

Instrumental Variable Estimation in Linear Regression Models with Censored and Mismeasured Covariates

by

Goyumi Lavanya Kumarasiri

A Thesis submitted to the Faculty of Graduate and Postdoctoral
Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics
University of Manitoba
Winnipeg

Copyright © 2026 by Goyumi Lavanya Kumarasiri

Abstract

This study deals with linear regression models where covariates are subject to both measurement error and censoring. We propose a two-step estimation process that combines the instrumental variable (IV) method with censored regression modeling and derive consistent estimators of the slope and intercept parameters. Our simulation studies suggest that the obtained estimators are unbiased and have lower root mean squared error compared to the estimators based on SIMEX and naive methods, across different model distributions, censoring scenarios, and sample sizes. This highlights the importance of accounting for both censoring and measurement error in parameter estimation in such linear regression models. Our estimation method is further validated using real data applications.

Keywords: Error in Covariates, Censored Covariates, Instrumental Variables, Measurement Error.

Acknowledgment Page

First and foremost, I would like to express my sincere gratitude to my supervisors Dr. Liqun Wang and Dr. Yuliya Martsynyuk for their guidance and support throughout this time. Their expertise and thoughtful feedback have made a significant impact on this research and have helped me grow as a researcher.

I would also like to thank my advisory committee, Dr. Sumeet Kalia and Dr. Kevin McGregor for taking the time to review the thesis and providing comments, suggestions and constructive feedback.

I am also thankful to the Department of Statistics at the University of Manitoba for the opportunity and providing a supportive academic environment. I would also like to acknowledge the financial support provided by Dr. Liqun Wang and the University of Manitoba throughout my studies.

Finally, I am deeply grateful to my family and close friends for their constant encouragement and for being there for me through thick and thin.

Dedication Page

I dedicate this thesis to my parents, for their love, support and dedication since I was a kid.

Contents

Contents	iii
List of Tables	vii
List of Figures	xi
1 Introduction	1
1.1 Background and Motivation	1
1.2 Literature Review	2
1.3 Objectives of the Study	6
1.4 Outline of the Thesis	7
2 Measurement Error Models and Correction Methods	9
2.1 Measurement Error Models	9
2.2 Instrumental Variable Method	12
2.3 Regression Calibration Method	19
2.4 Simulation Extrapolation Method (SIMEX)	22

3	Censored and Mismeasured Covariates	27
3.1	Censored Covariates	27
3.2	Model	29
3.3	Instrumental Variable Method	31
3.4	Regression Calibration Method (RC)	34
3.5	Simulation Extrapolation Method (SIMEX)	37
3.6	Naive Method	38
4	Simulation Studies	39
4.1	Experiment 1	43
4.2	Experiment 2	47
4.3	Experiment 3	51
4.4	Experiment 4	56
4.5	Experiment 5	61
4.6	Summary of the Simulation Results	65
4.7	Additional Experiments	67
4.7.1	Instrumental Strength	67
5	Real Data Applications	71
5.1	Mercury Level using Fish Consumption	71
5.2	Framingham Heart Study	77
6	Discussion and Conclusion	83

Bibliography	87
A Additional Simulation Studies	91
A.1 Departure from Normality	102
Index	106

List of Tables

4.1	Comparison of methods under 10% censoring and different sample sizes for Experiment 1.	45
4.2	Comparison of methods under 50% censoring and different sample sizes for Experiment 1.	45
4.3	Comparison of methods under 10% censoring and different sample sizes for Experiment 2.	48
4.4	Comparison of methods under 50% censoring and different sample sizes for Experiment 2.	49
4.5	Comparison of methods under 10% censoring and different sample sizes for Experiment 3.	53
4.6	Comparison of methods under 50% censoring and different sample sizes for Experiment 3.	54
4.7	Comparison of methods under 10% censoring and different sample sizes for Experiment 4.	58
4.8	Comparison of methods under 50% censoring and different sample sizes for Experiment 4.	58

4.9	Comparison of methods under 10% censoring and different sample sizes for Experiment 5.	63
4.10	Comparison of methods under 50% censoring and different sample sizes for Experiment 5.	63
4.11	Performance of the IV estimator under varying instrument strengths for Experiment 3 under 10% censoring.	69
5.1	Description of variables used in the NHANES dietary study. .	73
5.2	Comparison of IV, RC, and SIMEX estimation methods for NHANES dietary study	75
5.3	Description of variables used in the Framingham Heart Study analysis.	78
5.4	Comparison of IV, RC, and SIMEX estimation methods for FHS data	81
A.1	Comparison of IV, RC, SIMEX, and naive methods for $n = 50$, and 5% censoring under two parameter settings for Experiment 1. .	92
A.2	Comparison of methods for $n = 500$ with 5% censoring under two parameter settings for Experiment 1.	92
A.3	Comparison of methods for $n = 500$ with 10% censoring under two parameter settings for Experiment 1.	93
A.4	Comparison of methods for $n = 500$ with 50% censoring under two parameter settings for Experiment 1.	93

A.5	Comparison of IV, RC, SIMEX, and naive methods for $n = 50$ and 10% censoring under two parameter settings for Experiment 3.	95
A.6	Comparison of IV, RC, SIMEX, and naive methods for $n = 100$ and 10% censoring under two parameter settings for Experiment 3.	96
A.7	Comparison of IV, RC, SIMEX, and naive methods for $n = 200$ and 10% censoring under two parameter settings for Experiment 3.	97
A.8	Comparison of IV, RC, SIMEX, and naive methods for $n = 500$ and 10% censoring under two parameter settings for Experiment 3.	98
A.9	Comparison of IV, RC, SIMEX, and naive methods for $n = 50$ and 10% censoring under two parameter settings for Experiment 4.	99
A.10	Comparison of IV, RC, SIMEX, and naive methods for $n = 100$ and 10% censoring under two parameter settings for Experiment 4.	100
A.11	Comparison of IV, RC, SIMEX, and Naive methods for $n = 200$ and 10% censoring under two parameter settings for Experiment 4.	100
A.12	Comparison of IV, RC, SIMEX, and Naive methods for $n = 500$ and 10% censoring under two parameter settings for Experiment 4.	101

A.13 Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 1.	102
A.14 Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 2.	103
A.15 Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 3.	103
A.16 Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 4.	104
A.17 Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 5.	104

List of Figures

4.1	Comparison of RMSE values under the considered scenarios in Experiment 1.	47
4.2	Comparison of RMSE values under the considered scenarios in Experiment 2	51
4.3	Comparison of RMSE values under the considered scenarios in Experiment 3	55
4.4	Comparison of RMSE values under the considered scenarios in Experiment 4	60
4.5	Comparison of RMSE values under the considered scenarios in Experiment 5	65

Chapter 1

Introduction

1.1 Background and Motivation

Linear regression models are widely used in many scientific disciplines, including epidemiology, environmental science, public health, and economics. The underlying assumptions of the regression model have a major effect on the validity of the resulting statistical inference. According to one of the fundamental assumptions, it typically assumes that the covariates are measured exactly as they are, which means measured without error. However, when it comes to real world problems, this assumption is rarely satisfied since the covariates could be measured with error due to various reasons such as limitations in instruments, data collecting methods and reporting errors. This can be identified as a source of bias in statistical analysis. In addition to measurement error, censoring is another problem that frequently arises in statistical modeling due to detection limits ([Atem et al., 2017](#)) or the data including a substantial proportion of individuals with zero exposure, resulting in zero inflated distributions ([Wang et al., 2024](#)).

Both measurement error and censoring problems have been broadly studied individually. However, in many real world examples in biomedical and public health research, the measurement error and censoring problems occur simultaneously. For example, predicting the level of mercury in the blood of individuals using fish consumption, which is based on self reported dietary data is measured with error due to the inaccuracy of reporting. At the same time it contains an excess number of zeros (Kipnis et al., 2009). Adding more into the examples, in studies on assessing the relationship between blood glucose level and systolic blood pressure, the glucose measure can be measured with error due to laboratory variability and also can be subject to censoring due to limit of detection (Chen et al., 2011; Tsimihodimos et al., 2018). These examples highlight that covariates can be affected by both measurement error and censoring in the real world problems.

Motivated by these interesting problems, this study focuses on parameter estimation in regression models where a covariate is censored and measured with error at the same time, giving more attention to the instrumental variable method.

1.2 Literature Review

Measurement error in covariates is a common problem in regression analysis and many studies have been done in this area particularly in the fields of epidemiology, nutrition, environmental science, econometrics and many more. In the presence of measurement error in variables, the standard regression methods produce biased parameter estimates which can lead to misleading inference. Therefore, using measurement error correction methods is widely important. In

the early days studies have been done to identify the consequences of ignoring the errors in covariates and the major discovery was attenuation bias, where the regression coefficients are biased towards zero when measurement error is present in covariates. Early foundational work on measurement error done by [Fuller \(1987\)](#) identified the classical and Berkson measurement error structures and developed estimation procedures according to the model assumptions. To further identify the methods later, [Carroll et al. \(2006\)](#) expanded to a class of statistical methods to analyze regression models with measurement error in both linear and nonlinear cases. Many statistical methods have been proposed to correct measurement error in variables since it is a crucial factor to obtain unbiased estimates and valid statistical inference. Moreover, [Buonaccorsi \(2010\)](#) emphasized the practical aspect of measurement error models.

Among the several approaches designed to correct measurement error, the Regression Calibration (RC) method is the most widely used method in many applications due to its conceptual simplicity and practical implementation ([Spiegelman et al., 1997](#); [Carroll et al., 2006](#); [Spiegelman, 2014](#)). The RC method involves developing a calibration equation that estimates the true exposure value on the measured error-prone covariates. The RC method was initially developed for epidemiological studies involving measurement error in exposure variables. This method was further developed to use in the studies related to nutritional epidemiology ([Kipnis et al., 2009](#)). Another common method of addressing the measurement error problem is the Simulation Extrapolation Method (SIMEX). The SIMEX method was introduced and further developed by [Cook and Stefanski \(1994\)](#). Unlike the other error correction methods, the SIMEX method corrects the measurement error by adding extra noise to the observed data, studying the changes of the parameter estimates and

finally extrapolating these estimates back to the error-free case using different extrapolating functions. [Carroll et al. \(2006\)](#) further developed the SIMEX methodology to showcase its applicability to nonlinear problems.

Apart from the RC and SIMEX methods, another important method of error correction is the instrumental variable method (IV). The IV method relies on auxiliary variables which are correlated with error prone covariates but independent of both measurement error and regression error terms. [Buzas and Stefanski \(1996\)](#) developed instrumental variable estimators for generalized linear models with measurement error, demonstrating that IV-based estimation can produce consistent parameter estimates in the presence of mismeasured covariates. This method does not require the direct estimation of the true covariate and provides consistent estimators. [Carroll et al. \(2006\)](#) extended this method into nonlinear models while [Fuller \(1987\)](#) laid the foundation. The IV method is widely used in applications in econometrics ([Card, 2001](#)) in the early stage and later this method was employed in other applications such as biomedical applications ([Lu et al., 2023](#)). Generally, the IV method can be considered as a better method to be used when the appropriate auxiliary information is available.

Moreover, censored variables is another challenge in regression modeling. when we consider the censoring problem, it can be observed in both the response variable and predictor variables in many applications including biomedical and environmental studies. In this study the focus will be on censored covariates. Censored covariates give partial information about the variables which makes the usage of classical estimation methods challenging. Early work was done by [Rigobon and Stoker \(2007\)](#) on regression models with censored covariates.

Many statistical methods have been proposed to address this censoring issue such as likelihood based methods, imputation methods, weighting methods, and Bayesian methods (Kwon, 2026; Lotspeich et al., 2024). Apart from the parametric approaches, semiparametric approaches have gained attention since they improve robustness. More recent studies such as Atem et al. (2017); Ashner and Garcia (2024) expanded the methodologies to address the challenges of censored covariate problem.

In addition, the scenario where the covariates are censored and measured with error can be observed in real world applications where the problem gets more challenging. This issue commonly arises in biomedical and epidemiological studies where biomarkers could be measured with error and at the same time fall below a detection limit (Qin et al., 2026). These studies show that covariates that are simultaneously censored and measured with error occur naturally in many scientific applications. The traditional measurement error correction methods assume that the covariates are fully observed. Therefore, these methods give biased estimates. A limited number of studies have been done to introduce methods to address the censored and mismeasured covariate scenario particularly based on maximum likelihood estimation. Moreover, instrumental variable method has been used in models where both censoring and measurement error appear. Wang and Hsiao (2007) have studied censored regression models where explanatory variables are subject to additive measurement error and proposed a two step estimation procedure to address this problem. This approach yields to consistent estimates.

Extensive studies have been done on measurement error models as well as regression models with censored covariate problems separately but the combined scenario has received less attention. Some studies addressing this problem

include: (Song and Wang, 2025; Qin et al., 2026; Kipnis et al., 2009). Although regression models with censored responses and mismeasured covariates have been studied widely, a very few studies have focused on the case where a covariate is censored and mismeasured at the same time. This combined problem addresses a challenge in statistical inference. Motivated by the above mentioned research gap, this study focuses on linear regression models where a covariate is mismeasured and censored at the same time and develops an instrumental variable method to address this problem.

1.3 Objectives of the Study

The primary objectives of this study can be stated as below.

1. **Model Formulation.**

Formulation of the model where a covariate is censored and measured with error simultaneously.

2. **Methodology Development.**

Derivation of consistent estimates using a two step estimation procedure involving the instrumental variable method.

3. **Simulation Studies and Methodology Comparison.**

Conducting of simulation studies for the proposed methodology, RC, SIMEX, and Naive methods for different model settings, sample sizes, censoring levels, and comparison of results of the proposed methodology with the above mentioned methods.

4. Real World Application.

Application of these methods to real world examples to verify the methodology.

1.4 Outline of the Thesis

The remaining chapters of the thesis are organized as follows. Chapter 2 focuses on the measurement error problem discussing the estimation of the measurement error problem including the methodologies. Chapter 3 discusses the censored covariates and the main focus of this study which is about censored and mismeasured covariates and it is followed by the methodologies. Chapter 4 presents the simulation studies under various scenarios while the chapters 5 and 6 focus on the real data applications used and the discussion, respectively. Moreover, appendix A includes some additional simulations done on high reliability ratios and proposed methodology under departure of normality.

Chapter 2

Measurement Error Models and Correction Methods

This chapter discusses about the classical measurement error framework, underlying theoretical foundations and approaches that are being used to correct the measurement error bias.

2.1 Measurement Error Models

Classical Measurement Error Model

Consider a simple linear measurement error model.

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad (2.1)$$

where, Y is the response variable, X is the true unobserved covariate and ϵ is the random error satisfying $E[\epsilon] = 0$, and $var(\epsilon) = \sigma_\epsilon^2$.

Given that X is not directly observed, the error prone observed covariate, X^* can be written under the classical additive measurement error as,

$$X^* = X + e, \quad (2.2)$$

where e is the measurement error.

The common assumptions under the classical measurement error are:

A 2.1. *the measurement error distribution follows, $E(e) = 0$;*

A 2.2. *the measurement error is independent of the true covariate, $e \perp X$;*

A 2.3. *the measurement error is independent of the response error, $e \perp \epsilon$.*

Under these assumptions, the surrogate variable X^* is unbiased of the true covariate X .

Substituting X^* for X gives,

$$Y = \beta_0 + \beta_1 X^* + u, \quad (2.3)$$

where $u = \epsilon - \beta_1 e$.

In order to use the ordinary least squares estimation, it should satisfy, $Cov(X^*, u) = 0$. However, in this case,

$$Cov(X^*, u) = Cov(X + e, \epsilon - \beta_1 e) = -\beta_1 \sigma_e^2. \quad (2.4)$$

Since, $Cov(u, X^*) \neq 0$ in general, the least squares estimating equation for β ,

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_{X_i^*}) X_i^* = 0, \quad (2.5)$$

would be biased due to violation of the exogeneity condition. This is commonly known as the attenuation bias.

Theorem 2.1. (*Attenuation Bias*): *Under the classical measurement error model, the least squares estimator the the slope, $\hat{\beta}_1$, is inconsistent: $\hat{\beta}_1 \xrightarrow{P} \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2}$, as $n \rightarrow \infty$, where the quantity*

$$\lambda := \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2}, \quad (2.6)$$

is known as the attenuation factor or the reliability ratio, which ranges between 0 and 1.

Since the attenuation factor is less than 1, $\hat{\beta}_1$ underestimates the slope. The larger the variation of the measurement error, the smaller the attenuation factor and severity of the underestimation.

A measurement error model can also be prone to another type of error, the so-called Berkson measurement error, which occurs when the actual individual value varies around the assigned value (Fuller, 1987). This is common in studies where group level measurements are assigned to individuals. This method is not considered further since the main focus of this thesis is on the classical additive measurement error.

Measurement error can be classified as differential and non-differential as well. Throughout this thesis, non-differential measurement error is assumed, where the measurement error is independent of the response and the true covariate.

2.2 Instrumental Variable Method

The instrumental variable (IV) Method is a framework that addresses the measurement error problem using auxiliary variables known as instruments, which is independent of both the regression error and measurement error and has a high correlation with the unobserved true covariate. This estimation method is particularly important since it does not require the direct observation of the true covariate and the knowledge of the measurement error variance. The IV estimator gives consistent estimates for regression parameters even under the presence of mismeasured covariates, under the given conditions.

The instrumental variable method provides an attractive alternative when replicate measurements are present. This method is widely used in econometrics and have been adapted to measurement error problems in many areas, including biostatistics and epidemiology (Carroll et al., 2006). However, it is important to identify a valid instrument since the quality of the result depends on this.

Instrumental Variable Assumptions

Let, S be the instrumental variable which is directly observable. The instrument is valid if it satisfies the following assumptions.

A 2.4. (*Relevance*): *The instrument should be correlated with the true covariate.*

$$\text{Cov}(S, X) \neq 0. \quad (2.7)$$

This relevance condition represents that the instrument contains information about the unobserved covariate and the strength of this relationship plays a

critical role in the estimation. If the instruments are weakly correlated with the true covariate, it can lead to unstable estimates.

A 2.5. (*Exogeneity*): *The instrument should be independent of both the regression error ε and measurement error e :*

$$S \perp (\varepsilon, e). \quad (2.8)$$

By this assumption, $S \perp e$, and it ensures that the instrument is not contaminated by the measurement error. Any correlation between the instrument and measurement error undermines the objective and the covariate may become biased. [Carroll et al. \(2006\)](#) mentions that in linear models, the weaker covariance between the instrument and the measurement error is sufficient. On the other hand, the assumption $S \perp \varepsilon$ ensures that the instrument influences the response only through the true covariate.

The validity of the instrumental variable estimation depends on these assumptions. However, the exogeneity assumption is not directly testable for the observed data. In that case, the choice of the instrument can be stated using the domain knowledge or the data generating process. Moreover, if the instrument is weakly associated with the true covariate, it can lead to biased estimates and poor performance. Therefore, it is very important to pay attention to the validity of the instrument before proceeding with the instrumental variable method.

Given that the instrumental variable S satisfies the above mentioned assumptions, A 2.4 and A 2.5, the objective is to estimate the coefficient β_1 when the true covariate X is unobserved.

