141	1.039	0.155	0.161
	0.1	1.0	0.685	0.123	0.118	0.943	0.138	0.143
	0.3		0.669	0.126	0.121	0.996	0.144	0.149
	0.5		0.635	0.134	0.127	1.066	0.156	0.159
	0.1	2.0	0.523	0.109	0.102	0.916	0.140	0.141
	0.3		0.506	0.111	0.104	0.997	0.145	0.147
	0.5		0.468	0.116	0.108	1.101	0.156	0.158

The measurement error is Normal $(0, \sigma_U^2)$; Sample standard error, $SE\left(\hat{\beta}_X\right)$, is the standard deviation of Monte Carlo estimates; Average standard error, Average$\left(SE\left(\hat{\beta}_X\right)\right)$, is the average of Monte Carlo estimated standard errors.

Table 3 Coverage, Relative and Absolute Bias of Average Parameter Estimates for Naïve Method

Method	ρ	σ_U^2	Coverage $\hat{\beta}_X$	Relative Bias $\hat{\beta}_X$	Absolute Bias $\hat{\beta}_X$	Coverage $\hat{\beta}_Z$	Relative Bias $\hat{\beta}_Z$	Absolute Bias $\hat{\beta}_Z$
Naïve	0.1	0.1	95.1	-4.6	0.046	97.6	-0.8	0.008
	0.3		94.8	-5.2	0.052	97.1	-0.1	0.001
	0.5		93.9	-6.2	0.062	97	0.9	0.009
	0.1	0.3	83.5	-12.3	0.123	95.9	-2.3	0.023
	0.3		81.3	-13.3	0.133	97.1	-0.2	0.002
	0.5		78.3	-15.2	0.152	97.3	2.5	0.025
	0.1	0.5	65.4	-18.8	0.188	95	-3.46	0.035
	0.3		63.4	-20.0	0.200	96.7	-0.3	0.003
	0.5		59.5	-22.7	0.227	96.5	3.9	0.039
	0.1	1.0	24.8	-31.5	0.315	92.8	-5.7	0.057
	0.3		24.1	-33.1	0.331	96.4	-0.4	0.004
	0.5		20.8	-36.5	0.365	95.1	6.6	0.066
	0.1	2.0	1.9	-47.7	0.477	90.1	-8.4	0.084
	0.3		1.6	-49.4	0.494	96.2	-0.3	0.003
	0.5		0.9	-53.2	0.532	92.7	10.1	0.101

Table 4 Simulation Summary Statistics for $\hat{\beta}_X$ and $\hat{\beta}_Z$ with true $\beta_X = 1,\ \beta_Z = 1$ and sample size *n*= 150; True covariates are generated from bivariate normal distributions; measurement error variance is $\sigma_U^2 = 0.1$

Approach	ρ	$\hat{\beta}_X$	$SE\left(\hat{\beta}_X\right)$	Average$\left(SE\left(\hat{\beta}_X\right)\right)$	$\hat{\beta}_Z$	$SE\left(\hat{\beta}_Z\right)$	Average$\left(SE\left(\hat{\beta}_Z\right)\right)$
RCNO[§]	0.1	0.999	0.142	0.149	0.992	0.135	0.146
RCCORR [‡]		1.011	0.144	0.150	0.969	0.134	0.144
RSCCORR [*]		1.120	0.167	0.161	0.999	0.135	0.140
RSCADJ[°]		1.116	0.167	0.166	1.000	0.138	0.146
RCNO[§]	0.3	0.992	0.148	0.153	0.999	0.141	0.151
RCCORR [‡]		1.002	0.149	0.155	0.978	0.141	0.151
RSCCORR [*]		1.127	0.177	0.168	0.975	0.141	0.145
RSCADJ[°]		1.121	0.175	0.173	0.974	0.143	0.151
RCNO[§]	0.5	0.982	0.160	0.164	1.009	0.155	0.162
RCCORR [‡]		0.999	0.163	0.167	0.978	0.157	0.163
RSCCORR [*]		1.153	0.202	0.187	0.931	0.160	0.159
RSCADJ[°]		1.145	0.197	0.191	0.934	0.162	0.165

[§] Regression Calibration No Adjustment for Correlation
[‡] Regression Calibration Adjusted for Correlation
[*] Risk Set Calibration Corrected for Correlation
[°] Risk Set Calibration Corrected for Correlation and adjusted to improve efficiency

Sample standard error, $SE\left(\hat{\beta}_X\right)$, is the standard deviation of Monte Carlo estimates;

Average standard error, Average$\left(SE\left(\hat{\beta}_X\right)\right)$, is the average of Monte Carlo estimated standard errors.

Table 5 Coverage, Relative and Absolute Bias of Average Parameter Estimates for $\sigma_U^2 = 0.1$

Method	ρ	Coverage $\hat{\beta}_X$	Relative Bias $\hat{\beta}_X$	Absolute Bias $\hat{\beta}_X$	Coverage $\hat{\beta}_Z$	Relative Bias $\hat{\beta}_Z$	Absolute Bias $\hat{\beta}_Z$
RCNO[§]	0.1	95.9	-0.1	0.001	97.6	-0.8	0.008
RCCORR [‡]		95.8	1.1	0.011	95.8	-3.1	0.031
RSCCORR [*]		91.1	12.0	0.120	97.1	0.1	0.001
RSCADJ[°]		91.0	11.6	0.116	97.3	0.0	0.000
RCNO[§]	0.3	96.4	-0.8	0.008	97.1	0.1	0.001
RCCORR [‡]		96.5	0.3	0.002	97.1	-2.2	0.022
RSCCORR [*]		90.1	12.7	0.127	96.6	-2.5	0.025
RSCADJ[°]		90.6	12.1	0.121	96.8	-2.6	0.026
RCNO[§]	0.5	96.7	-1.8	0.018	97	0.9	0.009
RCCORR [‡]		97.0	-0.1	0.001	95.9	-2.2	0.022
RSCCORR [*]		89.1	15.3	0.153	92.9	-6.9	0.069
RSCADJ[°]		90.5	14.5	0.145	93.1	-6.6	0.066

Table 6 Mean Square Error, Relative Efficiency $\sigma_U^2 = 0.1$

Method	ρ	Error	Mean Square Error $\hat{\beta}_X$	Relative Efficiency $\hat{\beta}_X$	Mean Square Error $\hat{\beta}_Z$	Relative Efficiency $\hat{\beta}_Z$
RCNO[§]	0.1	0.1	0.020	0.991	0.018	1.000
RCCORR [‡]			0.021	1.025	0.019	1.034
RSCCORR [*]			0.042	2.079	0.018	0.997
RSCADJ[°]			0.041	2.033	0.019	1.041
RCNO[§]	0.3	0.1	0.022	0.985	0.020	1.000
RCCORR [‡]			0.022	0.996	0.020	1.024
RSCCORR [*]			0.047	2.128	0.021	1.031
RSCADJ[°]			0.045	2.030	0.021	1.063
RCNO[§]	0.5	0.1	0.026	0.962	0.024	1.000
RCCORR [‡]			0.027	0.986	0.025	1.043
RSCCORR [*]			0.064	2.383	0.030	1.259
RSCADJ[°]			0.060	2.220	0.031	1.269

observe RSCCORR and RSCADJ perform similarly. These two methods are very similar in that the calibration method is the same with the exception of the tails of the risk set. The resulting behavior of the estimates should be similar. There is a slight inflation in the estimate $\hat{\beta}_X$ as correlation increases. Larger levels of correlation result in a larger value of $\hat{\beta}_X$. For RSCCORR and RSCADJ, we observe that the estimate of the standard error has been inflated. The regression calibration approaches, with the exception of design point (0.5, 0.3) (correlation, measurement error), yield downwardly biased estimates for β_Z. We also observe that as correlation increases, the estimates obtained from risk set calibration approaches become more variable and downwardly biased.

Table 8 contains estimates which describe the coverage and accuracy of the methods. For both regression calibration methods, the results $\hat{\beta}_X$ and $\hat{\beta}_Z$ resulting from RCNO and RCCORR have reasonable coverage; the coverage is relatively close to the nominal rate. Alternatively, coverage for $\hat{\beta}_X$ is particularly poor for the risk set calibration methods and fair for $\hat{\beta}_Z$. This is due to the increase in the standard error and absolute bias estimates. It follows that the confidence interval used to determine the coverage has a shifted location yet not a wider interval. Again, we observe that as correlation increases, the absolute bias of β_X is smaller for RCCORR compared to RCNO. Also, when correlation is small, RSCCORR and RSCADJ perform well when estimating β_Z but not so when estimating β_X. Each has an absolute bias of 0.002 when estimating β_Z, yet the absolute bias in β_X is 0.462 and 0.446 for RSCCORR and

RSCADJ respectively. RCNO still underestimates β_X by at most 3.7%. Alternatively, the risk set calibration methods overestimate β_X by at least 44.6%.

The results in Table 9 provide a comparison of the method designed to correct for correlation and measurement error to ignoring both correlation and measurement error. The results found in Table 9 are not surprising based on what we have observed thus far. At this point, we observe that the MSE for $\hat{\beta}_X$ has increased considerably from what was observed in Table 6. More specifically, there is a 7 fold increase in MSE for the risk set calibration methods. While not as steep as what was observed in MSE for $\hat{\beta}_X$, the impact of increased measurement error also affects the MSE of $\hat{\beta}_Z$. There is a 1- to 4-fold increase in the estimates as compared to measurement error of 0.1. In methods used to estimate β_X, we observe that the relative efficiency, for both RCNO and RCCORR is less than 1 indicating that the Naïve method is inefficient compared to the regression calibration methods. Alternatively, RSCCORR and RSCADJ are considerably larger than 1. Hence, the risk set calibration methods are not as efficient as ignoring correlation and measurement error. The relative efficiency for methods used to estimate β_Z take on a different structure. Each ratio is greater than 1 indicating that the Naïve method is more efficient the RCCORR, RSCCORR and RSCADJ.

For $\sigma_U^2 = 0.5,$ we again see a similar situation as to the previous value of σ_U^2. The results are found in Tables 10, 11 and 12. As we have seen in previous tables, RCNO and RCCORR outperformed RSCCORR and RSCCADJ in estimating the true value of the

Table 7 Simulation Summary Statistics for $\hat{\beta}_X$ and $\hat{\beta}_Z$ with true $\beta_X = 1$, $\beta_Z = 1$ and sample size *n*= 150; True covariates are generated from bivariate normal distributions; measurement error variance is $\sigma_U^2 = 0.3$

Approach	ρ	$\hat{\beta}_X$	$SE\left(\hat{\beta}_X\right)$	Average$\left(SE\left(\hat{\beta}_X\right)\right)$	$\hat{\beta}_Z$	$SE\left(\hat{\beta}_Z\right)$	Average$\left(SE\left(\hat{\beta}_Z\right)\right)$
RCNO[§]	0.1	0.997	0.158	0.154	0.977	0.136	0.145
RCCORR [‡]		1.018	0.164	0.157	0.924	0.137	0.143
RSCCORR [*]		1.462	0.328	0.222	1.002	0.152	0.137
RSCADJ[°]		1.446	0.298	0.229	1.002	0.153	0.147
RCNO[§]	0.3	0.986	0.163	0.158	0.998	0.142	0.150
RCCORR [‡]		1.013	0.171	0.163	0.935	0.145	0.150
RSCCORR [*]		1.502	0.367	0.238	0.900	0.166	0.142
RSCADJ[°]		1.484	0.325	0.246	0.902	0.164	0.151
RCNO[§]	0.5	0.963	0.175	0.168	1.025	0.155	0.161
RCCORR [‡]		1.010	0.188	0.177	0.939	0. 164	0.164
RSCCORR [*]		1.649	0.535	0.287	0.726	0.253	0.168
RSCADJ[°]		1.609	0.400	0.296	0.740	0.220	0.180

[§] Regression Calibration No Adjustment for Correlation
[‡] Regression Calibration Adjusted for Correlation
[*] Risk Set Calibration Corrected for Correlation
[°] Risk Set Calibration Corrected for Correlation and adjusted to improve efficiency

Sample standard error, $SE\left(\hat{\beta}_X\right)$, is the standard deviation of Monte Carlo estimates;

Average standard error, Average$\left(SE\left(\hat{\beta}_X\right)\right)$, is the average of Monte Carlo estimated standard errors.