Moment Conditions for Instrumental Variable Estimation

Followed by the assumptions stated above, the instrumental variable estimator can be derived using a set of population moment conditions which emphasize the relationship between the instrument, response variable and the unobserved true covariate.

Consider the linear regression model (2.1) where X is not directly observable and suppose that S is a valid instrument satisfying the above discussed assumptions A 2.4 and A 2.5.

Rearrange the terms of the regression model (2.1) as,

$$Y - \beta_0 - \beta_1 X = \epsilon. \quad (2.9)$$

After multiplying both sides of the equation by the instrument S , it will lead to

$$S(Y - \beta_0 - \beta_1 X) = S\epsilon. \quad (2.10)$$

By taking the expectation of both sides of (2.10),

$$E[S(Y - \beta_0 - \beta_1 X)] = E[S\epsilon], \quad (2.11)$$

where $E[S\epsilon] = 0$, since $S \perp \epsilon$. Therefore, it gives the following population moment condition equation:

$$E[S(Y - \beta_0 - \beta_1 X)] = 0. \quad (2.12)$$

Since we consider the true covariate X to be unobserved, to account for the measurement error, after substituting (2.2) into (2.12) it gives,

$$E[S(Y - \beta_0 - \beta_1(X^* - e))] = 0. \quad (2.13)$$

After expanding the above equation and simplification, it will lead to,

$$\begin{aligned} E[S(Y - \beta_0 - \beta_1 X^* + \beta_1 e)] &= 0, \\ E[S(Y - \beta_0 - \beta_1 X^*)] + \beta_1 E[Se] &= 0. \end{aligned} \quad (2.14)$$

According to the assumption A 2.5, $E[Se] = 0$. Therefore, under measurement error, the moment condition for instrumental variable estimation can be presented as,

$$E[S(Y - \beta_0 - \beta_1 X^*)] = 0. \quad (2.15)$$

The above equation (2.15) can be used to estimate the unknown regression coefficient.

After expanding and rearranging the terms in (2.15), we have

$$\begin{aligned} E[SY] - \beta_0 E[S] - \beta_1 E[SX^*] &= 0, \\ \beta_1 E[SX^*] &= E[SY] - \beta_0 E[S], \\ \beta_1 &= \frac{E(SY) - \beta_0 E[S]}{E[SX^*]}. \end{aligned} \quad (2.16)$$

After centering all the variables before estimation, the equation reduces to,

$$\beta_1 = \frac{E[SY]}{E[SX^*]}. \quad (2.17)$$

According to the equation, (2.17), the regression coefficient is determined by two observable population moments.

This estimator can be represented using a covariance representation. The equations can be presented equivalently as, $E(SY) = Cov(S, Y)$ and $E(SX^*) = Cov(S, X^*)$.

Therefore, according to [Carroll et al. \(2006\)](#), the classical population instrumental variable estimator can be presented as,

$$\beta_1 = \frac{Cov(S, Y)}{Cov(S, X^*)}. \quad (2.18)$$

This covariance representation can be obtained directly using the regression model as well.

Theorem 2.2. (*Instrumental Variable Estimator*): *Given the linear regression model (2.1)-(2.2) and given that the instrument S satisfies the assumptions A 2.4 and A 2.5, the regression coefficient β_1 is given by*

$$\beta_{IV} = \frac{Cov(S, Y)}{Cov(S, X^*)}.$$

Proof: We have

$$Cov(S, Y) = Cov(S, \beta_1 X + \epsilon) = \beta_1 Cov(S, X) + Cov(S, \epsilon).$$

Since, $Cov(S, \epsilon) = 0$,

$$Cov(S, Y) = \beta_1 Cov(S, X). \quad (2.19)$$

According to the equation (2.2) and independence of S and e ,

$$Cov(S, X^*) = Cov(S, X + e) = Cov(S, X) + Cov(S, e) = Cov(S, X).$$

After substituting this result into (2.19),

$$Cov(S, Y) = \beta_1 Cov(S, X^*).$$

According to [Carroll et al. \(2006\)](#), this leads to the regression coefficient,

$$\beta_{IV} = \frac{Cov(S, Y)}{Cov(S, X^*)}. \quad (2.20)$$

Theorem 2.3. (*Consistency of the instrumental variable estimator*):

Let

$$\hat{\beta}_{IV} := \frac{\hat{Cov}(S, Y)}{\hat{Cov}(S, X^*)}$$

denote the instrument variable estimator of β_{IV} defined in (2.20). Given the linear regression model (2.1)-(2.2) and given that the instrumental variable assumptions are satisfied, as $n \rightarrow \infty$,

$$\hat{\beta}_{IV} \xrightarrow{P} \beta_{IV}.$$

Proof: Let $\hat{Cov}(S, Y)$ and $\hat{Cov}(S, X^*)$ denote the sample covariances between S and Y , and S and X^* , respectively. The instrumental variable estimator is given by,

By the Weak Law of Large Numbers,

$$\hat{Cov}(S, Y) \xrightarrow{P} Cov(S, Y)$$

and

$$\hat{Cov}(S, X^*) \xrightarrow{P} Cov(S, X^*),$$

where

$$Cov(S, Y) = \beta_X Cov(S, X) + Cov(S, \epsilon) = \beta_X Cov(S, X)$$

and

$$Cov(S, X^*) = Cov(S, X) + Cov(S, e) = Cov(S, X).$$

Therefore,

$$\frac{Cov(S, Y)}{Cov(S, X^*)} = \frac{\beta_{IV} Cov(S, X)}{Cov(S, X)} = \beta_{IV}.$$

Thus, according to Slutsky's theorem,

$$\hat{\beta}_{IV} = \frac{\hat{Cov}(S, Y)}{\hat{Cov}(S, X^*)} \xrightarrow{P} \frac{Cov(S, Y)}{Cov(S, X^*)} = \beta_{IV},$$

with β_{IV} of (2.20).

Theorem 2.4. (*Asymptotic Normality*): *When the conditions of Theorem 3 hold, the delta method implies:*

$$\sqrt{n}(\hat{\beta}_{IV} - \beta_{IV}) \xrightarrow{d} N(0, \Sigma_{IV})$$

where the asymptotic variance of the instrumental variable estimator is given

$$\text{by } \Sigma_{IV} = \frac{\text{Var}[(S - E[S])(Y - \beta_{IV}X^*)]}{\text{Cov}(S, X^*)^2}.$$

In practice, IV estimation is commonly computed using two stages and the resulting coefficient $\hat{\beta}_{2SLS}$, is called the two-stage least squares estimator. The steps can be summarized as follows.

1. Regress X on S and all other exogenous covariates in the model and obtain the fitted values for $\hat{X}$.
2. Regress Y on $\hat{X}$ and other covariates in the model.

This method is computationally convenient and can be extended to more broader problems that will be discussed in the coming chapters.

2.3 Regression Calibration Method

The regression calibration (RC) method is a widely used method to correct measurement error due to the computational simplicity and convenience of implementation. The key idea behind the RC method is to correct the measurement error by replacing the unobserved true covariate X by its conditional expectation given the observed data $\hat{X} = E[X|X^*]$.

Consider the same linear regression model (2.1), where X is error prone. Since replacing the true covariate X with the observed surrogate X^* produces biased estimates, the regression calibration method reduces the bias by replacing the unobserved surrogate with its conditional expectation given the observed covariates and it can be presented as follows:

$$\hat{X} = E(X|X^*). \quad (2.21)$$

The motivation of the regression calibration method can be explained using the conditional mean of the response variable under classical measurement error as follows:

$$E(Y|X^*) = E[E(Y|X, X^*)|X^*]. \quad (2.22)$$

The expression can be simplified as below since the response depends on the observed variable only through the true covariate:

$$E(Y|X, X^*) = E(Y|X).$$

Substituting to the regression model and (2.22) gives,

$$E(Y|X) = \beta_0 + \beta_1 X,$$

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

After taking the expectation it gives,

$$E(Y|X^*) = \beta_0 + \beta_1 E(X|X^*). \quad (2.23)$$

According to the above equation (2.23), it can be said that replacing X by $E(X|X^*)$ gives an approximation to the original regression model, which is a less biased approach. It requires additional information about the measurement error process to estimate this conditional expectation in practice. In that case, repeated measurements can be used to obtain the measurement error variance and a validation study can be used estimate the calibration relationship.

Derivation of the Calibration Function

Suppose, $X \sim N(\mu_X, \sigma_X^2)$ and $e \sim N(0, \sigma_e^2)$. It follows that $X^* \sim N(\mu_X, \sigma_X^2 + \sigma_e^2)$ and (X, X^*) are jointly normal.

Theorem 2.5. (*Calibration Estimator*): Suppose $\begin{pmatrix} X \\ X^* \end{pmatrix}$ follows a bivariate normal distribution with the mean vector $\begin{pmatrix} \mu_X \\ \mu_{X^*} \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \sigma_X^2 & Cov(X, X^*) \\ Cov(X, X^*) & \sigma_{X^*}^2 \end{pmatrix}$. Then,

$$E[X|X^*] = \mu_X + \frac{Cov(X, X^*)}{Var(X^*)}(X^* - \mu_{X^*}). \quad (2.24)$$

Proof: The covariance between the true covariate and the observed covariate is given by:

$$Cov(X, X^*) = Cov(X, X + e) = Var(X) + Cov(X, e).$$

Since $X \perp e$, the term $Cov(X, e) = 0$, it follows that,

$$Cov(X, X^*) = Var(X) = \sigma_X^2. \quad (2.25)$$

Similarly,

$$\begin{aligned} Var(X^*) &= Var(X + e) = Var(X) + Var(e), \\ Var(X^*) &= \sigma_X^2 + \sigma_e^2. \end{aligned} \quad (2.26)$$

After substituting (2.25) and (2.26) into (2.24), the calibration estimator can be presented as:

$$E[X|X^*] = \mu_X + \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2}(X^* - \mu_{X^*}).$$

Then the calibration equation is given by,

$$E[X|X^*] = \mu_X + \lambda(X^* - \mu_{X^*}), \quad (2.27)$$

with λ of (2.6). When the measurement error variance is very small, the reliability ratio approaches 1, and it will be approximately equal to the observed variable.

According to the methodology, substituting the conditional expectation into the regression model gives the following regression calibration model:

$$Y = \beta_0 + \beta_1 \hat{X} + \varepsilon. \quad (2.28)$$

The regression coefficients can be recovered using the ordinary least squares

method since $\hat{X}$ is observed. Let, $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ and $M = \begin{pmatrix} 1 & \hat{X}_1 \\ 1 & \hat{X}_2 \\ \vdots & \vdots \\ 1 & \hat{X}_n \end{pmatrix}$

After minimizing the residual sum of squares the calibration estimator, which is the ordinary least squares estimator computed using the calibrated covariate can be represented as,

$$\hat{\beta}_{RC} = (M'M)^{-1}M'Y. \quad (2.29)$$

In summary, following is the regression calibration algorithm.

1. Estimate the parameters $\mu_X, \sigma_X^2, \sigma_e^2$ of the measurement error model.
2. Compute the reliability ratio, λ .
3. Calculate the calibrated covariate, $\hat{X}$.
4. Fit the calibrated regression model by regressing Y on $\hat{X}$.
5. Obtain the regression calibration estimator $\hat{\beta}_{RC}$.

Under the classical measurement error, the regression calibration estimator is approximately consistent.

2.4 Simulation Extrapolation Method (SIMEX)

The SIMEX method provides a flexible alternative which is broadly applicable. Unlike the other error correction methods, the SIMEX method adopts a different strategy. The idea behind the SIMEX method is to correct for the measurement error by artificially adding a measurement error, studying how the parameter estimates change, and extrapolating back to the zero error case. This method directly depends on the assumed error distribution and works best when the

error structure is known. Moreover, the accuracy of the estimator depends highly on the choice of the extrapolation function.

The SIMEX procedure consists of two main steps, which includes the simulation stage and the extrapolation stage. Given the regression model (2.1), the steps of the SIMEX method are as follows.

1. Simulation

Let $\Lambda = (\lambda_1, \lambda_2, \dots, \lambda_K)$ where, $\lambda_k \geq 0$ and $k = 1, \dots, K$.

For each $\lambda \in \Lambda$ and for $b = 1, \dots, B$,

- (a) Generate an extra error $e^{*(b)} \sim N(0, \sigma_e^2)$ and the pseudo data,

$$X^{*(\lambda, b)} = X^* + \sqrt{\lambda}e^{*(b)}. \quad (2.30)$$

The factor $\sqrt{\lambda}$ introduces an additional measurement error while preserving the measurement error structure.

The variance of the additional measurement error is given by,

$$Var(\sqrt{\lambda}e^{*(b)}) = \lambda\sigma_e^2.$$

After substituting,

$$Var(X^{*(\lambda, b)}) = Var(X^*) + \lambda\sigma_e^2.$$

Since $Var(X^*) = \sigma_X^2 + \sigma_e^2$,

$$Var(X^{*(\lambda, b)}) = \sigma_X^2 + (1 + \lambda)\sigma_e^2. \quad (2.31)$$

- (b) Fit the regression $(Y, X^{*(\lambda)})$ for each simulated dataset and obtain $\hat{\beta}^{(\lambda, b)}$.

- (c) Repeat this simulation B times for each value of λ since the simulated measurement error is random, and take the average as follows.

$$\hat{\beta}(\lambda) = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)}(\lambda). \quad (2.32)$$

2. Extrapolation

According to the average parameter estimate (2.32), the collection of estimates can be shown as, $\{\hat{\beta}(\lambda_1), \hat{\beta}(\lambda_2), \dots, \hat{\beta}(\lambda_K)\}$. Then, fit a smooth function $\phi(\lambda)$.

In the SIMEX literature, several extrapolation functions have been proposed which include linear, quadratic, and reciprocal functions. Often a quadratic function is used as it is sufficiently flexible. The quadratic extrapolation function is given by,

$$\phi(\lambda) = \theta_0 + \theta_1\lambda + \theta_2\lambda^2.$$

where, the unknown parameters $\theta_0, \theta_1, \theta_2$ are estimated by fitting the equation to the points $(\lambda, \hat{\beta}(\lambda))$.

Then the fitted extrapolation function is evaluated at $\lambda = -1$.

When $\lambda = -1$ is substituted to (2.31), it gives

$$Var(X^{*(-1)}) = \sigma_X^2 + (1 - 1)\sigma_e^2 = \sigma_X^2.$$

This shows that after substitution it extrapolates back to the zero error case. Therefore, the SIMEX estimator can be defined as,

$$\hat{\beta}_{SIMEX} = \phi(-1). \quad (2.33)$$

Generally, the SIMEX estimator is an approximately consistent estimator and its performance relies on factors such as, validity of the assumed measurement error model, measurement error variance and the choice of the extrapolation function.

In the SIMEX method, additional variability is introduced, but averaging over a number of simulation replicates reduces this. Although, increasing the number of replications B improves stability, it becomes computationally expensive.

Overall, this chapter explained the concepts of the existing methodologies to correct for the errors in linear regression models, and these theoretical concepts and estimations are used as the fundamentals to develop the methods under censored and mismeasured covariates.

Chapter 3

Censored and Mismeasured Covariates

Following the measurement error correction methods from chapter 2, this chapter discusses the censored and mismeasured covariates, which is the main focus of this study. Here, the proposed instrumental variable estimation method for linear regression models is presented. Moreover, this chapter gives a brief introduction to the censored covariates and how the existing error correction methods can be modified after accounting for censoring.

3.1 Censored Covariates

Censoring is a very common scenario that arises in real world problems. In many cases, instead of the exact values, only partial information of the variables is available due to detection limits of instruments or data collection process. These variables are called censored variables. Missing data and censored data have different underlying interpretations. In missing data, no information is available on the unobserved values. Meanwhile, the censored data provides

partial information on the data. Therefore, censored data is useful for statistical inference. Focusing on the censored covariates, using standard regression methods when the covariates are not fully observed can lead to biased and inconsistent estimators.

Censoring can be classified into two categories as fixed and random censoring.

- Fixed Censoring - Occurs when the observations are censored at a fixed predetermined threshold and this censoring mechanism completely depends on the threshold. This is common in laboratory based studies due detection thresholds of instruments.
- Random Censoring - Occurs when the censoring threshold is not fixed and it differs among the observations. This type of censoring is common in longitudinal studies, where censoring times vary among individuals.

Moreover, there are three types of censoring, and they are as follows. Note that W is the observed value, η is the censoring indicator and C is the censoring threshold.

- Right Censoring - When the true value of the covariate exceeds the censoring threshold, where

$$W = \min(X^*, C) \quad \text{and} \quad \eta = I(X^* \leq C). \quad (3.1)$$

- Left Censoring - When the true value of the covariate lies below the censoring threshold, where

$$W = \max(X^*, C) \quad \text{and} \quad \eta = I(X^* \geq C). \quad (3.2)$$

- Interval Censoring - When the true value lies within a specific interval $C_1 < X^* \leq C_2$, where

$$W = \begin{cases} C_1, & X^* \leq C_1, \\ X^*, & C_1 < X^* < C_2, \\ C_2, & X^* \geq C_2, \end{cases} \quad \eta_L = I(X^* > C_1), \quad \eta_R = I(X^* < C_2). \quad (3.3)$$

To address this censored covariate problem, different methodological approaches have been introduced such as likelihood methods, imputation methods, complete case analysis, and Bayesian methods ([Lotspeich et al., 2024](#)). In this study the primary methodology used to address this problem is maximum likelihood estimation.

Furthermore, throughout this study, the non-informative censoring mechanism is assumed, which means the censored variable is independent of the measurement error as well as the regression error after conditioned on the true covariate.

3.2 Model

This section builds up a model to address the primary focus of this study which is censored and mismeasured covariates.

Consider the linear regression model,

$$Y = \beta'_x X + \beta'_z Z + \epsilon, \quad (3.3)$$

where, $Y \in \mathbb{R}$ is the response variable, $X \in \mathbb{R}^{p_x}$ and $Z \in \mathbb{R}^{p_z}$ are vectors of error-prone and error-free covariates respectively, ϵ is the random error and

β_x and β_z are unknown parameters. Note that, Z includes a column of ones representing the intercept β_0 .

Under the classical additive measurement error, the surrogate measurement for X is given by,

$$X^* = X + e, \quad (3.4)$$

where e is the measurement error independent of X .

However, X^* is censored, so we actually observe W . In this study, we consider two types of censoring mechanisms: one with fixed left censoring, where

$$W = \max(X^*, C), \quad (3.5)$$

and C is a known constant; the other with random right censoring, where

$$W = \min(X^*, C), \quad (3.6)$$

and C is a latent random variable following a certain distribution.

In addition, assume that a p_s -vector of instrumental variable (IV) S is available and is related to X as

$$X = \alpha_s S + \alpha_z Z + u, \quad (3.7)$$

where the random error u is independent of (S, Z) and has pdf $f_u(u; \theta)$.

In summary, two methodological challenges arise here where the observed covariate is affected by the measurement error and censoring at the same time. These challenges motivated to develop the methodologies discussed in the next sections.

3.3 Instrumental Variable Method

The proposed methodology for the above stated model is a two-step estimation procedure involving the instrumental variable method. The estimation procedure can be summarized as follows.

1. Estimation of the relationship between the latent covariate and the instrumental variable using a censored regression model.
2. Obtaining the fitted values using the estimated first stage model.
3. Estimation of the reduced form representation of the response related to the instrumental variable and the fully observed covariate.
4. Recovering the regression parameters by combining the reduced form and first stage estimates.

Instead of estimating the latent covariate directly from the observed surrogate, the proposed method uses an instrumental variable to construct the estimate of the covariate. The detailed explanation of the methodology is as follows:

First, substitute (3.7) into (3.4) and obtain the linear regression model of the latent covariate and the explanatory variables as follows:

$$X^* = \alpha_s S + \alpha_z Z + u + e. \quad (3.8)$$

Here, α_s, α_z represent the vectors of the regression coefficients associated with the instrumental variables and fully observed covariates, respectively.

Then, (3.8) and (3.5) or (3.6) form a censored linear model and, therefore, the unknown parameters (α_s, α_z) can be consistently estimated using either

fixed or random censoring methods according to the problem (Wang and Hsiao, 2007; Yi et al., 2020).

Under the **fixed censoring**, if the error term has a known parametric distribution, such as normal, then Tobit-type maximum likelihood method can be employed, which is commonly used when a continuous response or covariate is subject to the measurement error at a fixed threshold. Otherwise, a more robust censored least absolute deviation (CLAD) estimator or quantile regression methods can be used Wang and Hsiao (2007).

Under the **random censoring**, the parametric or semiparametric methods for the accelerated failure time (AFT) models can be used. The AFT model is a regression model designed for failure time or survival data where covariates have a multiplicative effect on the survival time. Moreover, the survival time is presented in a log linear form. In particular, the maximum likelihood method can be used if the error distribution has a known parametric form. More generally, semiparametric methods such as rank-based generalized estimating equations (GEE) are more robust and flexible (Jin et al., 2003).

The fitted values of the latent covariate, which serves as the basis of the second stage estimation are presented as, $\hat{X} = \hat{\alpha}_s S + \hat{\alpha}_z Z$, and after substituting $\hat{X}$ into (3.3), β_x and β_z can be estimated by the two-stage least squares method.

However, in this study, rather than using the two-stage least squares method, a more efficient estimator is derived using a reduced form representation Wang and Hsiao (2007).

Substituting (3.7) into (3.3) leads to the reduced form which contains only

the observable variables:

$$Y = \gamma'_s S + \gamma'_z Z + v, \quad (3.9)$$

where $\gamma_s = \alpha'_s \beta_x$, $\gamma_z = \alpha'_z \beta_x + \beta_z$ and $v = \beta'_x u + \varepsilon$. Therefore, γ_s and γ_z can be consistently estimated by the least squares method. Write $\gamma = (\gamma'_s, \gamma'_z)'$, $\beta = (\beta'_x, \beta'_z)'$, and their relations as $A\beta = \gamma$, where

$$A = \begin{pmatrix} \alpha'_s & 0 \\ \alpha'_z & I \end{pmatrix}.$$

For consistent estimators $\hat{\gamma}$ and $\hat{A}$, such that, as $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \Sigma_{\hat{\gamma}}), \quad (3.10)$$

we have that a minimum distance estimator for β is given by

$$\hat{\beta} = (\hat{A}' \hat{\Sigma}_{\hat{\gamma}}^{-1} \hat{A})^{-1} \hat{A}' \hat{\Sigma}_{\hat{\gamma}}^{-1} \hat{\gamma}. \quad (3.11)$$

Under standard regularity conditions, including independence of observations and validity of the instrumental variables and finite moments, the estimators obtained in the two-stage procedure are consistent and asymptotically normally distributed as

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, (A' \Sigma_{\hat{\gamma}}^{-1} A)^{-1}\right), \quad n \rightarrow \infty. \quad (3.12)$$

Therefore, this estimation procedure can be interpreted as a generalized two-stage estimation method, where the first stage estimates the relationship between the latent covariate and the instrumental variables under censoring, and the second stage recovers the parameters of the regression model. The

reduced form minimum distance estimator can be said as an efficient estimator since it used the estimated covariance structure of the reduced form coefficients rather than simply substituting the first stage estimates and recovering the structural parameters.

Note that throughout the asymptotic results, the covariance matrices are defined for scaled estimators. The covariance matrix of the unscaled estimator $\hat{\gamma}$ is $n^{-1}\Sigma_{\gamma}$. However, in the numerical analysis, when the *vcov()* R function is used, it directly estimates the covariance matrix of the unscaled estimator.

3.4 Regression Calibration Method (RC)

In order to compare this proposed methodology, existing methodologies were explored and designed according to the model. This section looks into the first comparison method, classical regression calibration (RC), which is a competing method for the proposed.

The key idea behind the RC method is to correct the measurement error by replacing unobserved true covariate X by its conditional expectation given the observed data, $\hat{X} = E[X|X^*, Z]$, which is known as the calibration function. The existing regression calibration method is derived under normality assumptions.

Given the regression model in section 2.3, the calibration function can be calculated as follows:

$$\hat{X}_i^{RC} = E(X_i | X_i^*, Z_i).$$

Case 1: Uncensored Observations

In this case, $W_i = X_i^*$. Therefore, the calibration estimator is

$$\hat{X}_i^{RC} = E(X_i \mid X_i^*, Z_i).$$

Under normality when $X \perp Z$,

$$\hat{X}_i^{RC} = \mu_X + \lambda(X_i^* - \mu_X) \quad , \quad \lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2}.$$

If $X_i \sim N(0, 1)$, then

$$\hat{X}_i^{RC} = \lambda X_i^*.$$

Case 2: Censored Observations

In this case, $W_i = c_i$ and the calibration estimator at right censoring can be presented as

$$\hat{X}_i^{RC} = E(X_i \mid X_i^* > c_i, Z_i).$$

Let

$$X_i^* = X_i + e_i \sim N(\mu_X, \sigma_{X^*}^2) \quad , \quad \sigma_{X^*}^2 = \sigma_X^2 + \sigma_e^2.$$

Define

$$a_i = \frac{c_i - \mu_X}{\sigma_{X^*}}.$$

Then, from properties of the truncated normal distribution, the conditional expectation is given by,

$$E(X_i^* \mid X_i^* > c_i) = \mu_X + \sigma_{X^*} \frac{\phi(a_i)}{1 - \Phi(a_i)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal probability density and cumulative distribution function, respectively.

Using the law of iterated expectation,

$$E(X_i \mid X_i^* > c_i) = E[E(X_i \mid X_i^*) \mid X_i^* > c_i].$$

After substituting the calibration equation we can obtain:

$$\hat{X}_i^{RC} = \mu_X + \lambda \sigma_{X^*} \frac{\phi(a_i)}{1 - \Phi(a_i)}. \quad (3.13)$$

Moreover, under left censoring the calibration equation can be written as

$$\hat{X}_i^{RC} = \mu_X + \lambda \sigma_{X^*} \frac{\phi(a_i)}{\Phi(a_i)}. \quad (3.14)$$

This estimator depends on the term $\frac{\phi(a_i)}{1 - \Phi(a_i)}$, and it is known as the Inverse Mills Ratio.

The regression model with the calibrated covariate can be presented as

$$Y_i = \beta_0 + \beta_x \hat{X}_i^{RC} + \beta_z Z_i + \varepsilon_i.$$

The regression coefficients can be estimated using the ordinary least squares method as presented below.

$$Y_i \sim \hat{X}_i^{RC} + Z_i,$$

$$\sum_{i=1}^n (1, \hat{X}_i^{RC}, Z_i)' [Y_i - (\beta_0 + \beta_x \hat{X}_i^{RC} + \beta_z Z_i)] = 0,$$

$$\hat{\beta}^{RC} = (M^\top M)^{-1} M^\top Y,$$

where,

$$M = \begin{pmatrix} 1 & \hat{X}_1^{RC} & Z'_1 \\ 1 & \hat{X}_2^{RC} & Z'_2 \\ \vdots & \vdots & \vdots \\ 1 & \hat{X}_n^{RC} & Z'_n \end{pmatrix}.$$

Following the RC extension to the model, the coming section focuses on the next comparison method, SIMEX and how it is designed for the model in this study.

3.5 Simulation Extrapolation Method (SIMEX)

This section describes how the SIMEX method can be used to address the censored and mismeasured covariates.

This is an approach that adds additional measurement error to the surrogate variable X^* at levels $\lambda \geq 0$, producing $W^{(\lambda)}$, re-estimates coefficients for each λ and extrapolates the resulting estimates to the case of no measurement error.

For each $\lambda \in \Lambda$ and for $b = 1, \dots, B$ where, $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$.

1. Generate an extra error $e^{*(b)}$.

$$X^{*(b)}(\lambda) = X^* + \sqrt{\lambda}e^{*(b)}, \text{ where } e^{*(b)} \text{ follows the same distribution as } e.$$

2. Obtain the observed surrogate after accounting for censoring as,

$$W^{*(b)}(\lambda) = \max(X^{*(b)}(\lambda), c),$$

with the corresponding censoring indicator $\eta_i = I(X^{*(b)}(\lambda) \geq c)$.

3. Estimate the regression parameters considering $(Y, X^{*(b)}(\lambda), Z)$.
4. For each simulated dataset, calculate estimate $\hat{\beta}^{(b)}(\lambda)$ using censored maximum likelihood or other semiparametric methods.
5. Repeat this simulation B times to obtain the average as follows:

$$\hat{\beta}(\lambda) = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{(b)}(\lambda). \quad (3.15)$$

Afterwards, fit a smooth function $\phi(\lambda)$ to the points $(\lambda, \hat{\beta}(\lambda))$ and a quadratic extrapolation function, $\phi(\lambda) = a_0 + a_1\lambda + a_2\lambda^2$ is used. Then extrapolate to $\lambda = -1$ and obtain, $\hat{\beta}_{SIMEX} = \phi(-1)$.

3.6 Naive Method

For the purpose of comparing the methods that account for the measurement error and censoring, a naive estimation approach, which ignores the measurement error and accounts only for the censoring problem, is implemented. This approach can lead to bias and inconsistent estimates. Here, X is assumed to be the true, observed covariate and the regression model is fitted as Y on (X, Z) accounting for censoring.

Chapter 4

Simulation Studies

This chapter presents the simulation studies conducted to evaluate the proposed IV method and compare this method with other existing methods. In this study, rather than assessing performance under one setting, five different simulation experiments were conducted. To evaluate the performance of the proposed estimation method, simulation studies were conducted under various setups and compared with the existing approaches of the Regression Calibration, SIMEX and naive methods, using different performance evaluation metrics, such as the mean estimate, standard deviation, root mean squared error, and coverage probability. In each simulation experiment, data was generated according to the regression model (3.3), where the covariate X is subject to measurement error and censoring.

As explained above, five experiments were done under the simulation studies. The Experiment 1 considered a baseline model with normal errors under fixed censoring and the Experiment 2 extended the same model to a random censoring scenario and random censoring was considered in the other experiments. The Experiments 3 and 4 considered models under non-normal response errors and

the Experiment 5 evaluated the performance under heteroskedastic errors in the model.

The simulation parameters were chosen to represent realistic regression settings. In the latent covariate model (3.7), the coefficient α_z was set to a value and the value of the coefficient α_s was selected after trying different values starting from smaller values to ensure that the instrumental variable has a considerable effect on the latent covariate. Moreover, the variance of the latent error term u was chosen to generate the expected correlation between the instrumental variable to satisfy the IV assumptions. In all the Experiments 1 to 5, the correlations were selected to be approximately between 0.7 to 0.8, and the measurement error variance e was selected to achieve a reliability ratio of 0.7 according to (2.6) explained in Chapter 2.

All five experiments were done for different sample sizes ($n = 50, 200, 500$) and censoring levels (10% and 50%) for 2000 replications. The performance of the estimators were evaluated using the mean estimate (Est), Monte Carlo standard deviation (MCSD), root mean squared error (RMSE), and coverage probability (CP) at the nominal level 0.95.

The proposed instrumental variable estimator was implemented using a two-step estimation. The first stage considered the relationship between $X^* = \alpha_s S + \alpha_z Z + u + e$, which is a censored regression model. When u and e are independent and normally distributed as $u \sim N(0, \sigma_u^2)$ and $e \sim N(0, \sigma_e^2)$, a tobit regression model is appropriate, while for the other scenarios semiparametric packages were used in the estimation.

In the simulation studies, to recover the first stage parameters, for both fixed and random censoring cases, when the errors are normally distributed,

the function *survreg()* under the package *survival* was used, which fits an Accelerated Failure Time (AFT) model using maximum likelihood estimation. The function *Surv()* was used to specify the censored response with *type*=“*left*” and *type*=“*right*” to indicate the left and right censored responses, respectively, and *dist*=“*gaussian*” to specify the error distribution. When the normality assumption is violated, the function *aftsrr()* was used under the package *aftgee*, which fits a semiparametric AFT model. The response was transformed using the exponential function and the observed response and indicator were specified using the *Surv()* function. The parameter estimation was done on the rank-based estimating equations, specified by the option *rankWeights*=“*gehan*”, which is the Gehan’s weighting scheme. The second stage parameters were recovered using a minimum distance estimation as explained in Chapter 3.

The regression calibration estimator was obtained by computing $\hat{X}_i^{RC} = E(X_i | W_i, Z_i)$, for both uncensored and censored observations, as explained in section 3.4, with X replaced by $\hat{X}^{RC}$. The estimates were recovered using the ordinary least squares estimation. Moreover, the SIMEX procedure was implemented as described in section 3.5. For each $\lambda \in \Lambda$ and for $b = 1, \dots, B$ where, $\lambda = seq(0.1, 1.0, by = 0.2)$, $B = 100$ was considered and the R packages *maxLik* and *censCov* were used in the estimation.

To emphasize the importance of accounting for the measurement error, a naive estimation approach was implemented which ignores the measurement error and accounts only for the censoring problem. This approach can lead to bias and inconsistent estimates. We assumed X to be the true observed covariate and fitted the regression model Y on (X, Z) accounting for censoring.

Performance Measures

The performance of the estimators were evaluated using the mean estimate (Est), Monte Carlo standard deviation(MCSD), root mean squared error (RMSE), and coverage probability (CP) based on ($R = 2000$) Monte Carlo replications.

Let $\hat{\beta}_r$ denote the estimate of a parameter β obtained from the r^{th} simulation replication, $r = 1, \dots, R$, and let

$$\bar{\beta} = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_r \quad (4.1)$$

denote the average estimate across all replications.

The mean estimate (Est) is calculated as

$$\text{Est} = \bar{\beta} = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_r. \quad (4.2)$$

The Monte Carlo standard deviation (MCSD) is a measure of variability of the estimator across repeated samples and is given by

$$MCSD = \sqrt{\frac{1}{R-1} \sum_{r=1}^R (\hat{\beta}_r - \bar{\beta})^2}. \quad (4.3)$$

The root mean squared error (RMSE) combines both bias and variability and is defined as

$$RMSE = \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\beta}_r - \beta)^2}. \quad (4.4)$$

The coverage probability (CP) evaluates the accuracy of the estimated confidence intervals. Let $CI_r = (L_r, U_r)$ denote the nominal 95% confidence interval obtained from the r^{th} replication. The coverage probability is calculated as

$$CP = \frac{1}{R} \sum_{r=1}^R I(L_r \leq \beta \leq U_r), \quad (4.5)$$

where $I(\cdot)$ is the indicator function.

For each simulation setting, the estimators with smaller RMSE values and coverage probabilities closer to 0.95 were considered to have good performance.

4.1 Experiment 1

The first simulation experiment is based on a baseline setting where it reflects studies where a continuous outcome is used. For example, the response can be the systolic blood pressure, the latent covariate can represent the true blood glucose concentration, while the fully observed covariate can be the age. The blood glucose level can be subject to the measurement error and censoring due to instrument errors and detection limits respectively. In this experiment, both the response and measurement errors are generated from normal distributions. The setup of Experiment 1 is as follows:

1. Generate $(S, Z)' \sim N(\mu, \Sigma)$ and $(\varepsilon, e)' \sim N(0, (\sigma_\varepsilon^2, \sigma_e^2))$.

$$\begin{pmatrix} S \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_s \\ \mu_z \end{pmatrix}, \begin{pmatrix} \sigma_s^2 & \rho\sigma_s\sigma_z \\ \rho\sigma_s\sigma_z & \sigma_z^2 \end{pmatrix}\right),$$

where $\mu_s = 0, \mu_z = 0, \rho = 0.4, \sigma_s^2 = 0.5$, and $\sigma_z^2 = 0.5$.

The means are set to be 0 and the variances and the correlation $\rho = 0.4$ are chosen to have a moderate association between the fully observed covariate and the instrumental variable.

2. Generate $X = \alpha_s S + \alpha_z Z + u$, where $u \sim N(0, \sigma_u^2)$, $\alpha_s = 2, \alpha_z = 1$, and $\sigma_u^2 = 1.5$. This led to the correlation between the instrumental variable and the latent covariate to be approximately $\text{Corr}(X, S) = 0.775$, satisfying the relevance assumption (A 2.4) explained in Chapter 2.
3. Generate $X^* = X + e$, where $e \sim N(0, \sigma_e^2)$ and $\sigma_e^2 = 3.2$. Let $W = \max(X^*, c)$, where $\sigma_e^2 = 3.2$ was chosen to achieve the reliability ratio of 0.7 explained in (2.6).
4. Generate $Y = \beta'_x X + \beta'_z Z + \varepsilon$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 = 1$.

The regression variance was fixed to 1 under a standard linear regression model to provide a baseline signal-to-noise ratio (0.811).