Table 8 Coverage, Relative and Absolute Bias of Average Parameter Estimates for $\sigma_U^2 = 0.3$

Method	ρ	Coverage $\hat{\beta}_X$	Relative Bias $\hat{\beta}_X$	Absolute Bias $\hat{\beta}_X$	Coverage $\hat{\beta}_Z$	Relative Bias $\hat{\beta}_Z$	Absolute Bias $\hat{\beta}_Z$
RCNO[§]	0.1	94.7	-0.3	0.003	95.9	-2.3	0.023
RCCORR [‡]		94.4	1.8	0.018	92.0	-7.6	0.076
RSCCORR[*]		50.3	46.2	0.462	93.2	0.2	0.002
RSCADJ[°]		54.0	44.6	0.446	94.5	0.2	0.002
RCNO[§]	0.3	94.8	-1.4	0.014	97.1	-0.2	0.002
RCCORR [‡]		94.6	1.3	0.013	93.7	-6.5	0.065
RSCCORR[*]		49.6	50.2	0.502	87.6	-10.0	0.100
RSCADJ[°]		53.5	48.4	0.484	89.2	-9.8	0.098
RCNO[§]	0.5	94.1	-3.7	0.037	97.3	2.5	0.025
RCCORR [‡]		95.1	1.0	0.010	93.4	-6.1	0.061
RSCCORR[*]		46.8	64.9	0.649	63.8	-27.4	0.274
RSCADJ[°]		49.8	60.9	0.609	67.7	-26.0	0.260

Table 9 Mean Square Error, Relative Efficiency $\sigma_U^2 = 0.3$

Method	ρ	Error	Mean Square Error $\hat{\beta}_X$	Relative Efficiency $\hat{\beta}_X$	Mean Square Error $\hat{\beta}_Z$	Relative Efficiency $\hat{\beta}_Z$
RCNO[§]	0.1	0.3	0.025	0.761	0.019	1.000
RCCORR [‡]			0.027	0.829	0.025	1.290
RSCCORR[*]			0.321	9.782	0.023	1.215
RSCADJ[°]			0.288	8.767	0.023	1.231
RCNO[§]	0.3	0.3	0.027	0.729	0.020	1.000
RCCORR [‡]			0.029	0.801	0.025	1.252
RSCCORR[*]			0.387	10.527	0.038	1.862
RSCADJ[°]			0.340	9.253	0.037	1.810
RCNO[§]	0.5	0.3	0.032	0.706	0.025	1.000
RCCORR [‡]			0.035	0.782	0.031	1.242
RSCCORR[*]			0.707	15.615	0.139	5.642
RSCADJ[°]			0.531	11.718	0.116	4.706

The estimates resulting from RCNO vary from 0.942 to 0.990 and 0.965 to 1.039 for $\hat{\beta}_X$ and $\hat{\beta}_Z$ respectively. The intervals (1.014, 1.024) and (0.880, 0.905) for $\hat{\beta}_X$ and $\hat{\beta}_Z$ respectively, represent the minimum and maximum values of the estimates produced by the RCCORR method. The estimates, $\hat{\beta}_X$ and $\hat{\beta}_Z$, from the RSCCORR method vary from (1.641, 1.837) and (0.673, 0.994) respectively. Finally, for estimates resulting from RSCADJ, the minimum maximum combination is (1.692, 1.822) and (0.655, 0.996) for $\hat{\beta}_X$ and $\hat{\beta}_Z$ respectively. Again, the methods make a correction for attenuation. RCCORR, RSCCORR and RSCADJ overestimate $\hat{\beta}_X$ and underestimate $\hat{\beta}_Z$ at each level of correlation with an exception occurring at design point (0.5, 0.5). Also note, the standard error for risk set calibration methods has become inflated.

While fair for both RCNO and RCCORR, coverage, found in Table 11 for the other methods is lower than the nominal level of 95. Interestingly, notice that the coverage for β_Z increases for RCCORR as correlation increases. While the risk set calibration methods overestimate β_X by more than 60%, the same methods underestimate true β_Z by at most 34.0%. With the exception of the lowest level of correlation, the absolute bias indicates that the regression calibration methods provide a more accurate approximation of the true values of β_X and β_Z. The absolute bias of $\hat{\beta}_X$ indicates that as correlation increases, RCCORR out performs all methods yet at the lowest level of correlation, RCNO provides a more accurate estimate. The performance of each RSCCORR and RSCADJ is similar. Now consider the estimates $\hat{\beta}_Z$. When correlation is 0.1, the resulting absolute bias indicates that RSCORR and RSCADJ

provide more precise estimates. As correlation increases, RCNO produces more accurate estimates of β_Z as shown by smaller values for absolute bias.

As before, there are similar patterns in the MSE and relative efficiency for both $\hat{\beta}_X$ and $\hat{\beta}_Z$ which are found in Table 12. All of the estimates of MSE have increased as compared to what has been observed at smaller levels of measurement error. Due to the increased variability and decrease in accuracy, the largest MSE estimate results from RSCCORR followed RSCADJ. The RCNO and RCCORR estimators are more efficient than Naive method for estimating β_X as the resulting relative efficiency is less than 1. On the contrary, the RSCCORR and RSCADJ methods are less efficient than Naïve. The ratio of mean square error of the risk set calibration methods and the Naïve method is greater than 1. When ρ=0.1, the estimates for β_Z have the opposite pattern. The relative efficiency of RCCORR is greater than that of RSCADJ. This pattern does not occur when correlation is 0.3 or 0.5.

We now progress to a higher level of measurement error, $\sigma_U^2 = 1.0$. Tables 13 through 15 contain the results for design point (0.1, 1.0), (0.3, 1.0) and (0.5, 1.0). In contrast to the performance that we have already observed, we now see in that RSCCORR and RSCADJ now underestimate the true parameter value, β_X while the estimates resulting from the RCNO and RCCORR methods produce familiar results. The RSCCORR estimates (standard error) vary in value from -0.64 (1.35) to -0.30 (2.30) while those for RSCADJ estimates (standard error) vary from -0.66 (0.57) to -0.11 (0.93). The large standard errors for the risk set calibration methods indicate that estimates have large variability. RCNO and RCCORR correct for attenuation. In contrast, risk set

Table 10 Simulation Summary Statistics for $\hat{\beta}_X$ and $\hat{\beta}_Z$ with true $\beta_X = 1$, $\beta_Z = 1$ and sample size *n*= 150; True covariates are generated from bivariate normal distributions; measurement error variance is $\sigma_U^2 = 0.5$

Approach	ρ	$\hat{\beta}_X$	$SE\left(\hat{\beta}_X\right)$	Average$\left(SE\left(\hat{\beta}_X\right)\right)$	$\hat{\beta}_Z$	$SE\left(\hat{\beta}_Z\right)$	Average$\left(SE\left(\hat{\beta}_Z\right)\right)$
RCNO[§]	0.1	0.990	0.173	0.158	0.965	0.137	0.144
RCCORR [‡]		1.024	0.187	0.164	0.880	0.141	0.142
RSCCORR[*]		1.800	0.771	0.329	0.994	0.194	0.134
RSCADJ[°]		1.798	0.508	0.343	0.996	0.181	0.148
RCNO[§]	0.3	0.974	0.178	0.162	0.997	0.142	0.150
RCCORR [‡]		1.018	0.195	0.170	0.898	0.150	0.150
RSCCORR[*]		1.837	0.942	0.371	0.812	0.239	0.144
RSCADJ[°]		1.822	0.595	0.380	0.818	0.203	0.157
RCNO[§]	0.5	0.942	0.189	0.172	1.039	0.155	0.161
RCCORR [‡]		1.014	0.217	0.185	0.905	0.173	0.166
RSCCORR[*]		1.641	1.837	0.483	0.673	0.640	0.218
RSCADJ[°]		1.692	0.893	0.476	0.655	0.369	0.225

[§] Regression Calibration No Adjustment for Correlation
[‡] Regression Calibration Adjusted for Correlation
[*] Risk Set Calibration Corrected for Correlation
[°] Risk Set Calibration Corrected for Correlation and adjusted to improve efficiency

Sample standard error, $SE\left(\hat{\beta}_X\right)$, is the standard deviation of Monte Carlo estimates;

Average standard error, Average$\left(SE\left(\hat{\beta}_X\right)\right)$, is the average of Monte Carlo estimated standard errors.

Table 11 Coverage, Relative and Absolute Bias of Average Parameter Estimates for $\sigma_U^2 = 0.5$

Method	ρ	Coverage $\hat{\beta}_X$	Relative Bias $\hat{\beta}_X$	Absolute Bias $\hat{\beta}_X$	Coverage $\hat{\beta}_Z$	Relative Bias $\hat{\beta}_Z$	Absolute Bias $\hat{\beta}_Z$
RCNO[§]	0.1	92.8	-1.1	0.01	95.0	-3.5	0.035
RCCORR [‡]		92.3	2.4	0.024	85.7	-12.0	0.120
RSCCORR[*]		35.1	80.0	0.800	86.4	-0.6	0.006
RSCADJ[°]		37.2	79.8	0.798	89.7	-0.4	0.004
RCNO[§]	0.3	93.4	-2.6	0.026	96.7	-0.3	0.003
RCCORR [‡]		92.4	1.8	0.018	89.5	-10.2	0.102
RSCCORR[*]		35.9	83.7	0.837	70.3	-18.8	0.188
RSCADJ[°]		41.7	82.2	0.822	73.1	-18.2	0.182
RCNO[§]	0.5	91.7	-5.8	0.058	96.5	3.9	0.039
RCCORR [‡]		92.6	1.4	0.014	89.8	-9.5	0.095
RSCCORR[*]		44.5	64.1	0.641	53.6	-32.7	0.327
RSCADJ[°]		53.5	69.2	0.692	58.6	-34.5	0.345

Table 12 Mean Square Error, Relative Efficiency $\sigma_U^2 = 0.5$

Method	Correlation	Error	Mean Square Error $\hat{\beta}_X$	Relative Efficiency $\hat{\beta}_X$	Mean Square Error $\hat{\beta}_Z$	Relative Efficiency $\hat{\beta}_Z$
RCNO[§]	0.1	0.5	0.030	0.572	0.020	1.000
RCCORR [‡]			0.036	0.677	0.034	1.715
RSCCORR[*]			1.234	23.511	0.038	1.884
RSCADJ[°]			0.895	17.043	0.033	1.639
RCNO[§]	0.3	0.5	0.032	0.556	0.020	1.000
RCCORR [‡]			0.038	0.659	0.033	1.631
RSCCORR[*]			1.588	27.272	0.092	4.584
RSCADJ[°]			1.030	17.685	0.074	3.685
RCNO[§]	0.5	0.5	0.039	0.539	0.026	1.000
RCCORR [‡]			0.047	0.652	0.039	1.525
RSCCORR[*]			3.785	52.174	0.517	20.220
RSCADJ[°]			1.276	17.591	0.255	9.989

calibration no longer provides a correction for attenuation as the estimates are very small. The variability of the estimates resulting from the risk set methods is large. Consider the estimates generated for β_Z. Again, estimates from RCCORR improve as correlation increases yet variability increases. Also, as correlation increases, RSCCORR and RSCADJ overestimate β_Z. All methods are affected by increased levels of correlation and measurement error but the effect is more pronounced in the risk set calibration methods.

The coverage rates, found in Table 14, for the estimates of β_X are below the nominal level. The coverage rates for the RCNO estimates of β_Z are at or slightly above nominal level while the rates for RCCORR, RSCCORR and RSCADJ are below the nominal level. Looking at the relative and absolute biases found in Table 14, we can inspect more closely the quality of the estimates. There is evidence to support that regression calibration methods do a better job at estimation of β_X. As correlation increases, the absolute bias indicates that the estimates of β_X resulting from RCCORR are more accurate than RCNO. Alternatively, RCNO outperforms RCCORR in estimation of β_Z as the absolute bias is smaller at each level of correlation. The absolute bias for β_X resulting from RSCCORR is larger than the absolute bias from the RSCADJ at small to moderate levels of correlation. The risk set calibration methods underestimate true β_X by more than 110% which substantiate the findings in Table 13.

Results found in Table 15 show that the RCNO estimator and RCCORR estimator are satisfactory for β_X. At each level of correlation, both regression calibration methods prove to be more efficient than the Naïve approach. The relative efficiency associated

with each of RSCCORR and RSCADJ indicate that the Naïve approach is more efficient than both risk set calibration methods. In each, RSCCORR and RSCADJ, the measure of bias and precision, MSE, indicate that the methods are not performing well. Again, the largest estimate of MSE occurs when using RSCCORR followed by RSCADJ. The variability and accuracy of these two methods has been diminished by the amounts correlation and measurement error. Also, the MSE for each method is greater than what has been observed at the previously mentioned levels of measurement error. It is here that we see the damaging effects of correlation and measurement error.

In order to finish studying the ability of our methods to estimate β_X and β_Z, we consider the last level of measurement error under consideration, $\sigma_U^2 = 2.0$. The results are found in Tables 16 through 18. As shown in Table 16 all methods under consideration underestimate the true parameter estimate, $\beta_X = 1$. The estimates $\hat{\beta}_X$ vary from -0.550 to 0.906 while the estimates for β_Z vary from 0.699 to 1.390. Note that for each method, the estimate of β_Z increases at each level of correlation. This indicates that the methods are correcting for attenuation. However, this pattern is not observed in the estimates $\hat{\beta}_X$. The correction for attenuation is apparent in the results from methods RCNO and RCCORR. The standard error estimates are much larger than what has appeared in earlier tables. The estimates are more variable at this high level of measurement error. While RCNO and RSCCORR outperform the risk set calibration methods in estimating β_X, we observe that the ability of these regression calibration methods has diminished in light of the increased levels of measurement error and correlation.

For the design points under consideration found in Table 17, we observe that

Table 13 Simulation Summary Statistics for $\hat{\beta}_X$ and $\hat{\beta}_Z$ with true $\beta_X = 1$, $\beta_Z = 1$ and sample size *n*= 150; True covariates are generated from bivariate normal distributions; measurement error variance is: $\sigma_U^2 = 1.0$

Approach	ρ	$\hat{\beta}_X$	$SE\left(\hat{\beta}_X\right)$	Average$\left(SE\left(\hat{\beta}_X\right)\right)$	$\hat{\beta}_Z$	$SE\left(\hat{\beta}_Z\right)$	Average$\left(SE\left(\hat{\beta}_Z\right)\right)$
RCNO[§]	0.1	0.949	0.205	0.164	0.943	0.138	0.143
RCCORR [‡]		1.011	0.245	0.175	0.794	0.156	0.141
RSCCORR[*]		-0.467	1.665	0.368	0.861	0.252	0.123
RSCADJ[°]		-0.113	0.933	0.359	0.853	0.187	0.144
RCNO[§]	0.3	0.927	0.208	0.168	0.996	0.144	0.149
RCCORR [‡]		1.004	0.255	0.182	0.826	0.171	0.152
RSCCORR[*]		-0.639	1.349	0.366	1.163	0.337	0.147
RSCADJ[°]		-0.343	0.857	0.354	1.085	0.242	0.162
RCNO[§]	0.5	0.879	0.216	0.176	1.066	0.156	0.159
RCCORR [‡]		1.000	0.296	0.201	0.844	0.208	0.173
RSCCORR[*]		-0.936	1.150	0.299	1.619	0.511	0.186
RSCADJ[°]		-0.660	0.572	0.311	1.481	0.308	0.205

[§] Regression Calibration No Adjustment for Correlation
[‡] Regression Calibration Adjusted for Correlation
[*] Risk Set Calibration Corrected for Correlation
[°] Risk Set Calibration Corrected for Correlation and adjusted to improve efficiency

Sample standard error, $SE\left(\hat{\beta}_X\right)$, is the standard deviation of Monte Carlo estimates;

Average standard error, Average$\left(SE\left(\hat{\beta}_X\right)\right)$, is the average of Monte Carlo estimated standard errors.