The simulation results for Experiment 1 are as follows:

Table 4.1: Comparison of methods under 10% censoring and different sample sizes for Experiment 1.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	0.981	0.130	0.131	0.937	0.990	0.133	0.133	0.945	0.991	0.084	0.085	0.948
	RC	0.928	0.134	0.152	0.918	0.938	0.157	0.169	0.936	0.934	0.099	0.119	0.941
	SIMEX	1.154	0.315	0.351	0.890	0.931	0.215	0.226	0.892	0.932	0.136	0.152	0.908
	Naive	0.889	0.497	0.509	0.576	0.880	0.354	0.374	0.835	0.881	0.224	0.254	0.856
$\beta_x = 1.2$	IV	1.191	0.082	0.082	0.931	1.195	0.055	0.055	0.946	1.198	0.035	0.035	0.948
	RC	1.109	0.121	0.152	0.928	1.128	0.057	0.092	0.935	1.112	0.036	0.095	0.941
	SIMEX	1.055	0.240	0.281	0.916	1.162	0.187	0.190	0.900	1.170	0.118	0.122	0.921
	Naive	0.921	0.399	0.487	0.543	0.717	0.231	0.535	0.570	0.721	0.146	0.501	0.600
$\beta_z = 2.0$	IV	1.971	0.107	0.111	0.941	1.981	0.070	0.071	0.944	1.988	0.044	0.046	0.945
	RC	1.956	0.151	0.157	0.922	1.961	0.114	0.120	0.938	1.960	0.072	0.082	0.943
	SIMEX	2.010	0.212	0.213	0.901	2.058	0.181	0.190	0.928	2.010	0.114	0.115	0.936
	Naive	2.143	0.435	0.458	0.692	2.411	0.247	0.480	0.690	2.423	0.156	0.451	0.718

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table 4.2: Comparison of methods under 50% censoring and different sample sizes for Experiment 1.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	0.980	0.225	0.226	0.928	0.992	0.166	0.166	0.942	0.995	0.105	0.105	0.947
	RC	0.918	0.260	0.274	0.910	0.938	0.204	0.217	0.931	0.941	0.129	0.142	0.939
	SIMEX	0.895	0.395	0.410	0.815	0.921	0.299	0.307	0.850	0.928	0.189	0.202	0.870
	Naive	0.875	0.610	0.623	0.551	0.887	0.427	0.460	0.825	0.891	0.270	0.291	0.852
$\beta_x = 1.2$	IV	1.186	0.215	0.216	0.930	1.198	0.164	0.164	0.943	1.196	0.104	0.104	0.946
	RC	1.095	0.245	0.267	0.912	1.128	0.176	0.194	0.930	1.121	0.111	0.136	0.938
	SIMEX	1.045	0.365	0.397	0.795	1.110	0.261	0.310	0.810	1.102	0.165	0.192	0.842
	Naive	0.865	0.860	0.923	0.430	0.888	0.723	0.788	0.490	0.891	0.457	0.565	0.520
$\beta_z = 2.0$	IV	1.972	0.155	0.158	0.932	1.988	0.119	0.120	0.940	1.990	0.075	0.076	0.941
	RC	1.945	0.205	0.214	0.918	1.957	0.163	0.169	0.932	1.961	0.103	0.110	0.939
	SIMEX	2.030	0.285	0.289	0.885	2.023	0.215	0.216	0.905	2.021	0.136	0.138	0.913
	Naive	2.345	0.670	0.754	0.565	2.311	0.531	0.615	0.620	2.324	0.336	0.461	0.652

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

The simulation results for the model with the normal errors under fixed censoring of 10% and 50% are presented in Tables 4.1 and 4.2, respectively. Under this setting, the proposed IV estimator produces estimates closest to

the true parameter value. For example, according to Table 4.1, under 10% censoring, the estimate of $\beta_x = 1.2$ is 1.191 for $n = 50$ and with increasing sample size to $n = 500$, the estimate becomes 1.198 which is much closer to the true value. Meanwhile, the respective RMSE values decrease from 0.082 to 0.035. A similar behavior is observed on the other parameters β_0 and β_z .

When the censoring proportion increases from 10% to 50%, the performance decreases for the proposed and competing methods but the proposed method shows relatively better results. For the IV method, even under 50% censoring, the estimate of β_x remains close to the true value. (1.186 for $n = 50$ and 1.196 for $n = 500$). Moreover, the coverage probability exceeds 0.93 across all sample sizes while competing methods show sensitivity to higher censoring levels. For the RC method, the coverage probability converges slowly to the nominal level, and especially, the coverage probability for the SIMEX method decreases drastically. For example, in Table 4.2 it can be observed that the coverage probability for β_x under $n = 50$ is only 0.795 and increases slowly with increasing sample size. As expected, the naive method shows the lowest performance with low coverage probabilities (0.430 for $n=50$ and 0.520 for $n=500$). The trend remains the same for other covariates as well.

The following graph shows a summary of how the RMSE values change across the scenarios considered.

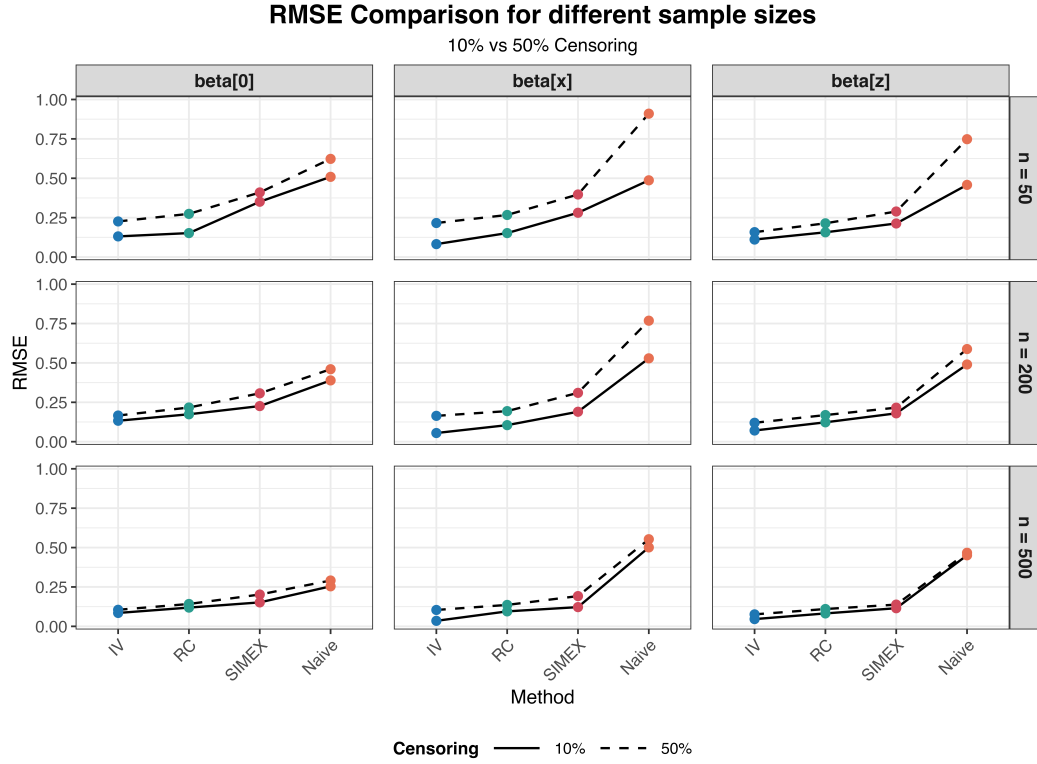


Figure 4.1: Comparison of RMSE values under the considered scenarios in Experiment 1.

4.2 Experiment 2

Experiment 2 extends Experiment 1 by replacing fixed censoring with random censoring. Random censoring mechanism can be commonly noticed in longitudinal and clinical studies due to various reasons. Therefore, it can be said that the Experiment 2 represents a more widely observed censoring mechanism

while keeping the setup of Experiment 1 as it is. As the Experiment 1, the response can be the systolic blood pressure, the latent covariate can represent the true blood glucose concentration, while the fully observed covariate can be the age. In this case, the blood glucose level can be considered to be randomly censored due to the difference of observation limits across individuals.

As explained above, Experiment 1 is kept as is and the censoring is changed to random censoring, where $c_i \sim u(c_0 - \delta, c_0 + \delta)$, $c_0 = 2.1$ (10% censoring), $c_0 = 1.2$ (50% censoring), and $\delta_c = 0.5$. Let $W = \min(X^*, c_i)$ and the censoring indicator $\Delta_i = 1; X_i^* \leq c_i$.

The simulation results for Experiment 2 are as follows:

Table 4.3: Comparison of methods under 10% censoring and different sample sizes for Experiment 2.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	0.983	0.188	0.187	0.925	0.988	0.144	0.143	0.943	0.992	0.091	0.091	0.946
	RC	0.910	0.202	0.221	0.916	0.932	0.171	0.184	0.928	0.931	0.108	0.128	0.936
	SIMEX	0.875	0.368	0.387	0.867	0.901	0.234	0.254	0.875	0.928	0.148	0.165	0.895
	Naive	0.840	0.542	0.565	0.791	0.879	0.371	0.390	0.825	0.878	0.235	0.265	0.844
$\beta_x = 1.2$	IV	1.195	0.132	0.132	0.928	1.195	0.066	0.066	0.944	1.197	0.042	0.042	0.946
	RC	1.080	0.142	0.187	0.910	1.105	0.071	0.119	0.928	1.108	0.045	0.102	0.936
	SIMEX	1.140	0.302	0.308	0.840	1.168	0.201	0.204	0.885	1.162	0.127	0.133	0.905
	Naive	0.670	0.510	0.735	0.470	0.711	0.255	0.549	0.560	0.715	0.161	0.511	0.587
$\beta_z = 2.0$	IV	1.975	0.162	0.164	0.930	1.978	0.081	0.082	0.942	1.989	0.051	0.052	0.944
	RC	1.922	0.256	0.268	0.915	1.943	0.128	0.140	0.932	1.958	0.081	0.091	0.938
	SIMEX	2.030	0.335	0.338	0.890	2.022	0.198	0.198	0.915	2.012	0.125	0.126	0.924
	Naive	2.521	0.540	0.750	0.608	2.438	0.270	0.515	0.675	2.431	0.171	0.464	0.701

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table 4.4: Comparison of methods under 50% censoring and different sample sizes for Experiment 2.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	0.982	0.255	0.256	0.920	0.994	0.187	0.187	0.940	0.996	0.118	0.118	0.943
	RC	0.902	0.295	0.312	0.895	0.920	0.225	0.239	0.922	0.937	0.142	0.155	0.929
	SIMEX	0.885	0.465	0.480	0.760	0.908	0.324	0.335	0.824	0.921	0.205	0.219	0.852
	Naive	0.845	0.640	0.660	0.480	0.870	0.452	0.470	0.800	0.885	0.286	0.308	0.834
$\beta_x = 1.2$	IV	1.192	0.265	0.266	0.922	1.196	0.187	0.187	0.939	1.198	0.118	0.118	0.943
	RC	1.072	0.280	0.310	0.900	1.111	0.199	0.218	0.920	1.115	0.126	0.152	0.926
	SIMEX	1.040	0.430	0.459	0.730	1.078	0.292	0.316	0.782	1.085	0.185	0.217	0.821
	Naive	0.620	0.980	1.139	0.350	0.831	0.778	0.861	0.481	0.881	0.492	0.586	0.511
$\beta_z = 2.0$	IV	1.970	0.190	0.192	0.924	1.988	0.136	0.136	0.938	1.991	0.086	0.086	0.940
	RC	1.915	0.255	0.270	0.905	1.961	0.182	0.186	0.925	1.957	0.115	0.123	0.932
	SIMEX	2.040	0.396	0.397	0.865	2.029	0.239	0.240	0.893	2.025	0.151	0.153	0.901
	Naive	2.560	0.760	0.944	0.500	2.423	0.572	0.711	0.611	2.331	0.362	0.491	0.635

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

The above Tables 4.3 and 4.4 include the results of the Experiment 1 setting under random censoring. The results show a similar trend as in the fixed censoring case, but the performance is slightly low compared to the fixed censoring case. According to Table 4.3, under 10% censoring, for the IV method, the estimate of $\beta_x = 1.2$ is 1.195 at $n = 50$ and with the increase of sample size, the estimate becomes 1.197, which is closer to the true value. Moreover, the lowest RMSE value and the highest coverage probabilities can be observed for IV estimates.

Moving from 10% censoring to 50% censoring, the trend remains the same as the Experiment 1. Even though the RC method performs closely with the IV method, the coverage probability is slightly less compared to the IV method under higher censoring for β_x (0.922 and 0.900 coverage probabilities for the IV method and RC method respectively). The trend is similar for the other parameters β_0 and β_z as well.

When random censoring is introduced to the same Experiment 1, compar-

atively large RMSE values and lower coverage probabilities can be observed for all the methods, while the IV estimator shows better performance out of the methods. For example, according to Table 4.1, under fixed censoring and $n = 500$ the coverage probability for the estimate $\beta_x = 1.2$ is 0.948, and Table 4.3 shows that under random censoring for the same scenario the coverage probability is 0.946, which is slightly low. Overall, according to Tables 4.1 and 4.3 for the coverage probability, the rate of convergence to the nominal level is lower when random censoring is considered.

The following graph shows a summary of how the RMSE values change across the scenarios considered.

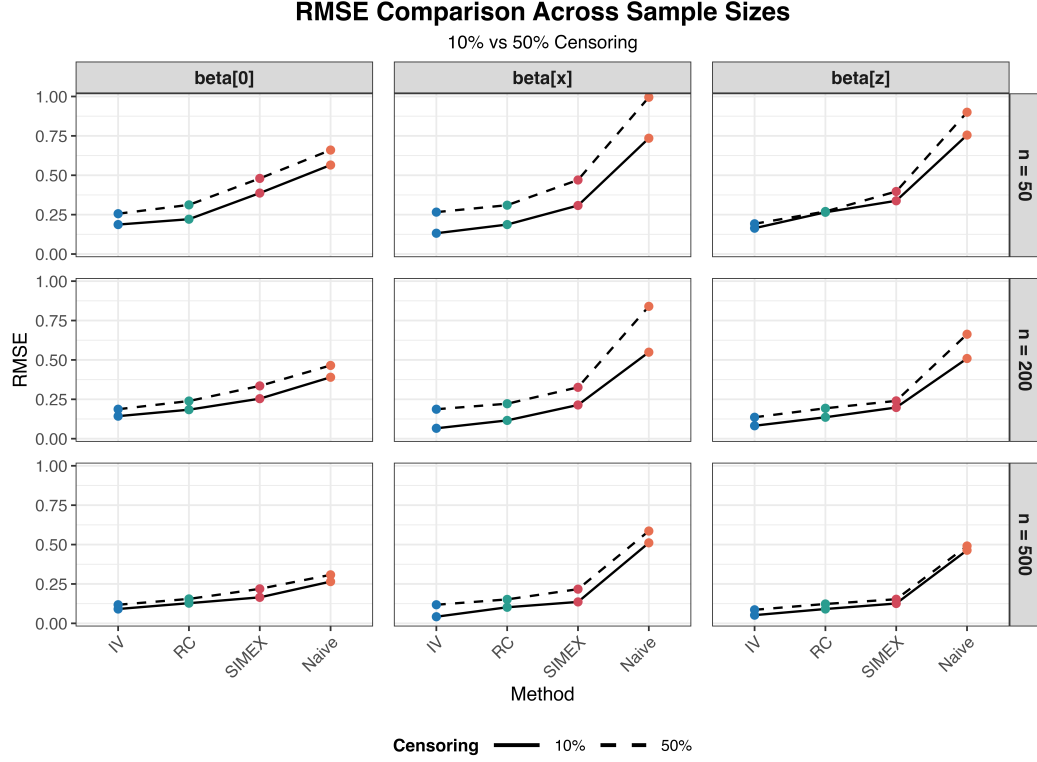


Figure 4.2: Comparison of RMSE values under the considered scenarios in Experiment 2

4.3 Experiment 3

Experiment 3 looks into a more complex simulation setting where two continuous and binary, fully observed covariates are present along with a randomly censored and mismeasured covariate, and a t -distributed response error. The same real world example explained in the above experiments can be adapted to this setting as the systolic blood pressure to be the continuous outcome and the blood glucose level to be the censored and mismeasured covariate. Additional risk factors such as age and smoking status can be included as continuous

and binary covariates respectively. Moreover, covariates like hemoglobin is directly related to blood glucose level and can be considered as an instrumental variable. Since these kind of biomarkers are positive and right skewed, the instrumental variable was generated from a log normal distribution. Further, the response error is generated from a non-normal, heavy tailed t -distribution. The simulation setup for Experiment 3 is as follows:

1. Generate $S \sim \text{LogNormal}(\mu, \sigma^2)$ where, $\mu = 0, \sigma^2 = 0.5$
2. Consider two fully observed covariates.

Generate $Z_1 \sim t(v), v = 5$ and $Z_2 \sim \text{Bernoulli}(p), p = 0.3$.

3. Generate $X = \alpha_s S + \alpha_{z1} Z_1 + \alpha_{z2} Z_2 + u$, where $u \sim N(0, \sigma_u^2)$, $\alpha_s = 2$, $\alpha_{z1} = 1$, $\alpha_{z2} = 1$, and $\sigma_u^2 = 4.5$, leading to the correlation between the instrumental variable and the latent covariate to be approximately $\text{Corr}(X, S) = 0.750$, satisfying the relevance assumption (A 2.4) explained in Chapter 2.
4. Generate $X^* = X + e$, where $e \sim N(0, \sigma_e^2)$ and $\sigma_e^2 = 6.5$, leading to the reliability ratio of 0.7 explained in (2.6).
5. Consider random right censoring.

Generate $c_i \sim u(c_0 - \delta, c_0 + \delta)$, $c_0 = 3.6$ (10% censoring), $c_0 = 1.5$ (50% censoring), $\delta_c = 0.5$ and let $W = \min(X^*, c_i)$
and the censoring indicator $\Delta_i = 1; X_i^* \leq c_i$.

6. Generate $Y = \beta_x X + \beta_{z1} Z_1 + \beta_{z2} Z_2 + \varepsilon = \beta'_x X + \beta'_z Z + \varepsilon$, with $\varepsilon \sim t(v)$ and $v = 5$. According to the error variance chosen, the signal to noise ratio is 0.796.

The simulation results for Experiment 3 are as follows:

Table 4.5: Comparison of methods under 10% censoring and different sample sizes for Experiment 3.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	1.020	0.224	0.225	0.925	1.015	0.114	0.115	0.943	1.010	0.074	0.075	0.950
	RC	0.970	0.262	0.264	0.917	0.975	0.121	0.124	0.941	0.980	0.108	0.110	0.947
	SIMEX	0.881	0.261	0.287	0.900	0.965	0.131	0.136	0.928	0.963	0.116	0.122	0.937
	Naive	0.820	0.396	0.435	0.830	0.832	0.330	0.370	0.880	0.830	0.297	0.342	0.901
$\beta_x = 1.2$	IV	1.195	0.151	0.151	0.924	1.198	0.121	0.121	0.942	1.200	0.076	0.076	0.946
	RC	1.160	0.178	0.182	0.910	1.151	0.120	0.130	0.930	1.161	0.109	0.116	0.941
	SIMEX	1.018	0.210	0.278	0.870	1.101	0.156	0.185	0.908	1.112	0.113	0.143	0.925
	Naive	0.721	0.343	0.589	0.507	0.823	0.187	0.421	0.545	0.954	0.260	0.358	0.607
$\beta_{z1} = 2.0$	IV	1.990	0.211	0.211	0.947	1.996	0.120	0.120	0.955	2.000	0.076	0.076	0.956
	RC	2.021	0.233	0.234	0.940	2.018	0.143	0.144	0.942	2.010	0.132	0.132	0.946
	SIMEX	1.851	0.257	0.297	0.921	1.960	0.209	0.213	0.938	1.960	0.183	0.187	0.945
	Naive	1.838	0.377	0.410	0.658	1.667	0.283	0.437	0.823	1.670	0.210	0.391	0.882
$\beta_{z2} = 3.2$	IV	3.218	0.277	0.278	0.948	3.185	0.152	0.153	0.951	3.190	0.096	0.096	0.953
	RC	3.181	0.322	0.323	0.930	3.210	0.178	0.178	0.941	3.210	0.137	0.137	0.942
	SIMEX	3.153	0.345	0.348	0.910	3.312	0.194	0.224	0.938	3.160	0.208	0.212	0.944
	Naive	3.402	0.424	0.470	0.868	3.160	0.423	0.425	0.882	3.332	0.316	0.343	0.900

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table 4.6: Comparison of methods under 50% censoring and different sample sizes for Experiment 3.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	1.015	0.275	0.276	0.922	1.005	0.182	0.182	0.940	1.008	0.115	0.115	0.948
	RC	0.940	0.285	0.291	0.905	0.950	0.186	0.193	0.928	0.955	0.111	0.120	0.942
	SIMEX	0.865	0.310	0.338	0.850	0.884	0.210	0.240	0.892	0.900	0.112	0.150	0.910
	Naive	0.815	0.470	0.505	0.763	0.830	0.406	0.440	0.860	0.830	0.400	0.435	0.880
$\beta_x = 1.2$	IV	1.188	0.215	0.216	0.919	1.191	0.142	0.142	0.938	1.198	0.096	0.096	0.943
	RC	1.105	0.190	0.212	0.902	1.120	0.128	0.151	0.925	1.130	0.112	0.132	0.939
	SIMEX	1.260	0.245	0.252	0.840	1.302	0.176	0.203	0.880	1.121	0.158	0.177	0.905
	Naive	0.700	0.420	0.653	0.420	0.820	0.248	0.454	0.481	0.820	0.340	0.510	0.600
$\beta_{z1} = 2.0$	IV	1.995	0.235	0.236	0.940	2.010	0.153	0.153	0.948	2.005	0.115	0.115	0.950
	RC	1.960	0.255	0.260	0.920	1.970	0.189	0.191	0.935	1.980	0.137	0.138	0.942
	SIMEX	1.900	0.315	0.330	0.900	1.950	0.240	0.245	0.930	1.950	0.199	0.205	0.940
	Naive	1.600	0.455	0.606	0.633	1.650	0.328	0.480	0.784	1.670	0.252	0.415	0.860
$\beta_{z2} = 3.2$	IV	3.195	0.255	0.255	0.942	3.180	0.163	0.164	0.949	3.200	0.130	0.130	0.948
	RC	3.180	0.335	0.336	0.918	3.200	0.260	0.260	0.925	3.186	0.164	0.165	0.940
	SIMEX	3.380	0.310	0.360	0.905	3.350	0.238	0.281	0.932	3.162	0.232	0.235	0.940
	Naive	3.120	0.610	0.615	0.720	3.150	0.485	0.488	0.786	3.320	0.373	0.392	0.890

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Tables 4.5 and 4.6 present how the methods behave under a t distributed response error. The proposed IV method continue to show low RMSE values and high coverage probabilities even under heavy tailed distributions. According to Table 4.5, under 10% censoring the IV method produced estimates 1.200, 2.000 and 3.190 for the parameters β_x, β_{z1} and β_{z2} respectively under $n = 500$ with a range of coverage probabilities 0.946-0.956. When the censoring level increases the effect of heavy tailed errors can be noticed but still the IV method performs well compared to the other competing methods. The RC estimator shows increased bias while the SIMEX method show less stable results with overestimated $\beta_x = 1.2$ for small sample sizes (in Table 4.6 for $n = 50$ the estimate is 1.260). As expected, the naive method shows the lowest results, where, for $n = 50$, the coverage probability is only 0.420.

It can be noticed that under 50% censoring, the effect of the heavy tailed errors become more prominent due to the higher variability in errors and the

higher censoring level. For the IV method, according to Table 4.6, the estimates of β_x ranged from 1.88 to 1.198 across the sample sizes with the RMSE values ranging from 0.216 to 0.096. Therefore, it can be said that even under a complex setting, the IV method works well under larger sample sizes. Even though the experiments are different from each other, the overall conclusion remains unchanged.

The following graph shows a summary of how the RMSE values change across the scenarios considered.

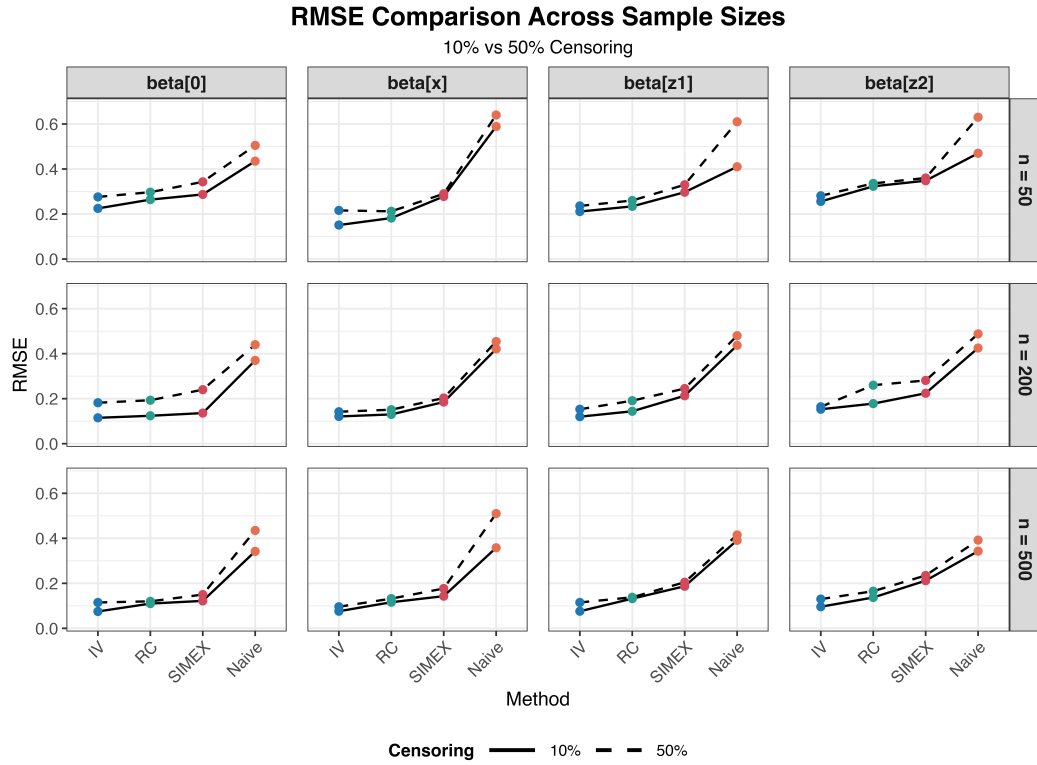


Figure 4.3: Comparison of RMSE values under the considered scenarios in Experiment 3

4.4 Experiment 4

Experiment 4 extends to a more complex setup, where the response error is considered to be positively skewed (centralized χ^2 distribution) and the variability of the error is not symmetric. The setup in Experiment 4 can be observed in the real world. Using the same example as the previous experiments, systolic blood pressure can be the response variable, while blood glucose level can represent the randomly censored and mismeasured covariate due to the varying threshold across measurements and laboratory errors, respectively. A repeated laboratory measurement can be assumed to be a normally distributed Instrument. This experiment includes a fully observed positively skewed covariate which can be a variable such as triglyceride concentration or protein level. Meanwhile, the fully observed binary variable can be the smoking status. The simulation setup for Experiment 4 is as follows:

1. Generate $S \sim N(\mu_s, \sigma_s^2)$ where, $\mu_s = 0, \sigma_s^2 = 1$.
2. Consider two fully observed covariates.

Generate $Z_1 \sim \text{LogNormal}(\mu_{z1}, \sigma_{z1}^2)$ where, $\mu_{z1} = 0, \sigma_{z1}^2 = 0.25$.

$Z_2 \sim \text{Bernoulli}(p)$ where, $p = 0.4$.

3. Generate $X = \alpha_s S + \alpha_{z1} Z_1 + \alpha_{z2} Z_2 + u$, where $u \sim N(0, \sigma_u^2)$, $\sigma_u^2 = 3.5$, $\alpha_s = 2$, $\alpha_{z1} = 0.3$, and $\alpha_{z2} = 0.5$, producing the correlation between the instrumental variable and the latent covariate to be approximately $\text{Corr}(X, S) = 0.760$, satisfying the relevance assumption (A 2.4) explained in Chapter 2.

4. Generate $X^* = X + e$, where $e \sim N(0, \sigma_e^2)$ and $\sigma_e^2 = 5.5$, leading to the reliability ratio of 0.7 explained in (2.6).
5. Consider random right censoring.

Generate $c_i \sim u(c_0 - \delta, c_0 + \delta)$ and let $W = \min(X^*, c_i)$,
 and the censoring indicator, $\Delta_i = 1; X_i^* \leq c_i$, where $c_0 = 3.9$ (10%
 censoring), $c_0 = 1.0$ (50% censoring) and $\delta = 0.5$.

6. Generate $Y = \beta'_x X + \beta'_{z1} Z_1 + \beta'_{z2} Z_2 + \varepsilon = \beta'_x X + \beta'_z Z + \varepsilon$, where $\varepsilon \sim \chi^2(v)$ with $v = 1$. Under the centralized chi-squared distribution, $\varepsilon = \chi^2(v) - v$. So, $E(\varepsilon) = 0$ and $var(\varepsilon) = 2v$. According to the error variance chosen the signal to noise ratio is 0.791.

The simulation results for Experiment 4 are as follows:

Table 4.7: Comparison of methods under 10% censoring and different sample sizes for Experiment 4.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	0.911	0.202	0.221	0.922	0.998	0.161	0.161	0.934	1.010	0.113	0.113	0.941
	RC	0.865	0.252	0.286	0.918	0.992	0.173	0.173	0.924	1.049	0.123	0.132	0.935
	SIMEX	0.781	0.289	0.362	0.891	0.942	0.197	0.205	0.901	0.965	0.194	0.197	0.937
	Naive	0.701	0.455	0.544	0.592	0.891	0.377	0.392	0.891	0.923	0.250	0.262	0.897
$\beta_x = 1.2$	IV	1.174	0.195	0.210	0.921	1.193	0.105	0.105	0.938	1.199	0.098	0.098	0.941
	RC	1.159	0.229	0.245	0.902	1.178	0.121	0.123	0.923	1.185	0.098	0.113	0.934
	SIMEX	1.142	0.437	0.441	0.855	1.152	0.198	0.204	0.864	1.154	0.166	0.172	0.887
	Naive	0.842	0.350	0.451	0.431	0.941	0.319	0.393	0.542	1.031	0.274	0.292	0.578
$\beta_{z1} = 2.0$	IV	2.006	0.208	0.208	0.924	1.996	0.120	0.120	0.937	2.002	0.065	0.065	0.941
	RC	2.060	0.261	0.268	0.919	2.034	0.130	0.134	0.928	1.978	0.126	0.128	0.932
	SIMEX	1.842	0.267	0.310	0.901	1.978	0.158	0.160	0.921	1.962	0.124	0.130	0.941
	Naive	1.854	0.390	0.416	0.654	1.908	0.336	0.348	0.822	1.774	0.171	0.292	0.788
$\beta_{z2} = 3.2$	IV	3.241	0.211	0.215	0.925	3.202	0.111	0.111	0.939	3.203	0.105	0.105	0.942
	RC	3.151	0.213	0.219	0.924	3.189	0.154	0.165	0.926	3.209	0.152	0.165	0.938
	SIMEX	3.375	0.269	0.386	0.898	3.143	0.193	0.240	0.918	3.158	0.154	0.199	0.924
	Naive	3.487	0.348	0.493	0.586	3.308	0.407	0.504	0.746	3.002	0.219	0.295	0.854

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table 4.8: Comparison of methods under 50% censoring and different sample sizes for Experiment 4.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	0.900	0.295	0.314	0.915	0.990	0.210	0.210	0.932	0.992	0.198	0.198	0.940
	RC	0.850	0.335	0.368	0.900	0.980	0.242	0.243	0.919	0.980	0.209	0.210	0.929
	SIMEX	0.760	0.405	0.471	0.855	0.900	0.304	0.320	0.886	0.920	0.216	0.230	0.919
	Naive	0.680	0.670	0.729	0.500	0.850	0.560	0.580	0.754	0.880	0.384	0.402	0.865
$\beta_x = 1.2$	IV	1.160	0.215	0.223	0.914	1.171	0.106	0.110	0.928	1.160	0.080	0.100	0.940
	RC	1.070	0.255	0.283	0.892	1.090	0.138	0.171	0.914	1.120	0.122	0.144	0.932
	SIMEX	0.940	0.340	0.450	0.820	0.987	0.227	0.311	0.862	0.920	0.128	0.308	0.876
	Naive	0.690	0.480	0.720	0.360	0.750	0.342	0.565	0.513	0.831	0.281	0.464	0.523
$\beta_{z1} = 2.0$	IV	2.010	0.225	0.226	0.918	2.020	0.131	0.133	0.925	1.996	0.092	0.092	0.940
	RC	2.000	0.270	0.271	0.908	1.987	0.150	0.151	0.916	1.990	0.178	0.178	0.930
	SIMEX	1.880	0.340	0.373	0.885	1.949	0.246	0.251	0.903	1.978	0.209	0.210	0.927
	Naive	1.800	0.550	0.588	0.620	1.870	0.439	0.458	0.784	1.912	0.285	0.298	0.786
$\beta_{z2} = 3.2$	IV	3.220	0.205	0.207	0.921	3.195	0.121	0.121	0.936	3.194	0.121	0.121	0.942
	RC	3.170	0.355	0.356	0.912	3.187	0.286	0.287	0.921	3.190	0.210	0.210	0.927
	SIMEX	3.120	0.330	0.395	0.890	3.151	0.261	0.300	0.911	3.182	0.217	0.283	0.904
	Naive	2.950	0.540	0.595	0.418	3.020	0.383	0.483	0.498	3.110	0.345	0.362	0.796

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Further investigation of performance under non-normal distributions extended to a positively skewed centralized χ^2 distribution. Although the RMSE values and coverage probabilities are low compared to those in the other experiments, the general conclusion remains the same.

Simulation results when the regression errors follow a non-normal χ^2 error distribution are presented in Tables 4.7 and 4.8. Given the more challenging setting with two fully observed covariates and non-normal skewed errors, the IV method continued to perform reasonably well compared to the other competing methods. For example, in Table 4.7, under 10% censoring and $n = 500$, for β_x, β_{z1} and β_{z2} the estimates are 1.199, 2.002 and 3.203 respectively and the estimates are close to the true value. Moreover, the RMSE values are below 0.11 and the coverage probabilities coverage to the nominal level.

More than the IV method, the effect of the skewed errors is noticeable in the competing methods and the naive method. The RC estimator showed greater bias and RMSE values compared to the IV method and according to Table 4.7, for 10% censoring, moving from $n = 50$ to $n = 500$, the RMSE values range from 0.245 to 0.113. Moreover, the SIMEX method showed more sensitivity showing a coverage probability of only 0.887 under 10% censoring and $n = 500$ for the β_x parameter according to Table 4.7. The performance of the other coefficients showed a similar trend.

When the censoring level increases to 50%, with the combination of skewed errors and higher censoring level, performance deteriorates for all methods, but the IV method stays relatively stable. For example, according to Table 4.8, under $n = 500$, for the coefficient β_x the RMSE values for IV, RC and SIMEX methods are 0.100, 0.144 and 0.308 respectively. As expected the naive method

showed the lowest performance with a RMSE value of 0.464 and a coverage probability of only 0.523 for the parameter β_x under $n = 500$ (Table 4.8). The performance was similar for the other coefficients as well.

The following graph shows a summary of how the RMSE values change across the scenarios considered.

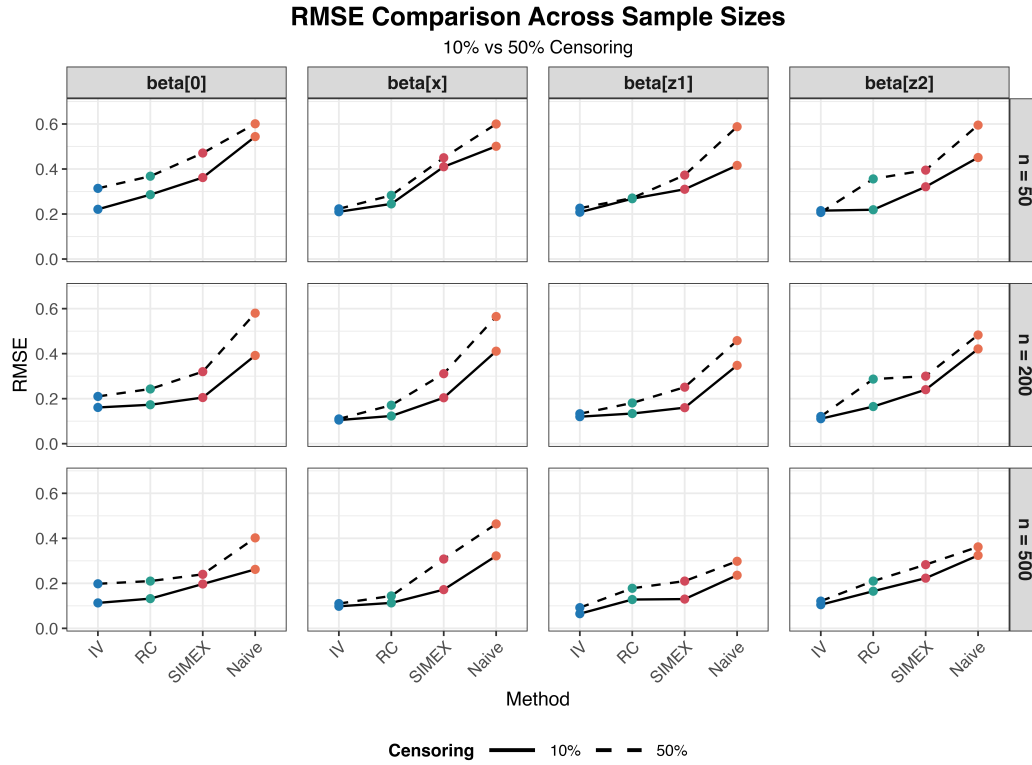


Figure 4.4: Comparison of RMSE values under the considered scenarios in Experiment 4

4.5 Experiment 5

Experiment 5 assesses the performance under a heteroskedastic response error, where the variability of the response depends on one of the covariates. This is a common situation in the real world where the outcome variability is different across population subgroups. This setup can be related to the same real world example discussed above where the outcome is systolic blood pressure, the latent covariate is blood glucose level and the fully observed covariate is age. To indicate heterogeneous populations, variance of the response error is modeled as an exponential function of the fully observed covariate. This experiment investigates the performance of the methods when the constant error variance assumption is violated. The setup for Experiment 5 is as follows:

1. Generate $S \sim N(\mu, \sigma^2)$ where, $\mu = 0, \sigma^2 = 1$ and $Z \sim N(\mu_z, \sigma_z^2)$ where, $\mu_z = 0, \sigma_z^2 = 1.5$.

2. Generate $X = \alpha_s S + \alpha_z Z + u; u \sim N(0, \sigma_u^2)$ where, $\alpha_s = 2, \alpha_z = 1, \sigma_u^2 = 1.5$.

α_z was set to 1 and $\alpha_s = 2, \sigma_u^2 = 1.5$ were chosen to produce a correlation between the instrumental variable and the latent covariate to be approximately $Corr(X, S) = 0.760$, satisfying the relevance assumption (A 2.4) explained in Chapter 2.