Table 14 Coverage, Relative and Absolute Bias of Average Parameter Estimates for $\sigma_U^2 = 1.0$

Method	ρ	Coverage $\hat{\beta}_X$	Relative Bias $\hat{\beta}_X$	Absolute Bias $\hat{\beta}_X$	Coverage $\hat{\beta}_Z$	Relative Bias $\hat{\beta}_Z$	Absolute Bias $\hat{\beta}_Z$
RCNO[§]	0.1	86.3	-5.1	0.051	92.8	-5.7	0.057
RCCORR [‡]		86.9	1.1	0.011	66.3	-20.6	0.206
RSCCORR[*]		27.6	-146.7	1.467	68.1	-13.9	0.139
RSCADJ[°]		26.1	-111.3	1.113	74.2	-14.7	0.147
RCNO[§]	0.3	85.7	-7.3	0.073	96.4	-0.4	0.004
RCCORR [‡]		88.2	0.4	0.004	76.5	-17.4	0.174
RSCCORR[*]		22.7	-163.9	1.639	70.8	16.3	0.163
RSCADJ[°]		20.0	-134.3	1.343	82.3	8.5	0.085
RCNO[§]	0.5	80.8	-12.1	0.121	95.1	6.6	0.066
RCCORR [‡]		86.8	0.0	0.000	81.7	-15.6	0.156
RSCCORR[*]		13.9	-193.6	1.936	39.4	61.9	0.619
RSCADJ[°]		6.5	-166.0	1.66	34.9	48.1	0.481

Table 15 Mean Square Error, Relative Efficiency $\sigma_U^2 = 1.0$

Method	ρ	Error	Mean Square Error $\hat{\beta}_X$	Relative Efficiency $\hat{\beta}_X$	Mean Square Error $\hat{\beta}_Z$	Relative Efficiency $\hat{\beta}_Z$
RCNO[§]	0.1	1.0	0.045	0.390	0.022	1.000
RCCORR [‡]			0.060	0.526	0.067	2.995
RSCCORR[*]			4.924	43.062	0.083	3.715
RSCADJ[°]			2.109	18.445	0.057	2.538
RCNO[§]	0.3	1.0	0.049	0.387	0.021	1.000
RCCORR [‡]			0.065	0.519	0.060	2.868
RSCCORR[*]			4.506	35.923	0.140	6.753
RSCADJ[°]			2.538	20.234	0.066	3.170
RCNO[§]	0.5	1.0	0.061	0.405	0.029	1.000
RCCORR [‡]			0.088	0.580	0.068	2.356
RSCCORR[*]			5.071	33.540	0.644	22.455
RSCADJ[°]			3.083	20.391	0.326	11.370

coverage is unacceptable except for $\hat{\beta}_Z$, produced by RCNO at the 0.3 level of correlation. Turning to the relative and absolute biases, we observe that RSCCORR and RSCADJ underestimate the true value of β_X by at least 130%. RCNO and RCCORR outperform the risk set calibration methods in estimating β_X. More specifically, RCCORR provides more accurate results than RCNO as the absolute bias is smaller for RCCORR than for RCNO. As has been observed in the other design points, RCNO performs best in estimating β_Z. followed by RSCCORR and RSCADJ. At the highest level of correlation, RCCORR produces more accurate results than RSCCORR and RSCADJ. The absolute bias is smaller. At low to moderate levels of correlation, RSCCORR and RSCADJ outperform RCCORR

Again, the relative efficiency for both RCNO and RCCORR is less than 1, again indicating that regression calibration methods are more efficient than the Naïve approach when estimating β_X. Conversely, RSCCORR and RSCADJ are less efficient than the Naive approach. The RCNO method is the best estimator of the true parameter value β_Z. At each level of correlation, the relative efficiency resulting from the RCNO method is approximately or less than 1; thus indicating that the RCNO method is barely more efficient than the Naïve method. At low to moderate levels of correlation, RSCCORR and RSCADJ outperform RCCORR in producing estimates for β_Z.

Conclusions

The methods under review can be classified as either regression calibration or risk

Table 16 Simulation Summary Statistics for $\hat{\beta}_X$ and $\hat{\beta}_Z$ with true $\beta_X = 1$, $\beta_Z = 1$ and sample size *n*= 150; True covariates are generated from bivariate normal distributions; measurement error variance is $\sigma_U^2 = 2.0$

Approach	ρ	$\hat{\beta}_X$	$SE\left(\hat{\beta}_X\right)$	Average$\left(SE\left(\hat{\beta}_X\right)\right)$	$\hat{\beta}_Z$	$SE\left(\hat{\beta}_Z\right)$	Average$\left(SE\left(\hat{\beta}_Z\right)\right)$
RCNO[§]	0.1	0.819	0.225	0.161	0.916	0.140	0.141
RCCORR [‡]		0.906	0.317	0.179	0.699	0.199	0.144
RSCCORR[*]		-0.398	0.739	0.108	0.858	0.206	0.115
RSCADJ[°]		-0.498	0.174	0.131	0.844	0.177	0.140
RCNO[§]	0.3	0.791	0.230	0.164	0.997	0.145	0.147
RCCORR [‡]		0.898	0.355	0.187	0.755	0.229	0.158
RSCCORR[*]		-0.362	0.897	0.103	1.098	0.553	0.125
RSCADJ[°]		-0.550	2.546	0.125	1.123	0.232	0.151
RCNO[§]	0.5	0.733	0.234	0.170	1.101	0.156	0.158
RCCORR [‡]		0.840	1.399	0.217	0.823	0.695	0.189
RSCCORR[*]		-0.298	2.299	0.088	1.362	1.918	0.136
RSCADJ[°]		-0.391	0.141	0.110	1.390	0.186	0.167

[§] Regression Calibration No Adjustment for Correlation
[‡] Regression Calibration Adjusted for Correlation
[*] Risk Set Calibration Corrected for Correlation
[°] Risk Set Calibration Corrected for Correlation and adjusted to improve efficiency

Sample standard error, $SE\left(\hat{\beta}_X\right)$, is the standard deviation of Monte Carlo estimates;

Average standard error, Average$\left(SE\left(\hat{\beta}_X\right)\right)$, is the average of Monte Carlo estimated standard errors.

Table 17 Coverage, Relative and Absolute Bias of Average Parameter Estimates for $\sigma_U^2 = 2.0$

Method	ρ	Coverage $\hat{\beta}_X$	Relative Bias $\hat{\beta}_X$	Absolute Bias $\hat{\beta}_X$	Coverage $\hat{\beta}_Z$	Relative Bias $\hat{\beta}_Z$	Absolute Bias $\hat{\beta}_Z$
RCNO[§]	0.1	63.4	-18.1	0.181	90.1	-8.4	0.084
RCCORR [‡]		13.7	-9.4	0.094	72.9	-30.1	0.301
RSCCORR[*]		13.7	-139.8	1.398	72.9	-14.2	0.142
RSCADJ[°]		0.0	-149.8	1.498	70.4	-15.6	0.156
RCNO[§]	0.3	59.8	-20.9	0.209	96.2	-0.3	0.003
RCCORR [‡]		69.5	-10.2	0.102	64.7	-24.5	0.245
RSCCORR[*]		13.7	-136.2	1.362	81.2	9.8	0.098
RSCADJ[°]		0.0	-155.0	1.550	70.4	12.3	0.123
RCNO[§]	0.5	52.6	-26.7	0.267	92.7	10.1	0.101
RCCORR [‡]		68.5	-16.0	0.160	78.9	-17.7	0.177
RSCCORR[*]		13.9	-129.8	1.298	39.4	36.2	0.362
RSCADJ[°]		0.0	-139.1	1.391	37.1	39.0	0.390

Table 18 Mean Square Error, Relative Efficiency $\sigma_U^2 = 2.0$

Method	ρ	Error	Mean Square Error $\hat{\beta}_X$	Relative Efficiency $\hat{\beta}_X$	Mean Square Error $\hat{\beta}_Z$	Relative Efficiency $\hat{\beta}_Z$
RCNO[§]	0.1	2.0	0.083	0.348	0.027	1.000
RCCORR [‡]			0.109	0.457	0.130	4.885
RSCCORR[*]			2.501	10.445	0.063	2.348
RSCADJ[°]			2.274	9.500	0.056	2.088
RCNO[§]	0.3	2.0	0.097	0.377	0.021	1.000
RCCORR [‡]			0.136	0.532	0.112	5.347
RSCCORR[*]			2.660	10.375	0.315	14.995
RSCADJ[°]			8.885	34.657	0.069	3.278
RCNO[§]	0.5	2.0	0.126	0.425	0.035	1.000
RCCORR [‡]			1.983	6.688	0.514	14.893
RSCCORR[*]			6.970	23.510	3.810	110.310
RSCADJ[°]			1.955	6.593	0.187	5.406

set calibration. The regression calibration methods seem to perform satisfactorily for small to moderate levels of correlation and measurement error. The other methods, RSCCORR and RSCADJ suffered under the impact of increased levels of correlation and measurement error.

The regression calibration methods consistently did a good job in estimating true values. While RCNO consistently underestimated the value of true β_X, the estimates including standard error remained stable at increased levels of correlation and measurement error. Consequently, making a correction for measurement error is better than ignoring measurement error and correlation. As levels of correlation increased, including a correction for correlation produced more accurate estimates. The RCCORR method, which adjusts for correlation, produced estimates of true β_X that had smaller absolute bias and reasonable estimates for MSE. While RCCORR consistently overestimated estimates of true β_X, the method also produced estimates that are stable. The estimates of true β_Z resulting from regression calibration methods too were stable.

The methods did encounter some problems as measurement error and correlation became large. When correlation and measurement error became large, the regression calibration estimates for true β_X, tended to be underestimated. RCCORR was outperformed by RCNO by producing estimates with small absolute bias; yet the variability increased. Nonetheless, that was not the case for estimates of true β_Z. RCNO provided less variable estimates. It is clear that the impact of the correlation/measurement error combination is greatest in the error prone variable and its effects are more pronounced at higher levels of correlation. It was not surprising to observe such erratic

results for the larger measurement error variances. As mentioned about the variability have great impact on the estimation processes.

The risk set calibration methods were not as steady as one would hope. In the presence of relatively large measurement error and correlation, the estimates were unstable. In some instances, where it was expected that the risk set calibration method would not estimate well, they actually outperformed the regression calibration methods. Although both RSCCORR and RSCADJ had problems with estimating the first parameter β_X, both typically did an adequate job at estimating β_Z. When RSCCORR and RSCADJ are pitted against the regression calibration methods, RCNO and RCCORR, the regression calibration methods, typically won. A major flaw of the risk set calibration methods was that they would severely underestimate β_X when correlation and measurement error were at high levels. In sum, although the risk set calibration methods have potential to be helpful in adjusting for correlation and measurement error in Cox proportional hazards modeling, they do not have stability that is hoped for.

It is important to indicate the measurement error more than correlation was the driving factor behind poor performance in the methods of discussion. Methods tended to perform better at the lowest level of correlation/measurement error. At these lower levels, the standard error and mean standard error was not overly inflated. Also, the coverage tended to be reasonable. Methods showed peak performance when both measurement error and correlation were less than 0.3. Methods performed poorly when design point levels were greater than 0.3.

CHAPTER 5

APPLICATION OF METHODS TO REAL UABMVS DATA

In order to illustrate regression calibration corrected for correlation and risk set calibration corrected for correlation the data from the UABMVS was used. In the study, elder drivers, returning to a Maryland Motor Vehicle Administration (MMVA) site (Burnie, Annapolis and Bel Air) to renew their drivers licenses were approached to participate in the study. Study participants were at least 55 years old. Those who agreed to participate completed the Gross Impairment Screening Battery (GRIMPS), a self-reported mobility instrument and Useful Field of View Subtest 2.[44]

Methods

Data

The GRIMPS battery is composed of cognitive and physical abilities measures. Cognitive Measures were assessed using the following tests: the Visual Closure subtest of the Motor-Free Visual Perception test (MVPT); Delayed Recall, Trail Making test and Symbol Scan. Physical Measures were assessed via the following tests: Rapid Walk, Tap Time, Arm Reach, Head/ Neck Rotation.

MVPT was used to detect visual pattern perception while the Delayed Recall Test was used to assess working memory.[44, 45] The Trail-making test, Part B, (TRAILS) was used to measure participants' general cognitive function. The test measures abilities to perform a directed visual search and to divide attention effectively.[44, 46] Symbol scan

was used to rule out neglect of one side of the visual field while driving.[44] Rapid Walk, Foot Tap and Arm Reach were used to assess lower and upper limb mobility respectively; while the Head/Neck Rotation assessed head/neck flexibility.[44, 45]

In addition to these measures, speed of processing was measured by the Useful Filed of View®, Subtest 2 (UFOV®).[44, 45, 47] Finally, each participant completed a self-report mobility questionnaire. The questionnaire assessed general mobility, driving behaviors, such as annual and weekly mileage and driving avoidance. Annual mileage was presented as a categorical variable with twelve levels. Subjects selected the most appropriate category to indicate their driving pattern. The categories were as follows: {less than 1000 miles per year, (1,001- 2500), (2,501-5000), (5,001- 7500), (7,501-10000), (10,001- 12500), (12,501-15000), (15,001- 17500), (17,501-20000), (20,001-25000), (25,001- 30000), greater than 30,000}. Miles per week was provided by the participant as a numerical constant. It is evident that the estimates for annual mileage are measured with error. Subjects were able to provide a best guess at their driving behavior. The cognitive tests, speed of processing as well as mobility measures have been shown to be related to elder drivers and impaired driving ability.[45, 48-50]

Outcome and Survival Time

The primary outcome of interest was an at-fault motor vehicle crash. This information was provided by the MMVA.[44] Survival time was defined as time from baseline assessment to either an at-fault motor vehicle crash, end of study or loss to follow up.

Application of Measurement Error Methods

The methods were illustrated by considering a subset of the data from UABMVS. The data used consisted of n=1804 subjects with complete baseline data, where m=88 at-fault crashes occurred. The median time on the study for this subset of data was 5.15 years with corresponding range (0.01, 6.07). For ease of analysis, all data were standardized.

Covariates Under Study

Three bivariate models including the error prone covariate, annual mileage, and an error free continuous covariate found to be predictive of at-fault motor vehicle crashing were used. The error free covariates of interest were participant age at assessment, denoted AGE, TRAILS and UFOV®. The hazards can be written as

$$\lambda\left(t \mid AGE, MILES\right) = \lambda_0\left(t\right)\exp\left(\beta_X AGE + \beta_Z MILES\right); \tag{5.1}$$

$$\lambda\left(t \mid UFOV, MILES\right) = \lambda_0\left(t\right)\exp\left(\beta_X UFOV + \beta_Z MILES\right); \tag{5.2}$$

$$\lambda\left(t \mid TRAILS, MILES\right) = \lambda_0\left(t\right)\exp\left(\beta_X TRAILS + \beta_Z MILES\right). \tag{5.3}$$

Annual mileage, which is frequently subject to measurement error, is commonly used as a descriptor of driving exposure. Measurement error variance was estimated using two estimates of annual mileage. The estimates of annual mileage were treated as replicate measurements, $\left(W_{i1}, W_{i2}\right)$, where the classic error model holds: $W_{ij} = X_i + U_{ij},\ i = 1,\ldots,n;\ j = 1,2.$ The first estimate was defined as the midpoint of the categorical representation of annual mileage in the Self Report Mobility Questionnaire.

The second estimate was based on the value of weekly miles driven. From the weekly miles driven estimate, an annual mileage estimate was determined as follows:

$$\frac{\text{MILES}}{\text{WEEK}} \times \frac{52.18}{\text{YEAR}}.$$

Results

Several subjects were omitted from this analysis due to missing baseline values. This reduced the number from approximately 1900 subjects to N=1804 for this analysis. Of this, there were m=88 at-fault motor vehicle crashes as provided by MMVA.

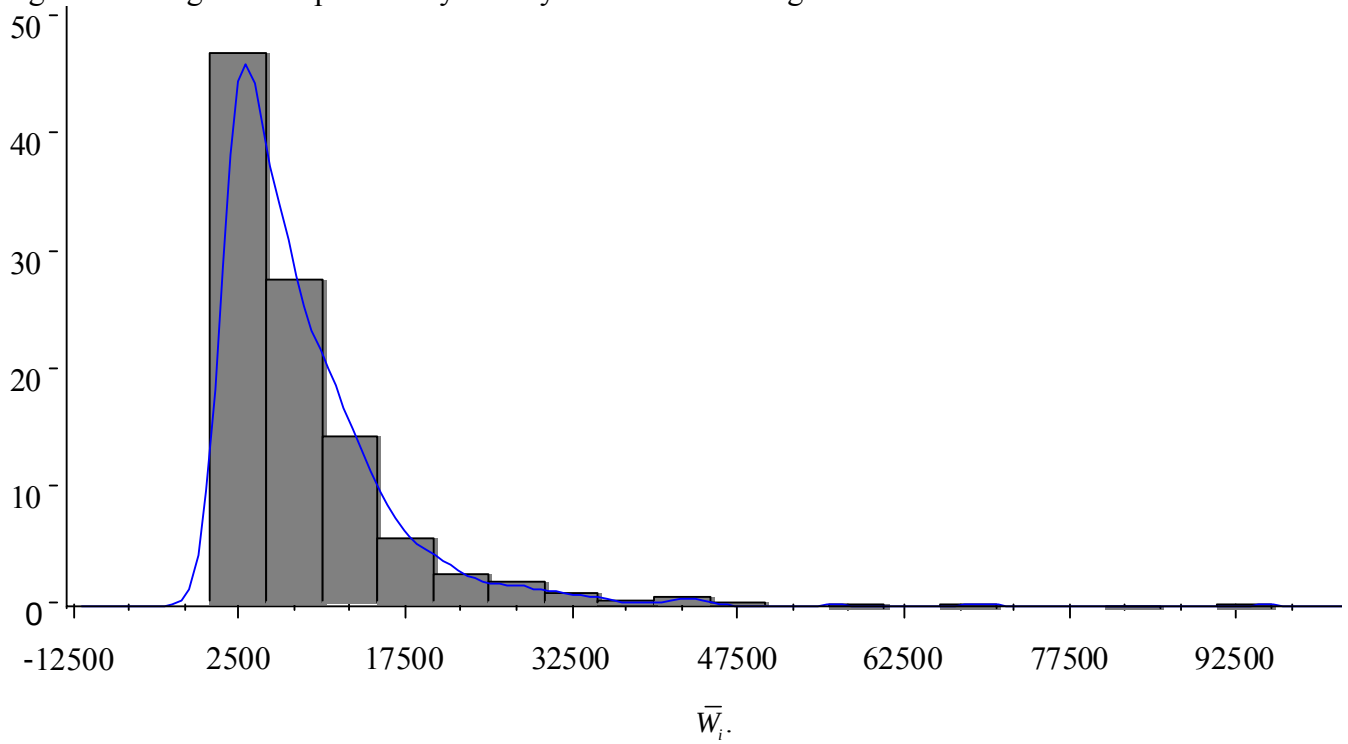
Figure 1 shows the baseline density of annual mileage using the naïve estimate

$$\bar{W}_{i\cdot} = \frac{W_{i1} + W_{i2}}{2},$$

the average of the replicates. Approximately 75% of the participants' average annual mileage was less than 10,734 miles. The measurement error was estimated using (2.14). Measurement error along with the correlation between annual miles and the prognostic variable of interest was considered small to moderate. Specifically, the estimated measurement error variance was $\sigma_U^2 = 0.311$ and the correlation between miles and age, UFOV® and TRAILS was -0.28, -0.17 and -0.15 respectively. These values correspond closely to those used in the simulation study. Measurement error and correlation in this example ranges from small to moderate.

Under models (5.1)-(5.3), comparisons of the naïve method to that of RCCORR and RSCCORR are found in Tables 19-21. The results for model (5.1) are found in Table 19. The estimated coefficient for mileage $\hat{\beta}_X$, is 0.338, 0.342, and 0.432 for the Naïve Approach, RCCORR and RSCCORR respectively. The corresponding standard errors are

Figure 1 Histogram and probability density function of Mileage estimate



* $\bar{W}_i.$, Average of annual mileage replicates; vertical axis represents percentage.

0.078, 0.167, and 0.190. The Naïve approach may underestimate the effect of driving exposure on time to an at fault motor vehicle crash. Accounting for correlation and measurement error via RCCORR and RSCCORR increase the effect sizes by 1.18% and 27.8% respectively. The results in the error free variable are not as dramatic yet follow a similar pattern. The estimated coefficient for age, $\hat{\beta}_Z$, is 0.168 when correlation and measurement error is ignored. The corresponding standard error (SE) is 0.015. The estimate resulting for RCCORR is similar to the naïve estimate and RSCCORR increases the parameter estimate by 13.1% from 0.168 to 0.190.

Estimates for $\hat{\beta}_X$, $\hat{\beta}_Z$ and their standard errors, under model (5.2), are found in Table 20. Using the Naïve method, the estimated coefficient for annual mileage is 0.331 with corresponding standard error 0.076. The estimate obtained from RCCORR is 0.335 and 0.419 from RSCORR with standard errors of 0.077 and 0.096 respectively.

The final model under discussion is (5.3). The estimated coefficient for MILES using the naïve method is 0.310 with standard error 0.076. The relative risk estimates based on the RCCORR estimates were 1.367 and 1.103 for $\hat{\beta}_X$, $\hat{\beta}_Z$ respectively. Additionally, the relative risk estimates for RSCCORR were 1.478 and 1.114. The RSCCORR relative risk estimates are 8.44% and 1.0% larger than the naïve estimate. The estimates for RCCORR are intermediate to that.

Discussion

In order to use either method, RCCORR or RSCCORR, the rare event assumption must hold. In this data set, there were 88 events; consequently, less than 5% of the

subjects experienced an at fault motor vehicle crash. The assumption of normal measurement error was investigated graphically.

$$W_{i1} - W_{i2} = \left(X_i + U_{i1}\right) - \left(X_i + U_{i2}\right) = U_{i1} - U_{i2}. \tag{5.4}$$

Under the assumption of normal measurement error, it follows that $W_{i1} - W_{i2}$ should follow a normal distribution. Figure 2 shows the density of $W_{i1} - W_{i2}$. The graph exhibits some skewness. This indicates that the error may not be normally distributed. Also, Figure 3 and Figure 4 are kernel estimates of the densities of W_{i1} and W_{i2}. Similarly, these plots too exhibit some skewness. It appears that annual miles may not be normally distributed.

The number of replicates should be taken into consideration in this setting. In the simulation study, we used three replicates which provided a strong estimate of measurement error and also allowed us to estimate parameters fairly well when correlation and measurement error was small. Replicates provide additional information in terms of estimating measurement error and the calibrated estimate of true *X*. Two or more replicates are sufficient when conducting regression calibration methods. Yet, it is important to note that observing only two replicates does not provide as much information as three or more. Small numbers of replicates, like two, reduces our ability to accurately asses the amount of measurement error. In this setting, the quality of resulting parameter estimates may not be as good as compared to observing more replicates.

It is evident from Tables 19 through 21 that failure to account for correlation and measurement error does not change the determination that MILES and time to an at-fault motor vehicle crash have an association. For each bivariate model, Wald statistics for

$\beta_X = 0$ are significant at the 0.05 level. For the model with UFOV as an error free covariate, Wald statistics for $\beta_Z = 0$ are significant. Alternatively, the remaining models show that AGE and TRAILS are not significant. However, the goal of this study was to show that correcting for correlation and measurement error will provide an accurate picture of the relationship between the covariates of interest and failure time. It is clear from each table in this section that failure to correct for correlation and measurement error may result in attenuated parameter estimates.

Table 19 Comparison of estimators $\hat{\beta}_X$, $\hat{\beta}_Z$ in the UABMVS Example; Sample size *n*=1804; SE is estimated standard error.

Method	Annual Mileage	Age
Naïve	0.338	0.168
SE	0.078	0.115
Regression Calibration	0.342	0.167
SE	0.167	0.115
Risk Set Calibration	0.432	0.190
SE	0.190	0.116

*Naïve regression based on average of observed replicates

* $\hat{\sigma}_u^2$ is estimated using replicate mileage estimates

Table 20 Comparisons of estimators $\hat{\beta}_X$, $\hat{\beta}_Z$ in the UABMVS Example; Sample size n=1804; SE is estimated standard error.

Method	Annual Mileage	Useful Field of View
Naïve	0.331	0.220
SE	0.076	0.099
Regression Calibration	0.335	0.219
SE	0.077	0.099
Risk Set Calibration	0.419	0.233
SE	0.096	0.100

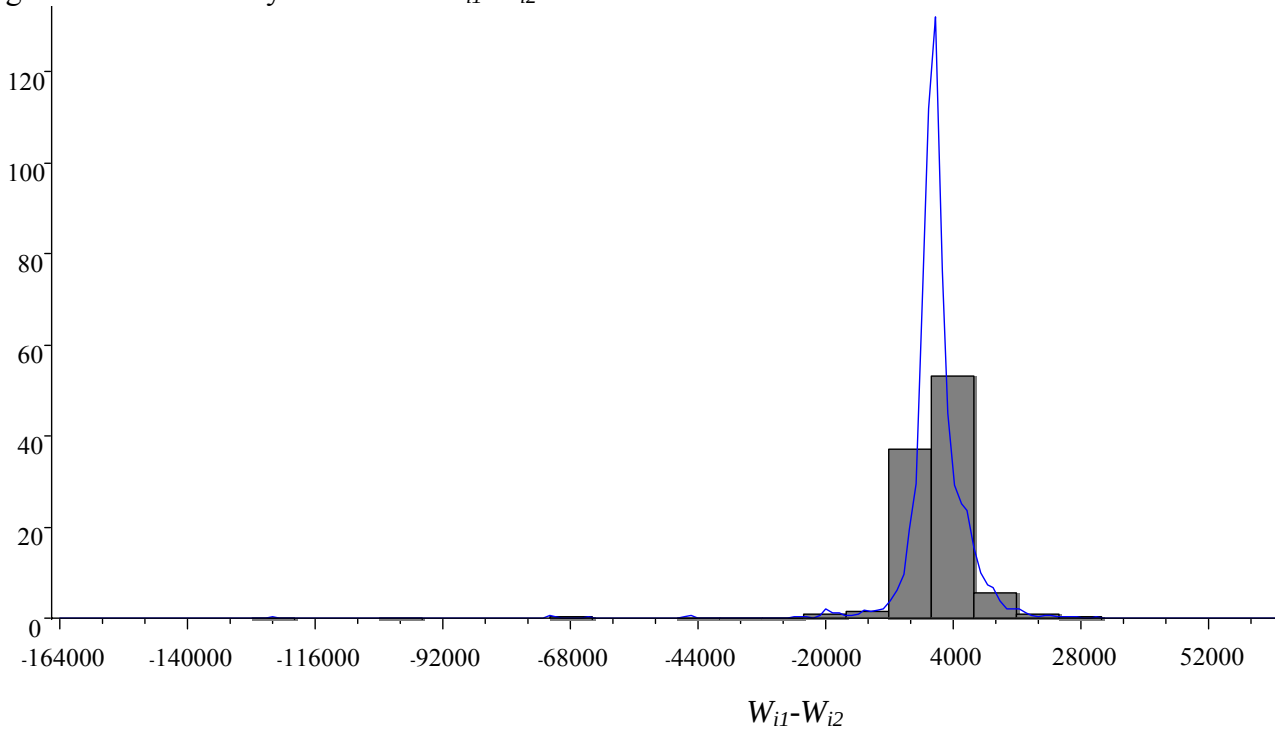
*Naïve regression based on average of observed replicates

* $\hat{\sigma}_u^2$ is estimated using replicate mileage estimates

Table 21 Comparison of estimators $\hat{\beta}_X$, $\hat{\beta}_Z$ in the UABMVS Example; Sample size n=1804; SE is estimated standard error.