3. Generate $X^* = X + e$ where, $e \sim N(0, \sigma_e^2)$ where, $\sigma_e^2 = 2.5$.

$\sigma_e^2 = 2.5$ was chosen to achieve the reliability ratio of 0.7 explained in (2.6).

4. Consider random right censoring.

Generate, $c_i \sim u(c_0 - \delta, c_0 + \delta)$ and let $W = \min(X^*, c_i)$.

Censoring indicator, $\Delta_i = 1; X_i^* \leq c_i$, where, $c_0 = 4.2$ (10% censoring), $c_0 = 1.2$ (50% censoring) and $\delta = 0.5$.

5. Generate $Y = \beta'_x X + \beta'_z Z + \varepsilon; \varepsilon_i \sim N(0, \sigma_{\varepsilon i}^2)$.

The regression errors were assumed to follow a normal distribution and the variance is considered to vary across individuals. The variance function is as follows:

$$\sigma_{\varepsilon i}^2 = \text{Var}(\varepsilon_i | Z_i) = \sigma_{\varepsilon}^2 \exp(\gamma_z Z_i),$$

where, $\gamma_z = 0.5$.

An exponential variance function was chosen to keep the error variance positive and to systematically increase with the fully observed covariate. $\gamma_z = 0.5$ was selected to have a moderate degree of heteroskedasticity.

The simulation results for Experiment 5 are as follows:

Table 4.9: Comparison of methods under 10% censoring and different sample sizes for Experiment 5.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	1.011	0.171	0.171	0.940	1.010	0.151	0.151	0.946	1.010	0.143	0.143	0.946
	RC	0.970	0.233	0.235	0.939	0.975	0.213	0.214	0.943	0.975	0.185	0.187	0.945
	SIMEX	0.901	0.251	0.270	0.890	0.910	0.239	0.256	0.910	0.910	0.214	0.232	0.930
	Naive	0.882	0.474	0.488	0.710	0.880	0.324	0.345	0.811	0.880	0.313	0.335	0.880
$\beta_x = 1.2$	IV	1.181	0.121	0.122	0.944	1.187	0.112	0.113	0.946	1.180	0.080	0.082	0.947
	RC	1.178	0.179	0.180	0.934	1.180	0.164	0.165	0.944	1.174	0.123	0.126	0.945
	SIMEX	1.175	0.212	0.213	0.840	1.105	0.191	0.213	0.900	1.097	0.171	0.198	0.917
	Naive	0.993	0.478	0.521	0.452	0.904	0.355	0.462	0.551	0.901	0.286	0.410	0.720
$\beta_z = 2.0$	IV	1.950	0.141	0.150	0.940	1.955	0.137	0.144	0.946	2.010	0.120	0.120	0.948
	RC	2.101	0.149	0.180	0.930	2.100	0.128	0.162	0.942	2.060	0.113	0.128	0.944
	SIMEX	2.103	0.173	0.201	0.890	2.050	0.192	0.198	0.920	1.995	0.176	0.176	0.935
	Naive	2.202	0.435	0.480	0.650	2.103	0.428	0.440	0.721	2.201	0.347	0.401	0.850

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table 4.10: Comparison of methods under 50% censoring and different sample sizes for Experiment 5.

Parameter	Method	$n = 50$				$n = 200$				$n = 500$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
$\beta_0 = 1.0$	IV	1.005	0.225	0.225	0.935	1.005	0.174	0.174	0.941	1.008	0.167	0.167	0.944
	RC	0.950	0.305	0.311	0.925	0.960	0.238	0.241	0.935	0.970	0.199	0.201	0.940
	SIMEX	0.885	0.345	0.369	0.865	0.900	0.260	0.279	0.900	0.910	0.231	0.248	0.915
	Naive	0.865	0.520	0.545	0.650	0.880	0.361	0.380	0.780	0.882	0.361	0.380	0.850
$\beta_x = 1.2$	IV	1.190	0.170	0.172	0.940	1.198	0.125	0.125	0.944	1.190	0.118	0.119	0.945
	RC	1.150	0.235	0.241	0.930	1.157	0.184	0.189	0.939	1.160	0.173	0.177	0.940
	SIMEX	1.120	0.285	0.296	0.751	1.140	0.224	0.232	0.880	1.150	0.225	0.230	0.910
	Naive	0.880	0.530	0.619	0.420	0.904	0.369	0.473	0.520	0.900	0.350	0.461	0.651
$\beta_z = 2.0$	IV	1.940	0.205	0.212	0.935	1.950	0.157	0.165	0.941	2.001	0.135	0.135	0.944
	RC	2.040	0.225	0.233	0.928	2.050	0.175	0.182	0.935	2.050	0.144	0.152	0.940
	SIMEX	2.120	0.255	0.282	0.895	2.111	0.203	0.232	0.910	1.952	0.192	0.198	0.925
	Naive	2.240	0.550	0.600	0.580	2.221	0.422	0.476	0.680	2.211	0.386	0.440	0.804

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Tables 4.9 and 4.10 present the results when the model is further extended to a heteroskedastic response error model. Even though the non constant variance increases the variability of the parameters, the IV method tends to

show a better performance relative to the competing methods and the naive method. It produces estimates close to the true parameter values and coverage probabilities which align with the nominal level. For instance, according to Table 4.9, under 10% censoring and $n = 500$, for the IV method, β_x parameter is between 1.180 and 1.187 across the sample sizes and the RMSE values are below 0.13. Moreover, the coverage probabilities exceed 0.94.

The effect of heteroskedasticity is more noticeable in the competing methods. Although the RC method performs reasonably well, the SIMEX method shows a decrease in performance especially for the parameter β_x where, under 50% censoring for $n = 50$, the coverage probability is only 0.751, as shown in Table 4.10. As expected, the naive method showed the lowest performance. With an increase in the censoring proportion, the performance of the methods was reduced, but the IV method continued to show a reasonable performance. For example, in Table 4.10, under 50% censoring, for $n = 500$ the RMSE values for IV, RC and SIMEX methods are 0.119, 0.177 and 0.241 respectively. Overall, it can be said that the IV method is relatively robust to the violation of the heteroskedasticity assumption.

The following graph shows a summary of how the RMSE values change across the scenarios considered.

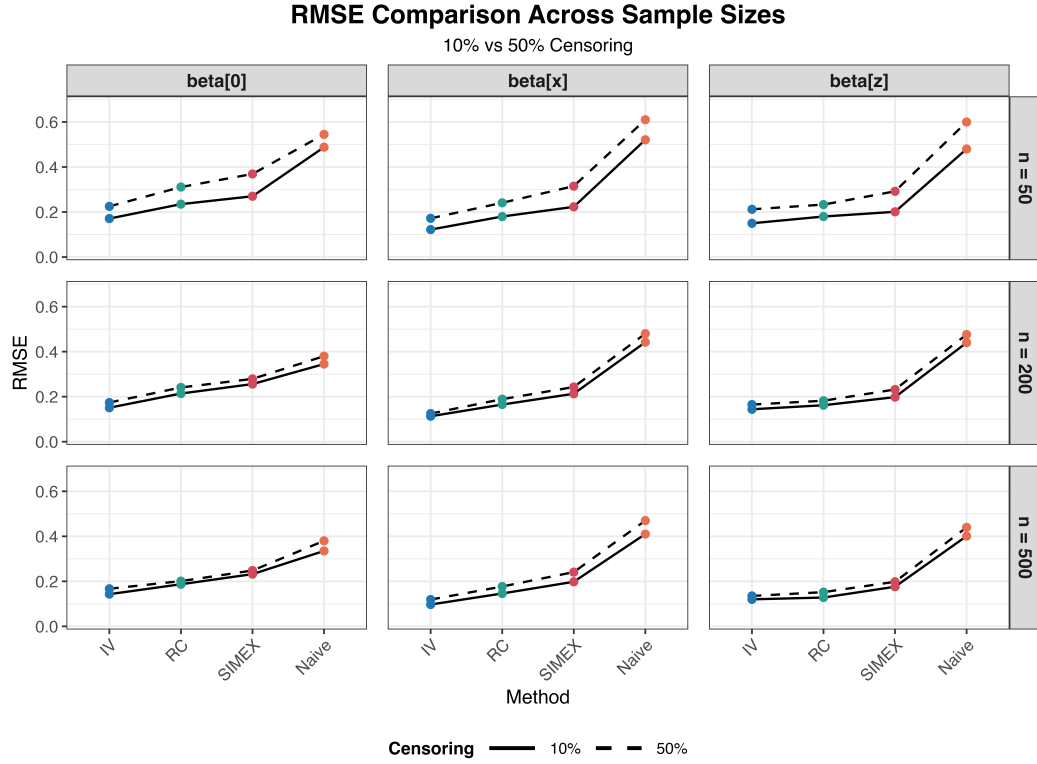


Figure 4.5: Comparison of RMSE values under the considered scenarios in Experiment 5

4.6 Summary of the Simulation Results

Five distinct simulation studies were done and presented in this chapter motivated by different real world scenarios. In order to evaluate the performance of the IV method, simulation studies were done under different combinations of models, error distributions, and censoring mechanisms.

A baseline setting with normally distributed errors with fixed censoring was considered in Experiment 1 and the Experiment 2 censored the same setup under random censoring since the censoring threshold can often vary across individuals. In both of these experiments, under fixed and random censoring, the proposed IV method maintained a stable performance with low bias, low RMSE and coverage probabilities close to the nominal levels indicating that this IV method is not sensitive to the censoring mechanism.

Experiments 3 and 4 examined the performance when the model error departs from normality. The error distributions considered were the heavy tailed t-distributed and positively skewed χ^2 distributed. Even though these settings are more challenging, still the IV method remained stable compared to the competing methods showing the method's robustness to departure from normality.

The Experiment 5 examined the effect of heteroskedastic errors where the variance changes across the observations. These error could arise due to heterogeneous populations. According to the simulation results under this model, even though a decrease of the performance can be observed, the IV method performed reasonably well and outperformed the other methods showing low RMSE values and coverage probabilities closer to the nominal level.

A consistent pattern can be observed across all the simulation experiments. Across all the estimation methods, with the increase of the sample size, smaller Monte Carlo standard deviation and RMSE values, and coverage probabilities closer to the nominal level, were observed. Moreover, with the increase of the censoring proportion, the performance decreased among all the methods. However, the IV method remained comparatively stable. The regression calibration

method performed closely to the IV method but the SIMEX method showed sensitivity to non-normal errors, smaller sample sizes and higher censoring proportions. As expected, the naive method exhibited the lowest performance with high bias, RMSE values and very low coverage probabilities showing the importance of accounting for the measurement error. Moreover, an interesting observation under the naive method is that, although the estimates should attenuate towards zero, some of the fully observed covariates were overestimated. This could be since the data is generated under not only the measurement error but censoring as well.

These simulation studies evaluated under different scenarios provide support for the proposed IV method. These findings support the proposed method and provides a reliable approach for linear regression models with a censored and mismeasured covariate.

4.7 Additional Experiments

4.7.1 Instrumental Strength

As mentioned in the previous Experiments 1-5, the value of α_s in the latent covariate model (3.7), was set to the value 2, to achieve a correlation of 0.7-0.8 between the latent covariate and the instrumental variable to satisfy the IV assumption. In this section, lower values for α_s are used to further examine the performance of the IV estimator when the instrumental strength is low. Therefore, an additional simulation study was conducted to examine the effect of the strength of the instrumental variable. Since the performance of the instrumental variable method depends on the strength of the instrument, to

examine how the IV estimator behaves under different values of α_s the following study was conducted.

For this analysis, the data generation process of Experiment 3 under 10% censoring was used, while varying the value of α_s . The values considered for α_s were 0.2, 1.0, 1.5, and 2.0, and the corresponding correlations are approximately 0.105, 0.465, 0.511, and 0.750, respectively. The remaining parameters were kept as they are and the simulations were conducted for sample sizes 50, 200, and 500. The following table presents the results.

Table 4.11: Performance of the IV estimator under varying instrument strengths for Experiment 3 under 10% censoring.

Parameter	Measure	$\alpha_s = 0.2$	$\alpha_s = 1.0$	$\alpha_s = 1.5$	$\alpha_s = 2.0$
$n = 50$					
$\beta_0 = 1.0$	Est	1.318	1.102	1.055	1.020
	MCSD	1.082	0.561	0.372	0.224
	RMSE	1.129	0.571	0.376	0.225
	CP	0.854	0.891	0.910	0.925
$\beta_x = 1.2$	Est	0.612	1.061	1.151	1.195
	MCSD	1.824	0.731	0.352	0.151
	RMSE	1.916	0.774	0.356	0.151
	CP	0.781	0.862	0.902	0.924
$\beta_{z1} = 2.0$	Est	2.521	2.118	2.047	1.990
	MCSD	1.652	0.712	0.434	0.211
	RMSE	1.732	0.722	0.437	0.211
	CP	0.842	0.903	0.926	0.947
$\beta_{z2} = 3.2$	Est	3.487	3.284	3.241	3.218
	MCSD	1.243	0.804	0.893	0.277
	RMSE	1.276	0.809	0.895	0.278
	CP	0.903	0.921	0.936	0.948
$n = 200$					
$\beta_0 = 1.0$	Est	1.263	1.041	1.068	1.015
	MCSD	0.873	0.824	0.893	0.114
	RMSE	0.912	0.825	0.896	0.115
	CP	0.909	0.910	0.920	0.943
$\beta_x = 1.2$	Est	0.741	1.180	1.177	1.198
	MCSD	1.430	0.293	0.241	0.121
	RMSE	1.502	0.294	0.242	0.121
	CP	0.840	0.905	0.915	0.942
$\beta_{z1} = 2.0$	Est	2.455	2.012	2.020	1.996
	MCSD	1.645	0.460	0.402	0.120
	RMSE	1.707	0.460	0.404	0.120
	CP	0.898	0.937	0.944	0.955
$\beta_{z2} = 3.2$	Est	3.426	3.208	3.210	3.185
	MCSD	1.146	0.789	0.826	0.152
	RMSE	1.168	0.789	0.826	0.153
	CP	0.955	0.952	0.946	0.951
$n = 500$					
$\beta_0 = 1.0$	Est	1.004	1.007	1.060	1.010
	MCSD	0.532	0.384	0.338	0.074
	RMSE	0.532	0.384	0.343	0.075
	CP	0.924	0.936	0.946	0.950
$\beta_x = 1.2$	Est	1.188	1.193	1.197	1.200
	MCSD	0.841	0.117	0.102	0.076
	RMSE	0.841	0.117	0.102	0.076
	CP	0.893	0.939	0.949	0.946
$\beta_{z1} = 2.0$	Est	2.008	1.999	1.997	2.000
	MCSD	0.867	0.189	0.187	0.076
	RMSE	0.867	0.189	0.187	0.076
	CP	0.903	0.958	0.949	0.956
$\beta_{z2} = 3.2$	Est	3.210	3.209	3.209	3.190
	MCSD	0.563	0.348	0.331	0.096
	RMSE	0.563	0.348	0.331	0.096
	CP	0.950	0.952	0.949	0.953

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

According to the results presented in Table 4.11, when the instrument is weak ($\alpha_s = 0.2$) for the coefficient $\beta_x = 1.2$ under $n = 50$ the RMSE is 1.923 and the coverage probability is only 0.781. However, in the same case, when the instrument is strong ($\alpha_s = 2$), the RMSE value decreases to 0.151 and the coverage probability increases to 0.924. This indicates that the performance of the IV estimator is sensitive to the strength of the instrument specifically for small sample sizes.

When the sample size increases, the IV estimator performs more stably across all levels of the instrument strength. Even when the instrument is relatively weak ($\alpha_s = 0.2$), the bias of the estimates is small, although the RMSE is larger compared to the stronger instrument settings. For example, in Table 4.11, under $n = 500$, the RMSE values for the coefficient β_X are 0.841 and 0.076 when $\alpha_s = 0.2$ and $\alpha_s = 2$, respectively. Moreover, as the instrument strength increases, the coverage probabilities remain reasonably close to the nominal level of 0.95 across most cases. This pattern can be observed in the other coefficients as well.

For all sample sizes considered, when $\alpha_s = 0.2$ and $\alpha_s = 1.0$, the RMSE values are relatively high and the coverage probabilities are lower compared to the results obtained for the SIMEX and RC methods in Experiment 3. Initially, Experiment 3 was conducted in a strong instrument setting. This observation verifies the importance of choosing a suitable instrumental variable that satisfies the relevance assumption (A 2.4).

Chapter 5

Real Data Applications

5.1 Mercury Level using Fish Consumption

Background

A common research area in public health is assessing nutritional data. In the context of this particular application, fish consumption is a major source of the mercury level in blood, which leads to adverse health effects, especially among women. Referring to the paper [Kipnis et al. \(2009\)](#), we evaluate the relationship between the level of mercury in blood of women of child bearing age (12-49 years) and fish consumption.

In this study, the true long term fish consumption is unobserved. Therefore, fish intake is measured by a 24 hour dietary recall but this may not be an accurate measure for the true long term fish consumption . Due to this reason and self reporting errors 24 hour fish consumption is subject to measurement error. On the other hand, fish consumption includes a large number of zeros since many individuals might not consume fish on the specific days. This leads

to a substantial proportion of zeros in the fish consumption variable. A zero observation in the fish consumption represents a censored value under the assumption that the recorded zero value corresponds to a value at or below the censoring threshold. In the analysis, we consider this interpretation and treat the lower bound observations of the 24 hour fish consumption as left censored. Since the 24 hour fish consumption is measured with error and censored at the same time, it makes this application ideal to this study.

The dataset used is publicly available and it was taken from NHANES (National Health and Nutrition Examination Survey) 2003-2004 which was constructed to assess the nutritional status of the population of USA.

The following data files were used to construct the dataset:

- Demographics file (DEMO).
- Laboratory file containing total blood mercury (LBXTHG).
- Dietary Individual Foods File – Day 1 (DR1IFF).
- Dietary Individual Foods File – Day 2 (DR2IFF).

Finally, all the files were merged using the unique participant id (SEQN) and the females between the age 12-49 years were extracted from the dataset. After cleaning and preprocessing, the final dataset consisted of 1921 observations.

Table 5.1: Description of variables used in the NHANES dietary study.

Variable	Description	Role
mercury	Blood mercury concentration level	Response
fish_oz_day1	Fish consumption from Day 1 recall	Censored and mismeasured covariate
fish_oz_day2	Fish consumption from Day 2 recall	Instrumental variable
age	Age of participant (years)	Fully observed Covariate
income	Household income level	Fully observed Covariate
race3	Race category (White, African American, Other)	Fully observed Covariate
educ	Education level (Some college or higher, No college)	Fully observed Covariate

The outcome is the total level of mercury in blood and due to the skewness of the distribution a log transformation was applied.

Fish intake was constructed manually using the individual food files using the USDA food codes for fish items consumed by each individual. Then the total fish consumption for the corresponding days (day 1 and day 2) were constructed separately.

The variable `fish_oz_day1`, which is the fish intake from day 1, was treated as the covariate which is mismeasured and censored at the same time while, `fish_oz_day2`, which is the fish intake from day 2, was considered as the instrumental variable for the fish intake from day 1 and it shows the below correlations with the day 1 fish intake and mercury level.