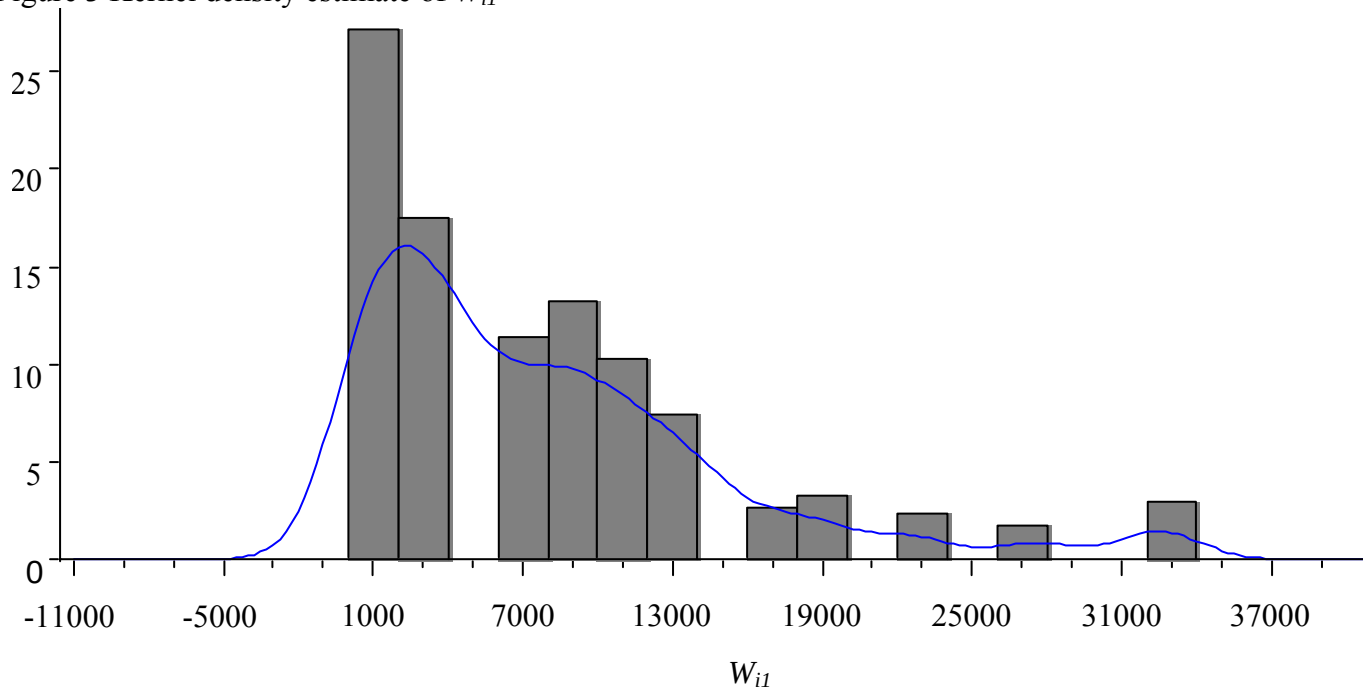
Method	Annual Mileage	TRAILS
Naïve	0.310	0.098
SE	0.076	0.095
Regression Calibration	0.314	0.098
SE	0.077	0.095
Risk Set Calibration	0.391	0.108
SE	0.096	0.096

Figure 2 Kernel density estimate of W_{i1}-W_{i2}



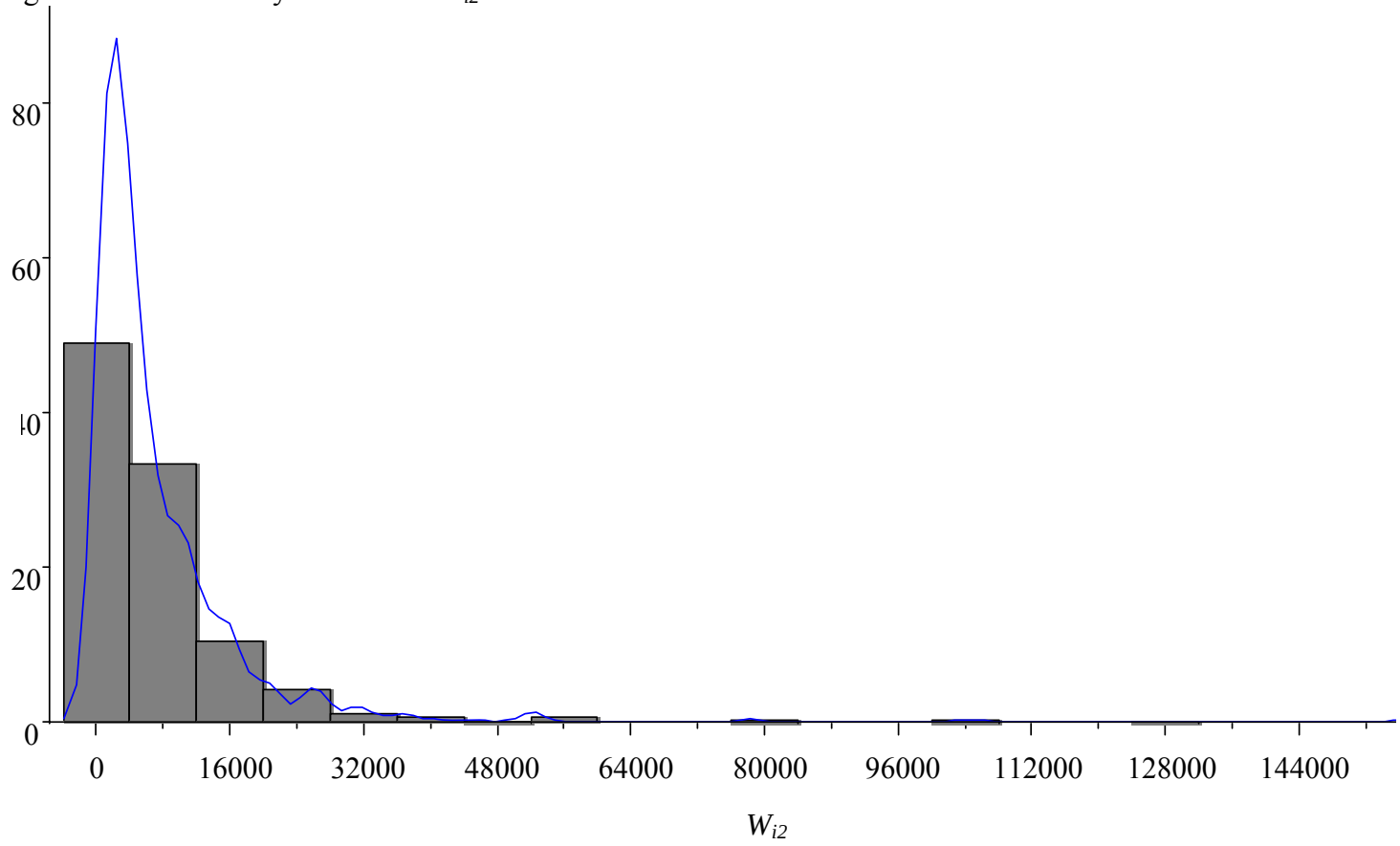
Vertical Axis represents percentage.

Figure 3 Kernel density estimate of W_{i1}



Vertical axis represents percentage.

Figure 4 Kernel density estimate of W_{i2}



Vertical Axis represents percentage.

CHAPTER 6

CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The problem under investigation in this dissertation was the assessment and correction of the effect of correlation and measurement error when conducting Cox proportional hazards modeling. More specifically, the research area was a comparative study of regression calibration and risk set calibration and their subsequent behavior and ability to adequately handle various levels of correlation and measurement error in parameter estimation. Here, the results are discussed. In addition, suggestions for further research, as well as specific recommendations for the replication of this research are provided.

Summary of Study

In Chapter 4, a simulation study was performed to compare failure to correct for correlation and measurement error to regression calibration and risk set calibration methods. More specifically, regression calibration not adjusted for correlation, regression calibration corrected for correlation, risk set calibration corrected for correlation and risk set calibration corrected for correlation and adjusted for small risk sets were compared to the Naïve approach, as well as to one another. The rationale for conducting this study was to add to the body of research concerning methods recommended for the analysis of continuous, correlated covariates where one covariate is error prone. The goal of this study was to investigate the quality of parameters produced in this setting and methods

used to improve the estimates in this setting. Covariate data consisting of bivariate normal correlated observations, survival times and event indicators were generated. Parameter estimation was made using the Naïve method, Regression calibration adjusted and not adjusted for correlation along with risk set calibration methods. The estimates derived via these methods were studied to assess whether the methods could correct the resultant attenuation effects.

As a result of the simulation study, it was observed that regression calibration methods and risk set calibration methods perform satisfactorily at low levels of correlation and measurement error. RCNO tended to outperform RCCORR at the lowest levels of correlation within a measurement error class. As correlation increased, RCCORR, which includes a correction for correlation would provide most accurate estimates yet slightly less efficient variable. The risk set calibration methods provided the least efficient variable estimates at each design point. At each design point, RSCADJ outperformed RSCCORR producing more consistent estimates yet each performed poorly in comparison to the regression calibration methods.

There are some conclusions which can be made about the methods under investigation. Overall, it is better to make a correction for the correlation/measurement error problem versus ignoring the problem. Regression calibration methods work well at low to moderate levels of correlation/measurement error. More specifically, at the lowest level of correlation, simply adjusting for measurement error provides reasonable results. Yet, additionally incorporating an adjustment for correlation improves estimates as correlation increases. The risk set calibration methods too provide reasonable results at the lowest level of correlation within a level of measurement error but the methods are

unstable. The instability is a result of small risk sets since at each event time, the method calibrates the error prone variable within the risk set to provide a more accurate estimate of the true, unobserved variable. This becomes problematic as the risk sets become small. Part of the method includes taking the inverse of the variance. These estimates, due to size of risk set, can become very large or small which produce calibrated estimates that are unreasonable causing the method to fail to converge. In order to adjust for this problem, RSCADJ was developed. This method does not calibrate the observations of the last twenty subjects remaining in the risk set. The average of the error prone estimates is used in its place. As a result, RSCADJ provided better estimates than RSCCORR yet these estimates generally had large MSE values.

In summary, regression calibration methods used to correct for correlation and measurement error can be sufficiently used when correlation and or measurement error is no greater 0.3. These methods can be easily implemented by a statistician of varied experience levels. When the measurement error/correlation combination is less than 0.3 and one has at least three replicates, the method works well at adjusting for the combination. Regression calibration will provide estimates that have small absolute bias and retain robust levels of coverage. When correlation or measurement error is greater than 0.3, different methods should be considered.

Suggestions for Future Research

Computing Resources

Using sing SAS IML, the simulation study was performed on a Pentium 4

computer with 1.25 GB of RAM and 37.5 GB hard drive. The original simulation took upwards of 2 hours to run. As the amount of correlation and error was increased, the amount of time to complete each design point run increased. This time expense can be eliminated with use of more efficient programming languages such as FORTRAN. The FORTRAN language would be useful because of its ability to handle mathematically complex functions. The compiler allows for intricate mathematical computing. Using a more efficient programming language would allow the investigation to look at mathematically dense methods to correct for correlation and measurement error. Also, FORTRAN allows for use of subroutines which would reduce the optimization time for finding parameter estimates.

Parameter Values

In this study, $\beta_X = 1$ and $\beta_Z = 1$ were used as the values of the true covariates. While this proved to be an interesting starting point, different parameter values can be investigated. A wide range of parameter values could be used. This examination would provide an indication of whether the parameter estimates influence the performance of the methods.

Censoring Mechanisms

Although censoring rates ranged from 35% to 45% for each simulation run, lower and higher amounts can be investigated. This investigation may provide an indication whether the amount of censoring truly affects the performance of the calibration methods empirically.

Model Specification

In the simulation study, the classic error model was used to describe the error prone estimate. While, this parameterization worked well in the simulation, different parameterizations can be explored. We observed in the application of the methods to the real data, the classic error assumption may not be appropriate. It follows that the ability of the methods to produce reasonable quality estimates under different parameterizations would be beneficial to the body of research.

Any of these suggestions has the ability to make a useful contribution to the body of research related to correcting the effects of the correlation/measurement error problem which commonly occurs in social science research settings.

LIST OF REFERENCES

[1] Klein JP, Moeschberger ML. *Statistics for Biology and health: Survival Analysis*. Second ed. New York: Springer -Verlag, 1997.

[2] Lawless JF. *Statistical Models and Methods for Lifetime Data*. 2nd ed: John Wiley & Sons, 2002.

[3] Andersen PK, Gill RD. Cox's Regression Model for counting processes: a large sample study. *The Annals of Statistics*. 1982;**10**: 1100-1120.

[4] Buzas JS. Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference*. 1998;**67**: 247-257.

[5] Fraser GE, Stram DO. Regression Calibration in Studies with Correlated Variables Measured with Error. *Am J Epidemiol*. 2001;**154**: 836-844.

[6] Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*: Chapman and Hall/CRC Press, 2003.

[7] Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models: A Modern Perspective*. Second ed. London: Chapman and Hall, 1995.

[8] Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B*. 1972;**34**: 187-220.

[9] Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. 1 ed: Wiley-Interscience, 1991.

[10] Miller RG. *Survival Analysis*. 1st ed: John Wiley & Sons Inc, 1997.

[11] Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd ed: Wiley Series in Probability and Mathematical Statistics, 2002.

[12] Fuller WA. *Measurement Error Models*: Wiley, 1987.

[13] Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*. 2006;**13**: 1265 - 1282.

[14] Neter J, Kutner MH, Wasserman W, Nachtsheim CJ. *Applied Linear Statistical Models*. 4th ed: Publisher: McGraw-Hill/Irwin, 1996.

[15] Prentice RL. Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika*. 1982;**Vol. 69**: 331-342.

[16] Zhou H, Pepe MS. Auxiliary Covariate Data in Failure Time Regression. *Biometrika*. 1995;**82**: 139-149.

[17] Zhou H, Wang CY. Failure Time Regression with Continuous Covariates Measured with Error. *Journal of the Royal Statistical Society Series B*. 2000;**62**: 657-665.

[18] Wang CY, Hsu L, Feng ZD, Prentice RL. Regression Calibration in Failure Time Regression. *Biometrics*. 1997;**53**: 131-145.

[19] Song X, Huang Y. On Corrected Score Approach for Proportional Hazards Model with Covariate Measurement Error. *UW Biostatistics Working Paper Series Working Paper 226*. 2004.

[20] Clayton D. *Models for the Longitudinal Analysis of Cohort and Case-Control Studies with Inaccurately Measured Exposures*. New York: Oxford Press, 1992.

[21] Cook JR, Stefanski LA. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*. 1994;**89**: 1314-1328.

[22] Stefanski LA, Cook JR. Simulation-Extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*. 1995;**90**: 1247-1256.

[23] Nakamura T. Proportional Hazards Model with Covariates Subject to Measurement Error. *Biometrics*. 1992;**48**: 829-838.

[24] Nakamura T. Corrected Score Function for Errors-in-Variables Models: Methodology and Application to Generalized Linear Models. *Biometrika*. 1990;**77**: 127-137.

[25] Kong FH, Gu M. Consistent Estimation in Cox Proportional hazards model with covariate measurement errors. *Statistica Sinica*. 1999;**9**: 953-969.

[26] Hu C, D.Y. L. Cox Regression with Covariate Measurement Error. *Scandinavian Journal of Statistics*. 2002;**29**: 637-655.

[27] Hu P, Tsiatis AA, Davidian M. Estimating the Parameters in the Cox Model When Covariate Variables Are Measured with Error. *Biometrics*. 1998;**54**: 1407-1419.

[28] Lin DY, Ying Z. Cox Regression with Incomplete Covariate Measurements. *Journal of the American Statistical Association*. 1993;**88**: 1341-1349.

[29] Paik MC, Tsai W-Y. On Using the Cox Proportional Hazards Model with Missing Covariates. *Biometrika*. 1997;**84**: 579-593.

[30] Chen HY, Little RJA. Proportional Hazards Regression with Missing Covariates. *Journal of the American Statistical Association*. 1999;**94**: 896-908.

[31] Lin DY, Wei LJ. The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*. 1989;**84**: 1074-1078.

[32] Xie SX, Wang CYW, Prentice RL. A Risk Set Calibration Method for Failure Time Regression by Using a Covariate Reliability Sample. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2001;**63**: 855-870.

[33] Lagakos SW. General Right Censoring and its Impact on the Analysis of Survival Data. *Biometrics*. 1979;**35**: 139-156.

[34] Leung K-M, Elashoff RM, Afifi AA. Censoring Issues In Survival Analysis. *Annual Review of Public Health*. 1997;**18**: 83-104.

[35] Liu Q, Pierce DA. A Note on Gauss-Hermite Quadrature (in Miscellanea). *Biometrika*. 1994;**81**: 624-629.

[36] Davis PJ, Rabinowitz P. *Methods of numerical integration*. 2nd ed. Orlando: Academic Press, 1984.

[37] Abramowitz M, Stegun IA. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. New York: Dover Publications, 1965.

[38] Davidian M, Gallant AR. The Nonlinear Mixed Effects Model with a Smooth Random Effects Density.**80**: 475-488.

[39] Liu K, Mazumdar S, Stone RA, Dew MA, Houck PR, Reynolds IIICF. Accounting for covariate measurement error in a Cox model analysis of recurrence of depression. *Journal of Psychiatric Research*. 2001;**35**: 177.

[40] Burton A, Altman DG, Royston P, Holder RL. The Design of Simulation Studies in Medical Statistics. *Statistics in Medicine*. 2006;**25**: 4279-4292.

[41] Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed: Lawrence Erlbaum, 1988.

[42] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*. 2005;**24**: 1713-1723.

[43] Casella G, Berger RL. *Statistical Inference*. 2nd ed: Duxbury Press, 2001.

[44] Ball KK, Roenker DL, Wadley VG*, et al.* Can High-Risk Older Drivers Be Identified Through Performance-Based Measures in a Department of Motor Vehicles Setting? *Journal of the American Geriatrics Society*. 2006;**54**: 77-84.

[45] Vance DE, Roenker DL, Cissell GM, Edwards JD, Wadley VG, Ball KK. Predictors of driving exposure and avoidance in a field study of older drivers from the state of Maryland. *Accident Analysis & Prevention*. 2006;**38**: 823.

[46] Goode KT, Ball KK, Sloane M*, et al.* Useful Field of View and Other Neurocognitive Indicators of Crash Risk in Older Adults. *Journal of Clinical Psychology in Medical Settings*. 1998;**5**: 425-440.

[47] Roenker DL, Cissell GM, Ball KK, Wadley VG, Edwards JD. Speed-of-Processing and Driving Simulator Training Result in Improved Driving Performance. *Human Factors*. 2003;**45**: 218-233.

[48] Richardson ED, Marottoli RA. Visual Attention and Driving Behaviors Among Community-Living Older Persons. *Journal of the Gerontology*. 2003;**58**: 832-836.

[49] Hu PS, Trumble DA, Foley DJ, Eberhard JW, Wallace RB. Crash risks of older drivers: a panel data analysis. *Accident Analysis & Prevention*. 1998;**30**: 569.

[50] Marottoli RA, Cooney LM, Wagner DR, Docette J, Tinetti MEl. Predictors of Automobile Crashes and Moving Violations among Elderly Drivers. *Annals of Internal Medicine*. 1994;**121**: 842-846.

APPENDIX A

IRB APPROVAL

UAB THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Institutional Review Board for Human Use

Form 4: IRB Approval Form
Identification and Certification of Research
Projects Involving Human Subjects

UAB's Institutional Review Boards for Human Use (IRBs) have an approved Federalwide Assurance with the Office for Human Research Protections (OHRP). The UAB IRBs are also in compliance with 21 CFR Parts 50 and 56 and ICH GCP Guidelines. The Assurance became effective on November 24, 2003 and expires on February 14, 2009. The Assurance number is FWA00005960.

Principal Investigator:	DUBE, TINA J
Co-Investigator(s):	
Protocol Number:	**X061221005**
Protocol Title:	*Doctoral Dissertation: Assessing and Correcting the Effects of Measurement Error Among Correlated Covariates in a Proportional Hazards Setting*

The IRB reviewed and approved the above named project on 04/18/07. The review was conducted in accordance with UAB's Assurance of Compliance approved by the Department of Health and Human Services. This Project will be subject to Annual continuing review as provided in that Assurance.

This project received EXPEDITED review.

IRB Approval Date: 4-18-07

Date IRB Approval Issued: 04/18/07

Marilyn Doss

Marilyn Doss, M.A.
Vice Chair of the Institutional Review
Board for Human Use (IRB)

Investigators please note:

The IRB approved consent form used in the study must contain the IRB approval date and expiration date.

IRB approval is given for one year unless otherwise noted. For projects subject to annual review research activities may not continue past the one year anniversary of the IRB approval date.

Any modifications in the study methodology, protocol and/or consent form must be submitted for review and approval to the IRB prior to implementation.

Adverse Events and/or unanticipated risks to subjects or others at UAB or other participating institutions must be reported promptly to the IRB.

470 Administration Building
701 20th Street South
205.934.3789
Fax 205.934.1301
irb@uab.edu

The University of
Alabama at Birmingham
Mailing Address:
AB 470
1530 3RD AVE S
BIRMINGHAM AL 35294-0104