- High correlated with day 1 fish intake (correlation = 0.671).
- Low correlation with the mercury level (correlation = 0.134).

Model

The linear regression model which indicates the relationship between the blood mercury level and fish intake can be written as, $Y = \beta_x X + \beta'_z Z + \epsilon$, where $Y = \log(\text{Mercury Level})$, $X = \text{True fish intake}$, and $Z = \text{Fully observed covariates}$.

Since the observed fish intake is subject to measurement error, it can be written as, $X^* = X + e$, where e is the measurement error. Moreover, $W = \max(X^*, 0)$ indicate the left censoring at 0. The association between the instrument and the fish intake can be written as, $X = \alpha_s S + \alpha_z Z + u$, where, S is Fish intake from day 2 and u is the random error.

The first stage regression coefficients were estimated using the *aftsrr()* function from the package *aftgee*, which fits a semiparametric AFT model and the parameter estimation was done using the Gehan's rank based weighting scheme. In this analysis, observed fish intake was transformed using an exponential function. The reduced form regression model, (3.9) was fitted and the estimated covariance matrix of the reduced form coefficients, $\hat{\Sigma}_\gamma$ was obtained using the *vcov()* function. Using the asymptotic covariance matrix of the IV estimator, $Var(\hat{\beta}) = (\hat{A}' \hat{\Sigma}_\gamma^{-1} \hat{A})^{-1}$, Monte Carlo standard deviations were obtained after taking the square roots of the diagonal elements. These Monte Carlo standard deviation were used to construct confidence intervals.

Results

The IV estimation method, regression calibration method, SIMEX method and naive method was applied to this dataset and the results for these methods are as follows:

Table 5.2: Comparison of IV, RC, and SIMEX estimation methods for NHANES dietary study

Method	Covariate	Estimate	SE	CI lower	CI upper
IV	fish_oz_day1	0.086	0.011	0.064	0.108
	age	0.011	0.005	0.001	0.021
	race3Other	-0.223	0.087	-0.394	-0.052
	race3White	-0.366	0.077	-0.517	-0.215
	income	0.050	0.011	0.028	0.072
	educ	0.124	0.016	0.093	0.155
RC	fish_oz_day1	0.077	0.017	0.044	0.110
	age	0.017	0.007	0.003	0.031
	race3Other	-0.187	0.090	-0.363	-0.011
	race3White	-0.297	0.109	-0.511	-0.083
	income	0.036	0.014	0.009	0.063
	educ	0.086	0.040	0.008	0.164
SIMEX	fish_oz_day1	0.122	0.022	0.079	0.165
	age	0.015	0.007	0.001	0.029
	race3Other	-0.231	0.112	-0.451	-0.011
	race3White	-0.267	0.115	-0.492	-0.042
	income	0.095	0.019	0.058	0.132
	educ	0.138	0.067	0.007	0.270
Naive	fish_oz_day1	0.037	0.025	-0.012	0.086
	age	0.029	0.021	-0.012	0.070
	race3Other	-0.093	0.061	-0.213	0.027
	race3White	-0.201	0.112	-0.421	0.019
	income	0.011	0.026	-0.040	0.062
	educ	0.081	0.038	0.007	0.155

Table 5.2 represents the estimates of the proposed IV, RC, SIMEX and

naive methods for the NHANES dietary study. All of the methods show a positive association between the day 1 fish consumption and the mercury level in blood.

For the fish intake variable, all correction methods produce larger estimates compared to the naive method ($\hat{\beta} = 0.037$) which indicates that the naive method underestimates the strength of the relationship between the blood mercury level and fish consumption. Moreover, the naive method gives a 95% confidence interval $(-0.012, 0.086)$, which includes zero within the interval. This indicates that the naive method gives insignificant estimates, while the correction methods produce significant estimates at 5% significance level. This indicates that ignoring measurement error attenuates the association between fish intake and blood mercury concentration.

When it comes to the effects of other fully observed covariates, age, income and education are positive under all the methods, which indicates that older individuals with higher income consume more fish leading to higher mercury concentration in blood. Although the magnitudes of the coefficients vary across the methods, the findings are consistent.

Furthermore, it can be observed that the attenuation effect of the measurement error mainly affects the fish intake variable, which is censored and measured with error while the fully observed covariates show changes in magnitude and direction.

5.2 Framingham Heart Study

Background

The framingham heart study is one of the most influential datasets in epidemiological studies. The application studied in this thesis is based on investigating the relationship between blood pressure and blood glucose level. High risk factors that elevate systolic blood pressure (SBP) are blood glucose level, obesity, and smoking.

Numerous studies have been done to understand the relationship between SBP and blood glucose level. Blood glucose levels are obtained using laboratory measures. Therefore, in practice, blood glucose levels are subject to measurement error. This measurement error in glucose measurements can arise due to various reasons such as laboratory variability and instrument precision. As a result, observed glucose levels can differ from the true underlying glucose levels.

In addition, laboratory measures such as glucose measurements are subject to detection limits. Measurements below a specific threshold are not presented as exact values, but reported as below the limit. Since originally the Framingham dataset did not contain censored glucose measurements, an artificial censoring mechanism was imposed. Apart from that, it satisfies the appropriate conditions to apply the proposed methodology since repeated glucose measurements are available.

Dataset

The dataset used is a subset of the Framingham Heart Study, taken from the R package *riskCommunicator* and it contains repeated measurements on individuals across several examination periods. After cleaning and preprocessing, the final dataset consisted of 3026 observations.

The following table presents the variables included in the dataset:

Table 5.3: Description of variables used in the Framingham Heart Study analysis.

Variable	Description	Role
SYSBP	Systolic blood pressure (mmHg)	Response
GLUCOSE	Blood glucose level measured at the first examination period	Censored and mismeasured covariate
GLUCOSE2	Blood glucose level measured at the second examination period	Instrumental variable
AGE	Age (years)	Fully observed covariate
SEX	Sex (Male/Female)	Fully observed covariate
BMI	Body mass index	Fully observed covariate
CURSMOKE	Current smoking status (0,1)	Fully observed covariate

In this analysis, observations corresponding to the first examination period were extracted into a separate dataset and the GLUCOSE variable from the second examination period was then extracted and merged with the baseline dataset according to the unique participant identification number.

Glucose measurement in the first examination is subject to measurement error and glucose measurement in the second examination is considered as the instrumental variable. It shows the below correlations with glucose measurement of examination period 1 and systolic blood pressure .

- High correlation with the glucose measurements of examination period 1 (correlation = 0.689).
- Low correlation with systolic blood pressure (correlation = 0.177).

Moreover, the idea behind taking a repeated measurement as an instrumental variable was that repeated measurements on the same individual are expected to show a high correlation since these measurements tend to remain stable throughout.

Since the original Framingham dataset does not contain censored glucose measurements, to account for censoring, artificial left censoring was imposed on the glucose variable to mimic a laboratory detection limit. A fixed censoring threshold corresponding to the 20th percentile of the glucose distribution was selected, and glucose values below this threshold were treated as left censored. The artificial censoring mechanism was introduced for the purpose of methodological evaluation, and the choice of the 20th percentile gives a moderate degree of censoring which is large enough to demonstrate the problem.

Model

The linear regression model which indicates the relationship between the systolic blood pressure (SYSBP) and glucose level can be written as, $Y = \beta_x X + \beta'_z Z + \epsilon$, where, $Y = \text{SYSBP}$, $X = \text{True glucose level}$, and $Z = \text{Fully observed covariates}$.

According to the model described in section 3.2, equation (3.7), $X^* = \text{Glucose level from examination period 1}$, and $S = \text{Glucose level from examination period 2}$.

The first stage regression coefficients were estimated using the *aftsrr()* function from the package *aftgee*, which fits a semiparametric AFT model and the parameter estimation was done using the Gehan's rank based weighting scheme. In this analysis, observed glucose measurements were transformed using an exponential function. The reduced form regression model, (3.9) was fitted and the estimated covariance matrix of the reduced form coefficients, $\hat{\Sigma}_\gamma$ was obtained using the *vcov()* function. Using the asymptotic covariance matrix of the IV estimator, $Var(\hat{\beta}) = (\hat{A}'\hat{\Sigma}_\gamma^{-1}\hat{A})^{-1}$, standard errors were obtained after taking the square roots of the diagonal elements. These standard errors were used to construct confidence intervals.

Results

The following table presents the results when glucose measurements are considered to be censored and mismeasured.

Table 5.4: Comparison of IV, RC, and SIMEX estimation methods for FHS data

Method	Covariate	Estimate	SE	CI lower	CI upper
IV	GLUCOSE	0.625	0.115	0.400	0.851
	AGE	0.774	0.042	0.692	0.855
	SEX	3.030	0.692	1.674	4.385
	BMI	1.324	0.087	1.153	1.494
	CURSMOKE	0.508	0.206	0.104	0.912
RC	GLUCOSE	0.126	0.144	0.156	0.408
	AGE	0.822	0.042	0.743	0.901
	SEX	2.741	0.691	1.387	4.095
	BMI	1.407	0.086	1.239	1.576
	CURSMOKE	0.509	0.229	0.061	0.957
SIMEX	GLUCOSE	0.183	0.165	0.140	0.506
	AGE	0.783	0.041	0.703	0.864
	SEX	2.808	0.612	1.608	4.008
	BMI	1.374	0.086	1.206	1.542
	CURSMOKE	0.251	0.115	0.025	0.477
Naive	GLUCOSE	0.054	0.085	-0.113	0.221
	AGE	0.818	0.030	0.759	0.877
	SEX	2.705	0.069	2.569	2.840
	BMI	1.405	0.086	1.236	1.574
	CURSMOKE	0.506	0.208	0.097	0.914

Table 5.4 represents the estimates from the proposed IV method, RC, SIMEX and naive method for the Framingham heart study. All the methods show a positive association between the glucose level of period 1 and the systolic blood pressure.

The estimated glucose effects for IV, SIMEX and RC methods are 0.6253,

0.1829 and 0.1255, respectively. It can be observed that the IV method produced the largest estimate while the estimate obtained from the naive method was only 0.0536, which is considerably a smaller estimate. This indicates attenuation bias when the measurement error is ignored. The IV estimator shows a stronger association between systolic blood pressure and blood glucose level, indicating that the true effect may be underestimated when the measurement error is ignored.

When it comes to the effects of other fully observed covariates, age, sex, BMI and smoking status are positive under all the methods indicating a positive association between systolic blood pressure and blood glucose level. Although some variation is observed, the estimates are relatively stable across the methods. This indicates that the measurement error correction has primarily affected the error prone glucose measurement.

Moreover, it can be observed that the standard errors related to the naive method are generally smaller compared to the corrected methods. For example, when it comes to the censored and error prone glucose measurement, the standard errors for the IV and naive methods are 0.1150 and 0.0846, respectively. This is due to the bias variance tradeoff which commonly arises in measurement error problems.

Both real data applications discussed in this chapter produce results consistent with the findings of the simulation studies discussed in Chapter 4. Both applications reflected the attenuation bias by producing larger estimates for the error prone covariate under corrected methods compared to the naive method. The proposed methodology performed as expected and indicated the practical usefulness of this methodology. Moreover, these applications highlight the importance of accounting for measurement error.

Chapter 6

Discussion and Conclusion

This study addressed linear regression models with censored and mismeasured covariates where a covariate is subject to censoring and measurement error simultaneously. This study was motivated particularly by the applications in dietary and biomedical studies. Given that limited attention has been given to this problem, this study developed an instrumental variable estimation procedure which accounts for both censoring and measurement error.

The theoretical development in this study extends the classical instrumental variable method by accounting for censoring along with a two step estimation procedure. This method combines the first stage estimation with a minimum distance estimator and recovers the structural regression parameters. Moreover, regression calibration and SIMEX methods were developed accordingly to make a comparison with the proposed methodology.

Through the simulation studies, the performance of the proposed estimator was evaluated under different sample sizes, censoring levels and model settings along with the other existing methods. Throughout the simulation studies, the proposed method outperformed the other existing methods by producing stable

results in different settings. This demonstrates the applicability of this method among realistic data settings. In this study, the importance of accounting for the measurement error was highlighted through the simulation results of the naive method.

As mentioned earlier, regression calibration and SIMEX methods were used as comparison methods. Among those methods, the regression calibration method was computationally simple, while the SIMEX method was implemented through repeated simulations. The computational time was the longest for the SIMEX method compared to the other methods.

The practical usefulness of this study was illustrated using the NHANES dietary study and Framingham Herat Study, and the proposed methodology showed successful results, showing the importance of this method which accounts for censoring and measurement error. Apart from these applications, this proposed methodology can be used in broader applications including laboratory based studies and studies on physical activity based on energy expenditure and step counts ([Qin et al., 2026](#)).

Overall, the development of methodology, numerical studies and practical applications contribute to the research on censored and mismeasured covariates and further developments in this area.

Even though, the proposed methodology performed well across the settings considered, there are limitations that should be addressed. The instrumental variable method is heavily dependent on the availability of a suitable and valid instrumental variable. Particularly, the instrumental variable should satisfy the relevance and exogeneity assumptions. Identifying a suitable instrument satisfying these assumptions in practical applications is challenging. Although

the relevance condition can be checked, the exogeneity condition is assumed based on the application.

This study can be extended to future research by incorporating multiple covariates which are censored and measured with error. Moreover, a doubly censored case can be adapted along with the measurement error, where the covariate is mismeasured and subject to both lower and upper detection limits. Extending this study to doubly censored case will broaden its applicability to real world problems.

Bibliography

- Ashner, M. C. and T. P. Garcia (2024). Understanding the implications of a complete case analysis for regression models with a right-censored covariate. *The American Statistician* 78(3), 335–344. (Cited on page 5.)
- Atem, F. D., J. Qian, J. E. Maye, K. A. Johnson, and R. A. Betensky (2017). Linear regression with a randomly censored covariate: application to an alzheimer’s study. *Journal of the Royal Statistical Society Series C: Applied Statistics* 66(2), 313–328. (Cited on pages 1 and 5.)
- Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications*. Chapman and Hall/CRC. (Cited on page 3.)
- Buzas, J. S. and L. A. Stefanski (1996). Instrumental variable estimation in generalized linear measurement error models. *Journal of the American Statistical Association* 91(435), 999–1006. (Cited on page 4.)
- Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69(5), 1127–1160. (Cited on page 4.)
- Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006).

- Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC. (Cited on pages 3, 4, 12, 13 and 16.)
- Chen, G., F. A. McAlister, R. L. Walker, B. R. Hemmelgarn, and N. R. Campbell (2011). Cardiovascular outcomes in Framingham participants with diabetes: the importance of blood pressure. *Hypertension* 57(5), 891–897. (Cited on page 2.)
- Cook, J. R. and L. A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association* 89(428), 1314–1328. (Cited on page 3.)
- Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons. (Cited on pages 3, 4 and 11.)
- Jin, Z., D. Lin, L. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. *Biometrika* 90(2), 341–353. (Cited on page 32.)
- Kipnis, V., D. Midthune, D. W. Buckman, K. W. Dodd, P. M. Guenther, S. M. Krebs-Smith, A. F. Subar, J. A. Tooze, R. J. Carroll, and L. S. Freedman (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* 65(4), 1003–1010. (Cited on pages 2, 3, 6 and 71.)
- Kwon, J. (2026). *Semiparametric Methods in Regression with Covariates Subject to Detection Limits*. Ph. D. thesis, University of California, Irvine. (Cited on page 5.)

- Lotspeich, S. C., M. C. Ashner, J. E. Vazquez, B. D. Richardson, K. F. Grosser, B. E. Bodek, and T. P. Garcia (2024). Making sense of censored covariates: Statistical methods for studies of huntington’s disease. *Annual Review of Statistics and its Application* 11(1), 255–277. (Cited on pages 5 and 29.)
- Lu, B., S. Thomson, S. Blommaert, M. Tadrous, C. C. Earle, and K. K. Chan (2023). Use of instrumental variable analyses for evaluating comparative effectiveness in empirical applications of oncology: a systematic review. *Journal of Clinical Oncology* 41(13), 2362–2371. (Cited on page 4.)
- Qin, C., L. Xue, U. Beyaztas, R. S. Zoh, M. Benden, J. Goldsmith, and C. D. Tekwe (2026). Correcting measurement error and zero inflation in functional covariates for scalar-on-function quantile regression. *arXiv preprint arXiv:2602.05784*. (Cited on pages 5, 6 and 84.)
- Rigobon, R. and T. M. Stoker (2007). Estimation with censored regressors: Basic issues. *International Economic Review* 48(4), 1441–1467. (Cited on page 4.)
- Song, X. and C.-Y. Wang (2025). A corrected score approach for proportional hazards model with error-contaminated covariates subject to detection limits. *Statistics in Medicine* 44(23-24), e70243. (Cited on page 6.)
- Spiegelman, D. (2014). Regression calibration in air pollution epidemiology with exposure estimated by spatio-temporal modeling. *Environmetrics* 24(8). (Cited on page 3.)
- Spiegelman, D., A. McDermott, and B. Rosner (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology.

- The American journal of Clinical Nutrition* 65(4), 1179S–1186S. (Cited on page 3.)
- Tsimihodimos, V., C. Gonzalez-Villalpando, J. B. Meigs, and E. Ferrannini (2018). Hypertension and diabetes mellitus: Coprediction and time trajectories. *Hypertension* 71(3), 422–428. (Cited on page 2.)
- Wang, C.-Y., J. d. D. Tapsoba, C. Duggan, and A. McTiernan (2024). Generalized linear models with covariate measurement error and zero-inflated surrogates. *Mathematics* 12(2), 309. (Cited on page 1.)
- Wang, L. and C. Hsiao (2007). Two-stage estimation of limited dependent variable models with errors-in-variables. *The Econometrics Journal* 10(2), 426–438. (Cited on pages 5 and 32.)
- Yi, Y., T. Ye, M. Yu, and J. Shao (2020). Cox regression with survival-time-dependent missing covariate values. *Biometrics* 76(2), 460–471. (Cited on page 32.)

Appendix A

Additional Simulation Studies

Additional experiments were conducted with a higher reliability ratio for the measurement error using two parameter settings. The error variance of the measurement error was chosen to achieve the reliability ratio of 0.85 for all the Experiments. The error variances σ_e^2 chosen here for Experiments 1, 3, and 4 (described in chapter 4) were 1.32, 2.68, and 2.26, respectively.