APPENDIX B

SIMULATION PROGRAMS

```
/*********************************************************************/
/*Author:   Tina Dube                                                */
/*Data Generation                                                    */
/*This program generates bivariate normal covariate data with a      */
/*prespecified amount of correlation parameter. Additionally,        */
survival data                                                        */
/*is generated based on true parameter estimates Betax=1 and Beta_Z=1*/
/*Co-Authored by http://www.biostat.umn.edu/~john-c/5421/notes.019   */
/*********************************************************************/
libname trial 'u:\Sas Library';

DATA MEANS;
      INPUT  MEAN @@;
      CARDS;
0 0 0 0 0
;
RUN;

%macro datagen(N,cohorts,CORRXZ,ERRORU);
%do X = 1 %to &cohorts;
      proc iml symsize=7000000 worksize=800000;
      reset storage=trial.catalog3;

            SIGMA={ 1.00 &CORRXZ 0.00 0.00 0.00,
            &CORRXZ 1.00 0.00 0.00 0.00,
            0.00 0.00 &ERRORU 0.00 0.00,
            0.00 0.00 0.00 &ERRORU 0.00,
            0.00 0.00 0.00 0.00 &ERRORU};
*PRINT SIGMA;*DEFINE THE MVN(X,Z,U);

            USE MEANS; *READ MEAN VECTOR;
            READ ALL INTO MU; *PRINT MU;
            P=NROW(SIGMA); *PRINT P; *CALCULATE NUMBER OF VARIABLES;
            CALL VNORMAL(DATAMATRIX,MU,SIGMA,&N,&X);
*PRINT DATAMATRIX;
            *GENERATE DATA MATRIX CALLED DATAMATRIX OF LENGTH N
            MU IS THE MEAN VECTOR AND SIGMA THE CORRESPONDING
COVARIANCE
            MATRIX &J IS THE SEED;
            CNAME={"Z" "XTRUE" "U1" "U2" "U3"}; *GENERATE DATA SET;
            CREATE EXPECT FROM DATAMATRIX [COLNAME=CNAME];
            APPEND FROM DATAMATRIX;
            *SHOW CONTENTS; *SHOW CONTENTS OF DATA SET;

            use EXPECT;
            read all ;
            U=U1||U2||U3; *PRINT U;*measurement error;
            K=NCOL(U); *PRINT K;*NUMBER OF REPLICATES;
            Z=Z; *PRINT Z;*error-free covariate;
            XTRUE=XTRUE; *PRINT XTRUE;*error prone covariate;
            X=REPEAT(xtrue,1,K); *PRINT X;
            W=X + U;

            *GENERATE THE REMAINDER OF DATA;
            DELTA=J(&N,1,.);
*DELTA indicator;
```

```
            time=J(&N,1,.);
*failure time;
            censor=J(&N,1,.);
*censoring time;
            time=J(&N,1,.);
*min(failure,censor);
            runi =J(&N,1,.);
*USED TO GENERATE SURVIVAL TIMES;
            runi2=J(&N,1,.);

            %do i = 1 %to &N;
*GENERATE FAiLURE AND CENSORiNG TiMES;
            runi[&i]=ranuni(123);
                runi2[&i]=ranuni(1234);
                time[&i]=-log(runi[&i])/(exp(z[&i]+XTRUE[&i]));
*Generate Survival times using beta1=1 and beta2=1;
                censor[&i]=-log(runi2[&i]);

                  if censor[&i] < time[&i]
                then do;
                    time[&i]=censor[&i];
                    DELTA[&i]=0;
                end;

                else do;
                    time[&i]=time[&i];
                    DELTA[&i]=1;
                end;
            %end;
*END I LOOP FOR GENERATING FAILURE AND CENSORING TIMES

      INSIDE PROC IML;

            ORIGDATA2=DELTA||TIME||XTRUE||W||Z;
                  *print ORIGDATA2;
            CNAME1={"DELTA" "time"  "X" "W1"  "W2"  "W3" "Z"  };
            CREATE trial.DTA_0105_2a&x FROM ORIGDATA2 [COLNAME=CNAME1];
            APPEND FROM ORIGDATA2;

            *SHOW CONTENTS; *SHOW CONTENTS OF DATA SET;
            show storage;

            store DELTA TIME W X XTRUE Z;

            cohortnumber=&x;
            print, "Cohort number:  " cohortnumber;

            *remove xtrue CENSOR CNAME CNAME1 DATAMATRIX
            K MU  ORIGDATA2
            P RUNI RUNI2 SIGMA U U1 U2 U3 XTRUE;

            free DELTA TIME W X  Z;

      QUIT;                                                    *END
PROC IML;
      RUN;
```

```
proc sort data=trial.DTA_0105_2a&x;
    by time;
run;

/*proc print data=trial.DTA_0105_2a&x;
run;*/
%end;
%mend datagen;
%datagen(150,1060,0.3,0.1);
*CALL THE MACRO;
```

```
/*****************************************************************/
/*Author:   Tina Dube                                            */
/*Module to generate the regression calibration estimates of the */
/*error prone covariate. This module generates the individual    */
/*variance components in order to complete the calibration method.*/
/*This module does not correct for correlation                   */
/*****************************************************************/

*options symbolgen mlogic;
proc iml symsize=7000000 worksize=800000;
start rc_mod(w ,wbar, z ,delta, time,rcmatrix);
n=100;*PRINT W WBAR Z DELTA TIME ;
riskind=j(n,n,.);
MUHATWATT=J(1,1,.);
MUHATZATT=J(1,1,.);
SIGMA_xz=J(1,1,.);
MUHATW=J(1,1,.);
MUHATZ=J(1,1,.);
sigmax=j(1,1,.);
sigmaxz=j(1,1,.);
sigmaz=j(1,1,.);
VAR_MAT=j(2,2,.);
MULT_MATRIX1=j(1,2,.);
MULT_MATRIX2=j(2,1,.);
RCMATRIX=J(N,1,.);ONEVEC=J(N,1,1);ONEVEC=ONEVEC`;*PRINT ONEVEC;
k=3;
wbarmat=repeat(wbar,1,3);
wbar_w=w-wbarmat;
/*****************************************************************/;
do I = 1 to N;
    do J = 1 to N;
        if time[i] <= time[j]  then riskind[i,j]=1 ;
        else if time[i] > time[j] then riskind[i,j]=0;
    end;
end;
/***************************calculate sigma u***************/;
sigma_u=ssq(wbar_w)/n*(k-1);

*print sigma_u;
/*************calculate mu hat x ***********************************/;
MUHATW=ONEVEC*WBAR/n;
muhatx=muhatw;*print muhatx; *print muhatw;
muhatwmat=j(n,1,muhatw);*print muhatwmat;
/******************calculate mu hat z ****************************/;
MUHATZATT=ONEVEC*Z/n;*print MUHATZATT;
muhatz=z`[,:];*print muhatz;
muhatzmat=j(n,1,muhatz);*print muhatzmat;

*print MUHATZATT;
/************************calculate sigma XZ ***********************/;
wbardiff=wbar-muhatwmat;
zbardiff=z-muhatzmat;
sigmaxz1=(wbardiff#zbardiff)[##];
sigmaxz=sigmaxz1/(n-1);*print sigmaxz;
/*********************calculate sigma x ***************************/;
c=sigma_u/k;
sigmax1=(wbardiff#wbardiff)[##];
```

```
sigmax=sigmax1/(n-1)-c;*print Sigmax;
/************************calculate sigma Z **********************/;
sigmaz1=(zbardiff#zbardiff)[##];
sigmaz=sigmaz1/(n-1);
**********************calibrated  values**************************/;
    VAR_MAT[1,1]=c+SIGMAX;
    VAR_MAT[1,2]=0;
    VAR_MAT[2,1]=0;
    VAR_MAT[2,2]=SIGMAZ;      *PRINT VAR_MAT;
    INVVAR_MAT=GINV(VAR_MAT);
    MULT_MATRIX1[1,1]=SIGMAX ;
      MULT_MATRIX1[1,2]=0 ;                *print MULT_MATRIX1;

    do i = 1 to N;
      *subject index;
        MULT_MATRIX2[1,1]=WBAR[i]- muhatwmat[i];
        MULT_MATRIX2[2,1]=Z[i]-muhatzmat [i];
        RCMATRIX[i]=muhatwmat[i]+MULT_MATRIX1*INVVAR_MAT*MULT_MATRIX2;
    END;

      *PRINT RCMATRIX;
*SHOW CONTENTS; *SHOW CONTENTS OF DATA SET;
 /*DM "Output; Clear; Log; Clear";*/

finish rc_mod;

reset storage=trial.catalog2;
store module=rc_mod;
quit;
run;

proc iml;
      reset storage=trial.catalog2;
      show storage;
quit;
run;
proc iml;
      reset storage=trial.catalog2;
      show storage;
            remove module=rc_mod;
            show storage;
quit;
run;
```

```
/*****************************************************************/
/*Author:   Tina Dube                                            */
/*Module to generate the regression calibration estimates of the */
/*error prone covariate. This module generates the individual    */
/*variance components in order to complete the calibration method.*/
/*Correction for correlation has been implemented into this module*/
/*****************************************************************/
*options symbolgen mlogic;
proc iml symsize=7000000 worksize=800000;

start rc_mod(w ,wbar, z ,delta, time,rcmatrix);

n=150;*PRINT W WBAR Z DELTA TIME ;
riskind=j(n,n,.);
MUHATWATT=J(1,1,.);
MUHATZATT=J(1,1,.);
SIGMA_xz=J(1,1,.);
MUHATW=J(1,1,.);
MUHATZ=J(1,1,.);
sigmax=j(1,1,.);
sigmaxz=j(1,1,.);
sigmaz=j(1,1,.);
VAR_MAT=j(2,2,.);
MULT_MATRIX1=j(1,2,.);
MULT_MATRIX2=j(2,1,.);
RCMATRIX=J(N,1,.);ONEVEC=J(N,1,1);ONEVEC=ONEVEC`;*PRINT ONEVEC;
k=3;
wbarmat=repeat(wbar,1,3);
wbar_w=w-wbarmat;
/*****************************************************************/
do I = 1 to N;
    do J = 1 to N;
        if time[i] <= time[j]  then riskind[i,j]=1 ;
        else if time[i] > time[j] then riskind[i,j]=0;
    end;
end;
/*********calculate sigma u*****************************************/
sigma_u=ssq(wbar_w)/n*(k-1);

*print sigma_u;
/*****calculate mu hat x ***************************************/;
MUHATW=ONEVEC*WBAR/n;
muhatx=muhatw;*print muhatx; *print muhatw;
muhatwmat=j(n,1,muhatw);*print muhatwmat;
/************************calculate mu hat z ************************/;
MUHATZATT=ONEVEC*Z/n;*print MUHATZATT;
muhatz=z`[,:];*print muhatz;
muhatzmat=j(n,1,muhatz);*print muhatzmat;

*print MUHATZATT;
/********************calculate sigma XZ ****************************/;
wbardiff=wbar-muhatwmat;
zbardiff=z-muhatzmat;
sigmaxz1=(wbardiff#zbardiff)[##];
sigmaxz=sigmaxz1/(n-1);*print sigmaxz;
/**********calculate sigma x ***************************************/;
c=sigma_u/k;
```

```
sigmax1=(wbardiff#wbardiff)[##];
sigmax=sigmax1/(n-1)-c;*print Sigmax;
/**************calculate sigma Z ***************************/;
sigmaz1=(zbardiff#zbardiff)[##];
sigmaz=sigmaz1/(n-1);
***************calibrated  values*****************************/;
    VAR_MAT[1,1]=c+SIGMAX;
    VAR_MAT[1,2]=SIGMAXZ;
    VAR_MAT[2,1]=SIGMAXZ;
    VAR_MAT[2,2]=SIGMAZ;      *PRINT VAR_MAT;
    INVVAR_MAT=GINV(VAR_MAT);
    MULT_MATRIX1[1,1]=SIGMAX ;
      MULT_MATRIX1[1,2]=SIGMAXZ ;                *print MULT_MATRIX1;

    do i = 1 to N;
      *subject index;
        MULT_MATRIX2[1,1]=WBAR[i]- muhatwmat[i];
        MULT_MATRIX2[2,1]=Z[i]-muhatzmat[i];
        RCMATRIX[i]=muhatwmat[i]+MULT_MATRIX1*INVVAR_MAT*MULT_MATRIX2;
    END;

      *PRINT RCMATRIX;
*SHOW CONTENTS; *SHOW CONTENTS OF DATA SET;
 /*DM "Output; Clear; Log; Clear";*/

finish rc_mod;

reset storage=trial.catalog3;
store module=rc_mod;
quit;
run;

proc iml;
      reset storage=trial.catalog3;
      show storage;
quit;
run;

/*proc iml;
      reset storage=trial.catalog3;
      show storage;
            remove module=rc_mod;
            show storage;
quit;
run;
```

```
/*********************************************************************/
/*Author:   Tina Dube                                                */
/*Module to calibrated data using the risk set calibration method    */
/*Covariates are calibrated at each event time                       */
/*********************************************************************/
*options symbolgen mlogic;
proc iml symsize=7000000 worksize=800000;
start rsc_mod(w ,wbar, z ,delta, time,rscmatrix);
n=150;
riskind=j(n,n,.);
MUHATWATT=J(N,1,.);
MUHATZATT=J(N,1,.);
SIGMA_xz=J(N,1,.);
sigma_x=j(n,1,.);
sigma_z=j(n,1,.);
VAR_MAT=j(2,2,.);
MULT_MATRIX1=j(1,2,.);
MULT_MATRIX2=j(2,1,.);
RSCMATRIX=J(N,N,.);
k=3;
wbarmat=repeat(wbar,1,3);
wbar_w=w-wbarmat;
/*********************************************************************/
;
do I = 1 to N;
    do J = 1 to N;
        if time[i] <= time[j]  then riskind[i,j]=1 ;
        else if time[i] > time[j] then riskind[i,j]=0;
    end;
end;
riskindsum=riskind[,+];
riskindsum2=riskindsum;
riskindsum2[N]=2;


/*****************calculate sigma u*********************************/;
sigma_u=ssq(wbar_w)/n*(k-1);

*print sigma_u;
/************************calculate mu hat x at t*****************/;
MUHATWATT=RISKIND*WBAR/RISKINDSUM;

/****************calculate mu hat z at t**************************/;
MUHATZATT=RISKIND*Z/RISKINDSUM;

*print MUHATZATT;
/*************calculate sigma XZ at t************************/;
QR=J(N,1,.);
MUZ=REPEAT(MUHATZATT,1,N);MUZ=MUZ`;ZMAT2=REPEAT(Z,1,N);ZDIFF=ZMAT2-MUZ;
MUW=REPEAT(MUHATWATT,1,N);MUW=MUW`;WBARMAT2=REPEAT(WBAR,1,N);WDIFF=WBAR
MAT2-MUW;
Q=ZDIFF#WDIFF;
Q=Q`;
Q2=RISKIND#Q;
Q2A=Q2*J(N,1,1);

sigma_xz=Q2A/(RISKINDSUM2-1);
```

```
sigma_xz[N]=0;

*print sigma_xz;
/****************calculate sigma x at t****************************/;
QT1=J(N,1,.);
c=sigma_u/k;
CMAT2=REPEAT(C,N,1);
QS=WDIFF#WDIFF;
QS=QS`;
QT=RISKIND#QS;
QT1=QT*J(N,1,1);
sigma_x=(QT1/(RISKINDSUM2-1))-CMAT2;
sigma_x[N]=-SIGMA_U/K;

*print Sigma_x;
/******************calculate sigma Z at t************************/;
QYT1=J(N,1,.);
QY=ZDIFF#ZDIFF;
QY=QY`;
QYT=RISKIND#QY;
QYT1=QYT*J(N,1,1);
sigma_z=QYT1/(RISKINDSUM2-1);
sigma_z[N]=0;

*print sigma_z;
*********************calibrated  values*****************************/;
do j = 1 to N;
      *time index;
    VAR_MAT[1,1]=c+SIGMA_X[j];
    VAR_MAT[1,2]=SIGMA_XZ[j];
    VAR_MAT[2,1]=SIGMA_XZ[j];
    VAR_MAT[2,2]=SIGMA_Z[j];
    INVVAR_MAT=GINV(VAR_MAT);
    MULT_MATRIX1[1,1]=SIGMA_X[j] ;
      MULT_MATRIX1[1,2]=SIGMA_XZ[j] ;                *print
MULT_MATRIX1;

    do i = 1 to N;
      *subject index;
        MULT_MATRIX2[1,1]=WBAR[i]- MUHATWATT[j];
        MULT_MATRIX2[2,1]=Z[i]-MUHATZATT[j] ;              *print
VAR_MAT mult_matrix1 mult_matrix2;

RSCMATRIX[j,i]=MUHATWATT[j]+MULT_MATRIX1*INVVAR_MAT*MULT_MATRIX2;
    END;
END;
*SHOW CONTENTS; *SHOW CONTENTS OF DATA SET;
 /*DM "Output; Clear; Log; Clear";*/
*****************************************************************;
finish rsc_mod;
*****************************************************************
*****************************;
reset storage=trial.catalog3;
store module=rsc_mod;
quit;
run;
```

```
proc iml;
      reset storage=trial.catalog3;
      show storage;
quit;
run;

/*proc iml;
      reset storage=trial.catalog3;
      show storage;
            remove module=rsc_mod;
            show storage;
quit;
run;
```

```
/*********************************************************************/
/*Author:   Tina Dube                                                */
/*Module to calibrated data using the risk set calibration method    */
/*Covariates are calibrated at each event time. This module adjusts  */
/*for small risk set size by adjusting when there are twenty         */
/*subjects left in the risk set. The remaining subjects will maintain*/
/*their original data.                                               */
/*********************************************************************/
*options symbolgen mlogic;
proc iml symsize=7000000 worksize=800000;
/*********************************************************************/;
start rsc_mod_adj20(w ,wbar, z ,delta, time,rscmatrix);
/*********************************************************************/;
n=150;
riskind=j(n,n,.);
MUHATWATT=J(N,1,.);
MUHATZATT=J(N,1,.);
SIGMA_xz=J(N,1,.);
sigma_x=j(n,1,.);
sigma_z=j(n,1,.);
VAR_MAT=j(2,2,.);
MULT_MATRIX1=j(1,2,.);
MULT_MATRIX2=j(2,1,.);
RSCMATRIX=J(N,N,.);
k=3;
wbarmat=repeat(wbar,1,3);
wbar_w=w-wbarmat;
/*********************************************************************/;
do I = 1 to N;
    do J = 1 to N;
        if time[i] <= time[j]  then riskind[i,j]=1 ;
        else if time[i] > time[j] then riskind[i,j]=0;
    end;
end;
riskindsum=riskind[,+];
riskindsum2=riskindsum;
riskindsum2[N]=2;


/*******************calculate sigma u*****************************/;
sigma_u=ssq(wbar_w)/n*(k-1);

*print sigma_u;
/*********************calculate mu hat x at t*******************/;
MUHATWATT=RISKIND*WBAR/RISKINDSUM;

/*********************calculate mu hat z at t******************/;
MUHATZATT=RISKIND*Z/RISKINDSUM;

*print MUHATZATT;
/*******************calculate sigma XZ at t******************/;
QR=J(N,1,.);
MUZ=REPEAT(MUHATZATT,1,N);MUZ=MUZ`;ZMAT2=REPEAT(Z,1,N);ZDIFF=ZMAT2-MUZ;
MUW=REPEAT(MUHATWATT,1,N);MUW=MUW`;WBARMAT2=REPEAT(WBAR,1,N);WDIFF=WBAR
MAT2-MUW;
Q=ZDIFF#WDIFF;
Q=Q`;
```

```
Q2=RISKIND#Q;
Q2A=Q2*J(N,1,1);

sigma_xz=Q2A/(RISKINDSUM2-1);
sigma_xz[N]=0;

*print sigma_xz;
/*********************calculate sigma x at t************************/;
QT1=J(N,1,.);
c=sigma_u/k;
CMAT2=REPEAT(C,N,1);
QS=WDIFF#WDIFF;
QS=QS`;
QT=RISKIND#QS;
QT1=QT*J(N,1,1);
sigma_x=(QT1/(RISKINDSUM2-1))-CMAT2;
sigma_x[N]=-SIGMA_U/K;

*print Sigma_x;
/******************calculate sigma Z at t**************************/;
QYT1=J(N,1,.);
QY=ZDIFF#ZDIFF;
QY=QY`;
QYT=RISKIND#QY;
QYT1=QYT*J(N,1,1);
sigma_z=QYT1/(RISKINDSUM2-1);
sigma_z[N]=0;

*print sigma_z;
************************calibrated  values***********************/;
do j = 1 to 130;
      *time index;
    VAR_MAT[1,1]=c+SIGMA_X[j];
    VAR_MAT[1,2]=SIGMA_XZ[j];
    VAR_MAT[2,1]=SIGMA_XZ[j];
    VAR_MAT[2,2]=SIGMA_Z[j];
    INVVAR_MAT=GINV(VAR_MAT);
    MULT_MATRIX1[1,1]=SIGMA_X[j] ;
      MULT_MATRIX1[1,2]=SIGMA_XZ[j] ;                *print
MULT_MATRIX1;

    do i = 1 to n;
      *subject index;
        MULT_MATRIX2[1,1]=WBAR[i]- MUHATWATT[j];
        MULT_MATRIX2[2,1]=Z[i]-MUHATZATT[j] ;               *print
VAR_MAT mult_matrix1 mult_matrix2;

RSCMATRIX[j,i]=MUHATWATT[j]+MULT_MATRIX1*INVVAR_MAT*MULT_MATRIX2;
    END;
END;

do j=131 to n;
do i=1 to n;
RSCMATRIX[j,i]=WBAR[i];
end;
end;
```

```
*SHOW CONTENTS; *SHOW CONTENTS OF DATA SET;
 /*DM "Output; Clear; Log; Clear";*/

finish rsc_mod_adj20;

reset storage=trial.catalog3;
store module=rsc_mod_adj20;
quit;
run;

proc iml;
      reset storage=trial.catalog3;
      show storage;
quit;
run;

/*proc iml;
      reset storage=trial.catalog3;
      show storage;
            remove module=rsc_mod_adj20;
            show storage;
quit;
run;
```

```
/*****************************************************************/
/*Author:   Tina Dube                                            */
/*Module to calculate the Loglikelihood and 1st and 2nd order    */
/*derivative Set Calibration Adjusted for Small Risk Sets        */
/*Using the data generated by the earlier program, this program
/*generates the first and second order derivatives for the
loglikelihood function,                                          */
/*is generated based on true parameter estimates Betax=1 and Beta_Z=1*/
/*Co-Authored by http://www.biostat.umn.edu/~john-c/5421/notes.019   */
*****************************************************************/
proc iml symsize=7000000 worksize=800000;

/*************MODULE TO COMPLETE LOGLIKELIHOOD AND DERIVATIVES*******/
start loglike(beta,wbar, z,p,RSCMATRIX,time,delta,l,dl,d2l,evals);
evals=evals+1;
h=1e-6;
n=150;
riskind=j(n,n,.);
*print delta wbar z p RSCMATRIX time delta;

do i = 1 to n;
    do j = 1 to n;
        if time[i] <= time[j]  then riskind[i,j]=1 ;
        else if time[i] > time[j] then riskind[i,j]=0;
    end;
end;
      *print riskind;

/*Section 2 Calculate the log likelihood for the ith observation****/;
L1=0;
do j=1 to N;*j=time;
    jloglike_1=delta[j]*( (beta[1]*RSCMATRIX[j,j])+ (beta[2]*z[j]) );
    L1=L1+jloglike_1;
end;
*print L1;

LogL2_sum=0;
do j=1 to n;*j= time;
    L2=0;*reset summation indication at each time;
        do i=1 to n;*i=subject;
        iLL1= riskind[j,i]*exp( beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]
);
        L2=L2+iLL1;                                    *print j i l2 iLL1;
        end;
    ilogL2=delta[j]*log(L2);
    LogL2_sum=LogL2_sum+ilogL2;
end;

L=L1-LogL2_sum;                                        *print L;
*likelihood;
/*DM "Output; Clear; Log; Clear";*/
/*Jth and JKth derivatives for the ith Subject;*** *************/
dl=j(p,1,0);
Deriv_B1=0;
Deriv_B2=0;
do j=1 to n;*j=time;
    iderivb1=delta[j]*RSCMATRIX[j,j];
```

```
    iderivb2=delta[j]*z[j];
    Deriv_B1=Deriv_B1+iderivb1;
    Deriv_B2=Deriv_B2+iderivb2;
end;           *first derivatives for both beta 1 and beta 2 first sum;


ratiosum_B1=0;
ratiosum_B2=0;

d2l=j(2,2,0);
secondderiv_B1=0;
secondderiv_B2=0;
secondderiv_B1B2=0;

do j=1 to n;*j= time;
    jPL_B1_numer=0;
    jPL_B2_numer=0;
    jPL_B1B2_numer=0;
    jPL_B1_numer2=0;
    jPL_B2_numer2=0;
    jPL_B1B2_numer2=0;
      jPL_denom=0;

    jPL_B1_numer2=0;
    jPL_B2_numer2=0;*reset summation indication at each time;
        do i=1 to n;*i=subject;
        inumerx2=   RSCMATRIX[j,i]*RSCMATRIX[j,i]*riskind[j,i]*exp(
beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]   );
        inumerz2=                    z[i]*z[i]*riskind[j,i]*exp(
beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]   );
        inumerx=                  RSCMATRIX[j,i]*riskind[j,i]*exp(
beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]   );
        inumerz=                           z[i]*riskind[j,i]*exp(
beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]   );
        idenom=                                 riskind[j,i]*exp(
beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]   );
        inumerxz=             RSCMATRIX[j,i]*z[i]*riskind[j,i]*exp(
beta[1]*RSCMATRIX[j,i]+beta[2]*z[i]   );

        jPL_B1_numer2=jPL_B1_numer2+inumerx2;
        jPL_B1_numer=jPL_B1_numer+inumerx;
        jPL_denom=jPL_denom+idenom;

        jPL_B2_numer2=jPL_B2_numer2+inumerz2;
        jPL_B2_numer=jPL_B2_numer+inumerz;

        jPL_B1B2_numer2=jPL_B1B2_numer2+inumerxz;

        end;


      jratio_B1=       delta[j]*jPL_B1_numer/jPL_denom;
    jratio_B2=       delta[j]*jPL_B2_numer/jPL_denom;
    ratiosum_B1=     ratiosum_B1+jratio_B1;
    ratiosum_B2=     ratiosum_B2+jratio_B2;
```

```
    jratio_B11a=jPL_B1_numer2/jPL_denom;
    jratio_B11b=(jPL_B1_numer/jPL_denom)**2;
    isum_B1=delta[j]*(jratio_B11b-jratio_B11a);
    secondderiv_B1= secondderiv_B1+  isum_B1;


    jratio_B22a=jPL_B2_numer2/jPL_denom;
    jratio_B22b=(jPL_B2_numer/jPL_denom)**2;
    isum_B2=delta[j]*(jratio_B22b-jratio_B22a);
    secondderiv_B2= secondderiv_B2+  isum_B2;


    jratio_B12a=jPL_B1B2_numer2/jPL_denom;
    jratio_B12b=(jPL_B1_numer*jPL_B2_numer)/(jPL_denom)**2;
    isum_B1B2=delta[j]*(jratio_B12b-jratio_B12a);
    secondderiv_B1B2= secondderiv_B1B2+  isum_B1B2;
end;

DL_Beta1=Deriv_B1-ratiosum_B1;*first partial derivative beta 1;
DL_Beta2=Deriv_B2-ratiosum_B2;*first partial derivative beta 2;
dl[1]= DL_Beta1;
dl[2]= DL_Beta2;                                          *print dl;

d2l[1,1]= secondderiv_B1;
d2l[1,2]= secondderiv_B1B2;
d2l[2,1]= secondderiv_B1B2;
d2l[2,2]= secondderiv_B2;                                 *print d2l;


/*DM "Output; Clear; Log; Clear";*/
*************************************************************************
*****************************;
finish loglike;
*************************************************************************
*****************************;
      reset storage=trial.catalog2;
      store module=loglike;

      *show storage;

quit;run;

proc iml;
      reset storage=trial.catalog2;
      show storage;
print riskind;
quit;
run;
/*proc iml;
      reset storage=trial.catalog2;
      show storage;
            remove module=loglike;
            show storage;
quit;
run;
```

```
/*********************************************************************/
/*Author:   Tina Dube                                                */
/*Risk Set Calibration Adjusted for Small Risk Sets                  */
/*Using the data generated by the earlier program, this program
calculates the likelihood estimates by calling the log-
Likelihood module and the risk set calibration module. bivariate normal
covariate data with a      */
/*prespecified amount of correlation parameter. Additionally,        */
survival data                                                        */
/*is generated based on true parameter estimates Betax=1 and Beta_Z=1*/
/*Co-Authored by http://www.biostat.umn.edu/~john-c/5421/notes.019   */
*********************************************************************/
options nonotes PAGENO=1;
title '    ';
run;
%macro CalcLike(N,cohorts);
%do X = 1 %to &COHORTS;
proc iml symsize=70000000 worksize=8000;
      reset storage=trial.catalog3;
      * Read data into IML ;

      reset storage=trial.catalog3;
      USE trial.DTA_0105_2a&x;

      read all var {X}into X;
    read all var {W1 W2 W3} into W;
      k=ncol(W);
      WBAR=W[,:];
      read all var {z} into Z;
    read all var {time} into TIME;                *read survival time
matrix;
    read all var {delta} into delta;                      *read DELTA
matrix;

      store x;free x;
      store w;free w;
      store wbar;free wbar;
    store z;free z;
    store time;free time;
    store delta;free delta;

load w wbar z delta time;
load module=rsc_mod_adj20;
run rsc_mod_adj20(w ,wbar, z ,delta, time,rscmatrix);
*print rscmatrix;
      *check module processes;
store rscmatrix;free rscmatrix;

print "finished calibrating data for cohort: " &x ".";

*Calculate the partial likelihood estimates;
eps=1e-8; *tolerance limits for increments;
diff=eps+1; *preset diff>eps;
oldl=-1e20; *present old likelihood very small;
p=2; *set number of parameters;
beta={1.00, 1.00}; *initialize parameter vector;
evals=0; *preset number of function evals;
```

```
*The following do loop stops when the increments in beta are
sufficiently small, or when number of interactions reaches 20 provided
the log likelihood increases;


print, "Likelihood estimates for cohort number:" &x ".";
cohortnumber=&x;

load RSCMATRIX;
do iter= 1 to 20 while (diff > eps);
      load module=loglike_20;
    run loglike_20(beta,wbar, z,p,RSCMATRIX,time,delta,l,dl,d2l,evals);
    invd2l=inv(d2l);
    factor=1; *print factor;
    tbeta=beta-invd2l *dl; *The Newton Step;

tl=l;
tdl=dl;
td2l=d2l;
*Go to step halving if the likelihood does not increase
if (l < oldl) then do;
        factor=0.5 * factor; * print factor;
        tl=oldl - 0.1*abs(oldl);
        do halves = 1 to 10 while (tl < oldl);
        tbeta=beta-factor *invd2l * dl; *step halving;
            load module=loglike_20;
        run loglike_20(tbeta,wbar,
z,p,RSCMATRIX,time,delta,tl,tdl,td2l,evals);
        factor=0.5*factor; * print factor;
        * print, "Step halving: " iter halves tbeta tl oldl tdl td2l;
        end;
            if (tl < oldl) then do;
                * print, "No convergence after " halves "step-
halves..." iter halves tbeta tl oldl tdl td2l;
            end;
end;

beta=tbeta;tina=invd2l*dl;dube=factor*tina;
diff=max(abs(dube));
l=tl; oldl=tl; dl=tdl; d2l=td2l;
* print, iter beta l dl d2l diff;
end;
if (diff> eps) then do;
    * print, "No convergence after " iter " iterations..." iter halves
tbeta tl oldl tdl td2l;
end;

serralt=j(1,p,0);
serr=j(p,1,0);
do i = 1 to p;
    serr[i]=sqrt(-invd2l[i,i]); *standard errors= square root of
diagonal;
end;

minvd2l=-invd2l;
```

```
m2l=-2*l;

*print, "Number of function evaluations :" evals;
covar=-invd2l;
*print, " -2 * loglikelihood: " m2l;
*print, " Estimated Coefficients, standard errors: ",beta serr;
*print, " Estimated Covariance matrix of coefficients: ",minvd2l;

betaRSC=beta`;
*print beta betarsc;
serr=serr`;smallvec=j(1,2,1);
bias=betaRSC-smallvec;

BETA=&x||betaRSC||serr||bias;

print beta;

cname4={"cohortnumber" "XRSC" "ZRSC" "serr1" "serr2" "biasx" "biasz"};
create betavec&x from beta [COLNAME=CNAME4];
append from beta;
*show storage;
*show space;

free RISKIND RSCMATRIX TIME WBAR Z beta;

QUIT; *END PROC IML;
RUN;

%END;                                                    *END X LOOP RUN
IML ;

%do J = 2 %to &COHORTS;
PROC APPEND BASE=betavec1 DATA=betavec&J;
RUN;
QUIT;
%END;

%mend CalcLike;
      *END CalcLike MACRO;
%CalcLike(150,1060);
```