Table A.1: Comparison of IV, RC, SIMEX, and naive methods for $n = 50$, and 5% censoring under two parameter settings for Experiment 1.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.4, 0.5)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.094	0.112	0.112	0.948	0.988	0.241	0.241	0.946
	RC	0.091	0.150	0.150	0.945	0.932	0.163	0.177	0.931
	SIMEX	0.118	0.301	0.302	0.901	1.132	0.400	0.421	0.884
	Naive	0.147	0.507	0.509	0.676	0.911	0.511	0.519	0.621
β_x	IV	0.400	0.002	0.002	0.951	1.201	0.007	0.007	0.949
	RC	0.401	0.183	0.183	0.938	1.207	0.217	0.217	0.942
	SIMEX	0.380	0.080	0.082	0.921	1.133	0.227	0.237	0.908
	Naive	0.190	0.072	0.222	0.643	0.763	0.115	0.452	0.576
β_z	IV	0.493	0.037	0.038	0.952	1.989	0.163	0.163	0.947
	RC	0.487	0.152	0.153	0.943	1.987	0.182	0.183	0.931
	SIMEX	0.502	0.157	0.157	0.934	2.003	0.195	0.195	0.926
	Naive	0.690	0.166	0.252	0.782	3.058	0.688	1.262	0.524

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.2: Comparison of methods for $n = 500$ with 5% censoring under two parameter settings for Experiment 1.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.4, 0.5)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.099	0.034	0.034	0.953	0.998	0.046	0.046	0.953
	RC	0.096	0.041	0.041	0.950	0.995	0.056	0.056	0.949
	SIMEX	0.091	0.066	0.067	0.936	0.973	0.077	0.082	0.927
	Naive	0.132	0.092	0.097	0.914	0.961	0.122	0.128	0.885
β_x	IV	0.400	0.010	0.010	0.953	1.200	0.012	0.012	0.952
	RC	0.398	0.052	0.052	0.949	1.205	0.064	0.064	0.947
	SIMEX	0.387	0.071	0.072	0.935	1.164	0.080	0.088	0.928
	Naive	0.271	0.148	0.196	0.760	0.903	0.042	0.300	0.744
β_z	IV	0.499	0.016	0.016	0.954	1.999	0.018	0.018	0.954
	RC	0.495	0.042	0.042	0.951	1.996	0.050	0.050	0.950
	SIMEX	0.501	0.061	0.061	0.943	2.004	0.070	0.070	0.941
	Naive	0.673	0.042	0.178	0.806	2.118	0.284	0.308	0.688

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.3: Comparison of methods for $n = 500$ with 10% censoring under two parameter settings for Experiment 1.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.4, 0.5)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.097	0.048	0.048	0.951	0.995	0.060	0.060	0.950
	RC	0.094	0.056	0.056	0.947	0.991	0.067	0.068	0.945
	SIMEX	0.082	0.086	0.088	0.926	0.946	0.096	0.110	0.913
	Naive	0.121	0.122	0.124	0.889	0.944	0.152	0.162	0.841
β_x	IV	0.399	0.016	0.016	0.951	1.199	0.018	0.018	0.950
	RC	0.396	0.063	0.063	0.946	1.210	0.075	0.076	0.943
	SIMEX	0.373	0.086	0.090	0.922	1.145	0.102	0.116	0.910
	Naive	0.246	0.159	0.221	0.706	0.871	0.000	0.341	0.671
β_z	IV	0.498	0.021	0.021	0.952	1.998	0.025	0.025	0.951
	RC	0.492	0.051	0.052	0.948	1.994	0.061	0.061	0.947
	SIMEX	0.500	0.081	0.081	0.938	2.009	0.091	0.091	0.934
	Naive	0.646	0.142	0.204	0.758	2.164	0.316	0.356	0.612

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.4: Comparison of methods for $n = 500$ with 50% censoring under two parameter settings for Experiment 1.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.4, 0.5)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.094	0.067	0.067	0.948	0.990	0.078	0.079	0.947
	RC	0.090	0.074	0.075	0.942	0.984	0.088	0.089	0.938
	SIMEX	0.068	0.114	0.118	0.900	0.902	0.116	0.152	0.885
	Naive	0.154	0.160	0.169	0.832	0.912	0.210	0.228	0.771
β_x	IV	0.397	0.024	0.024	0.949	1.197	0.028	0.028	0.948
	RC	0.392	0.078	0.078	0.940	1.219	0.089	0.091	0.936
	SIMEX	0.349	0.106	0.118	0.896	1.112	0.120	0.149	0.883
	Naive	0.214	0.183	0.261	0.624	0.823	0.104	0.391	0.578
β_z	IV	0.496	0.030	0.030	0.949	1.996	0.036	0.036	0.948
	RC	0.488	0.065	0.066	0.943	1.990	0.077	0.078	0.940
	SIMEX	0.497	0.102	0.102	0.928	2.018	0.113	0.114	0.919
	Naive	0.618	0.216	0.246	0.692	2.242	0.358	0.432	0.486

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Taking into account the results presented in Tables A1-A4, under both parameter settings, the proposed IV method continued to show mean estimates closer to the true parameter, lower RMSE values, and coverage probabilities closer to the nominal level which are consistent with the results in chapter 4. When comparing the results in chapter 4 in Tables 4.1, and 4.2 which show the results of the case where $n = 500$ under 10%, and 50% censoring respectively, it can be observed that the RMSE values are comparatively high, and coverage probabilities are fairly low compared to the results in Tables A3 and A4. For example, according to Table 4.1, for the IV method, when $n = 500$, under 10% censoring, for the coefficient β_x the RMSE values is 0.035 and the coverage probability is 0.948. In comparison, for the same case and the same parameter setting Table A3 shows a RMSE value of 0.018 and the coverage probability of 0.950.

Table A.5: Comparison of IV, RC, SIMEX, and naive methods for $n = 50$ and 10% censoring under two parameter settings for Experiment 3.

		$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (0.1, 0.4, 0.5, 0.2)$				$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (1.0, 1.2, 2.0, 3.2)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.102	0.213	0.213	0.926	1.012	0.220	0.220	0.938
	RC	0.095	0.224	0.224	0.931	0.991	0.249	0.249	0.947
	SIMEX	0.082	0.261	0.262	0.903	0.931	0.282	0.291	0.896
	Naive	0.060	0.309	0.312	0.861	0.851	0.327	0.360	0.742
β_x	IV	0.392	0.135	0.135	0.924	1.205	0.221	0.221	0.941
	RC	0.409	0.142	0.142	0.925	1.208	0.240	0.240	0.938
	SIMEX	0.300	0.137	0.170	0.882	0.810	0.109	0.405	0.801
	Naive	0.180	0.119	0.250	0.692	0.820	0.242	0.450	0.412
β_{z1}	IV	0.509	0.261	0.261	0.954	2.011	0.260	0.260	0.950
	RC	0.497	0.170	0.170	0.941	1.998	0.238	0.238	0.936
	SIMEX	0.505	0.190	0.190	0.926	1.880	0.307	0.330	0.889
	Naive	0.510	0.210	0.210	0.901	1.950	0.357	0.360	0.861
β_{z2}	IV	0.186	0.260	0.260	0.955	3.191	0.420	0.420	0.943
	RC	0.208	0.332	0.332	0.934	3.212	0.486	0.486	0.936
	SIMEX	0.190	0.310	0.310	0.912	3.380	0.477	0.510	0.874
	Naive	0.175	0.339	0.340	0.884	3.180	0.560	0.560	0.736

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.6: Comparison of IV, RC, SIMEX, and naive methods for $n = 100$ and 10% censoring under two parameter settings for Experiment 3.

		$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (0.1, 0.4, 0.5, 0.2)$				$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (1.0, 1.2, 2.0, 3.2)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.101	0.145	0.145	0.942	1.008	0.166	0.166	0.946
	RC	0.103	0.160	0.160	0.937	1.005	0.179	0.179	0.948
	SIMEX	0.091	0.176	0.176	0.919	0.963	0.207	0.211	0.924
	Naive	0.070	0.194	0.196	0.876	0.881	0.222	0.252	0.773
β_x	IV	0.396	0.082	0.082	0.931	1.203	0.130	0.130	0.944
	RC	0.397	0.133	0.133	0.931	1.198	0.186	0.186	0.938
	SIMEX	0.310	0.132	0.160	0.901	0.950	0.102	0.270	0.846
	Naive	0.180	0.096	0.240	0.694	0.920	0.242	0.370	0.461
β_{z1}	IV	0.502	0.175	0.175	0.954	2.005	0.230	0.230	0.951
	RC	0.495	0.145	0.145	0.943	1.996	0.161	0.161	0.938
	SIMEX	0.505	0.135	0.135	0.938	1.960	0.267	0.270	0.913
	Naive	0.498	0.150	0.150	0.909	2.000	0.290	0.290	0.878
β_{z2}	IV	0.196	0.210	0.210	0.956	3.195	0.340	0.340	0.948
	RC	0.188	0.297	0.297	0.949	3.186	0.390	0.390	0.947
	SIMEX	0.195	0.245	0.245	0.927	3.350	0.435	0.460	0.901
	Naive	0.225	0.269	0.270	0.892	3.210	0.500	0.500	0.773

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.7: Comparison of IV, RC, SIMEX, and naive methods for $n = 200$ and 10% censoring under two parameter settings for Experiment 3.

		$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (0.1, 0.4, 0.5, 0.2)$				$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (1.0, 1.2, 2.0, 3.2)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.101	0.100	0.100	0.950	1.003	0.116	0.116	0.951
	RC	0.100	0.110	0.110	0.950	1.001	0.123	0.123	0.950
	SIMEX	0.094	0.125	0.125	0.936	0.975	0.149	0.151	0.939
	Naive	0.080	0.152	0.153	0.888	0.921	0.164	0.182	0.823
β_x	IV	0.398	0.057	0.057	0.946	1.201	0.090	0.090	0.947
	RC	0.404	0.117	0.117	0.944	1.204	0.130	0.130	0.943
	SIMEX	0.320	0.089	0.120	0.917	1.030	0.059	0.180	0.884
	Naive	0.170	0.047	0.235	0.697	0.991	0.300	0.366	0.475
β_{z1}	IV	0.497	0.120	0.120	0.955	2.002	0.160	0.160	0.951
	RC	0.495	0.104	0.104	0.945	1.994	0.116	0.116	0.945
	SIMEX	0.500	0.095	0.095	0.944	1.980	0.184	0.185	0.929
	Naive	0.498	0.105	0.105	0.917	2.000	0.200	0.200	0.902
β_{z2}	IV	0.204	0.150	0.150	0.956	3.198	0.245	0.245	0.950
	RC	0.203	0.219	0.219	0.934	3.205	0.300	0.300	0.948
	SIMEX	0.202	0.165	0.165	0.938	3.120	0.331	0.340	0.919
	Naive	0.198	0.185	0.185	0.901	3.210	0.370	0.370	0.781

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.8: Comparison of IV, RC, SIMEX, and naive methods for $n = 500$ and 10% censoring under two parameter settings for Experiment 3.

		$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (0.1, 0.4, 0.5, 0.2)$				$(\beta_0, \beta_x, \beta_{z1}, \beta_{z2}) = (1.0, 1.2, 2.0, 3.2)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.104	0.069	0.069	0.959	1.002	0.076	0.076	0.959
	RC	0.105	0.070	0.070	0.957	1.004	0.078	0.078	0.958
	SIMEX	0.096	0.081	0.081	0.946	0.985	0.095	0.096	0.947
	Naive	0.085	0.094	0.095	0.917	0.945	0.105	0.123	0.848
β_x	IV	0.401	0.036	0.036	0.955	1.200	0.055	0.055	0.952
	RC	0.398	0.075	0.075	0.946	1.196	0.082	0.082	0.948
	SIMEX	0.330	0.057	0.090	0.931	1.080	0.117	0.110	0.912
	Naive	0.160	0.088	0.224	0.701	0.910	0.213	0.360	0.491
β_{z1}	IV	0.502	0.050	0.050	0.952	2.000	0.095	0.095	0.949
	RC	0.504	0.054	0.054	0.948	2.004	0.072	0.072	0.945
	SIMEX	0.499	0.060	0.060	0.949	1.990	0.115	0.115	0.941
	Naive	0.500	0.065	0.065	0.932	2.000	0.120	0.120	0.923
β_{z2}	IV	0.202	0.113	0.113	0.955	3.201	0.180	0.180	0.951
	RC	0.198	0.125	0.125	0.939	3.196	0.150	0.150	0.953
	SIMEX	0.199	0.110	0.110	0.944	3.190	0.210	0.210	0.938
	Naive	0.195	0.136	0.136	0.931	3.205	0.240	0.240	0.824

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Tables A5 to A8 present the results of the additional simulations conducted for Experiment 3 under two parameter settings and different sample sizes. The pattern of the results are consistent with the results presented in chapter 4 in Experiment 3. Across both parameter settings, the IV estimator shows the lowest RMSE values and coverage probabilities close to the nominal level. While the RC method shows a competing performance, the SIMEX method shows sensitivity to the sample size. When comparing the results for Experiment 3 in Chapter 4 (Table 4.5) with Tables A5, A7 and A8, it can be noticed that the Tables A5, A7, and A8 shows better performance. For example, for the IV method, according to Table 4.5, for $n = 500$, under 10% censoring for the coefficient β_X , the RMSE value is 0.076 and the coverage probability is

0.946 while for the same case, according Table A8, the RMSE is 0.055 and the coverage probability is 0.952.

Table A.9: Comparison of IV, RC, SIMEX, and naive methods for $n = 50$ and 10% censoring under two parameter settings for Experiment 4.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.5, 0.4)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.102	0.156	0.156	0.942	1.006	0.206	0.206	0.940
	RC	0.102	0.167	0.167	0.935	0.990	0.220	0.220	0.941
	SIMEX	0.094	0.185	0.185	0.912	0.956	0.277	0.281	0.896
	Naive	0.087	0.206	0.206	0.752	0.911	0.319	0.331	0.791
β_x	IV	0.522	0.233	0.234	0.936	1.247	0.406	0.409	0.938
	RC	0.533	0.294	0.296	0.925	1.218	0.386	0.386	0.923
	SIMEX	0.259	0.063	0.249	0.781	0.691	0.165	0.534	0.765
	Naive	0.005	0.054	0.498	0.532	0.587	0.894	1.084	0.323
β_z	IV	0.376	0.336	0.337	0.945	1.956	0.476	0.478	0.940
	RC	0.397	0.370	0.370	0.937	2.011	0.460	0.460	0.931
	SIMEX	0.375	0.389	0.390	0.903	1.947	0.492	0.495	0.752
	Naive	0.377	0.329	0.330	0.684	2.499	0.206	0.540	0.452

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.10: Comparison of IV, RC, SIMEX, and naive methods for $n = 100$ and 10% censoring under two parameter settings for Experiment 4.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.5, 0.4)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.097	0.112	0.112	0.944	1.004	0.141	0.141	0.948
	RC	0.091	0.124	0.124	0.940	0.990	0.155	0.155	0.945
	SIMEX	0.087	0.143	0.144	0.921	0.966	0.191	0.194	0.924
	Naive	0.082	0.160	0.161	0.817	0.931	0.225	0.236	0.890
β_x	IV	0.508	0.177	0.177	0.946	1.218	0.257	0.258	0.951
	RC	0.513	0.238	0.238	0.928	1.219	0.299	0.300	0.932
	SIMEX	0.279	0.094	0.240	0.824	0.671	0.308	0.612	0.753
	Naive	0.005	0.102	1.199	0.543	2.801	0.325	1.634	0.331
β_z	IV	0.392	0.224	0.224	0.955	1.978	0.322	0.323	0.941
	RC	0.391	0.233	0.233	0.945	1.988	0.336	0.336	0.936
	SIMEX	0.346	0.250	0.256	0.911	2.044	0.325	0.328	0.904
	Naive	0.321	0.251	0.263	0.772	1.685	0.161	0.354	0.471

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.11: Comparison of IV, RC, SIMEX, and Naive methods for $n = 200$ and 10% censoring under two parameter settings for Experiment 4.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.5, 0.4)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.103	0.075	0.075	0.951	1.007	0.096	0.096	0.951
	RC	0.106	0.099	0.099	0.944	1.008	0.103	0.103	0.946
	SIMEX	0.096	0.097	0.097	0.936	0.973	0.139	0.142	0.938
	Naive	0.088	0.121	0.122	0.843	0.936	0.169	0.181	0.902
β_x	IV	0.500	0.125	0.125	0.947	1.205	0.176	0.176	0.954
	RC	0.513	0.152	0.153	0.935	1.220	0.195	0.196	0.946
	SIMEX	0.330	0.045	0.175	0.861	0.680	0.102	0.530	0.786
	Naive	0.013	0.045	0.489	0.554	0.140	0.064	1.062	0.331
β_z	IV	0.405	0.154	0.154	0.957	2.000	0.106	0.106	0.949
	RC	0.402	0.117	0.117	0.952	2.001	0.215	0.215	0.941
	SIMEX	0.388	0.164	0.164	0.931	1.871	0.104	0.163	0.911
	Naive	0.312	0.154	0.177	0.803	1.983	0.225	0.226	0.601

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.12: Comparison of IV, RC, SIMEX, and Naive methods for $n = 500$ and 10% censoring under two parameter settings for Experiment 4.

		$(\beta_0, \beta_x, \beta_z) = (0.1, 0.5, 0.4)$				$(\beta_0, \beta_x, \beta_z) = (1.0, 1.2, 2.0)$			
	Method	Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
β_0	IV	0.102	0.053	0.053	0.956	1.003	0.065	0.065	0.952
	RC	0.105	0.054	0.054	0.952	1.004	0.069	0.069	0.950
	SIMEX	0.097	0.067	0.067	0.939	0.980	0.094	0.096	0.941
	Naive	0.091	0.084	0.085	0.923	0.951	0.121	0.132	0.914
β_x	IV	0.500	0.074	0.074	0.959	1.202	0.078	0.078	0.959
	RC	0.501	0.093	0.093	0.949	1.207	0.118	0.118	0.948
	SIMEX	0.376	0.122	0.174	0.884	0.760	0.132	0.459	0.803
	Naive	0.022	0.050	0.481	0.573	0.151	0.028	1.050	0.356
β_z	IV	0.398	0.042	0.042	0.959	1.995	0.064	0.064	0.955
	RC	0.400	0.111	0.111	0.954	2.002	0.067	0.067	0.951
	SIMEX	0.405	0.050	0.050	0.947	2.034	0.067	0.075	0.938
	Naive	0.422	0.148	0.150	0.901	1.993	0.142	0.142	0.643

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Tables A9 to A12 present the simulation results of Experiment 4 under two parameter settings and different sample sizes. The results are consistent with the results presented in chapter 4 under Experiment 4. Across both the parameter settings the IV method shows the best overall performance with mean estimates closer to the true parameter value, lower RMSE values and coverage probabilities closer to the nominal level. While the RC method shows a competing performance, the SIMEX method shows lower coverage probabilities for small sample sizes. For example, for the IV estimator according to Table A9, for the coefficient β_x the coverage probability is 0.781 under $n = 50$. As expected, the naive method shows the lowest performance. When comparing the results for Experiment 4 in Chapter 4 (Table 4.7) with Tables A9, A11 and A12, it can be noticed that the Tables A9, A11 and A12 shows better performance. For example, for the IV method, according to Table 4.7, for

$n = 500$, under 10% censoring for the coefficient β_X , the RMSE value is 0.098 and the coverage probability is 0.941 while for the same case, according Table A12, the RMSE is 0.078 and the coverage probability is 0.959.

A.1 Departure from Normality

In order to evaluate the performance of the IV method under the violation of normality of the errors u and e , the distribution of each error was changed one at a time as $e \sim t(5)$ and $u \sim t(5)$ while the other distributions were kept the same under 10% censoring. The results obtained are as follows:

Table A.13: Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 1.

n	Parameter	$e \sim t(5)$				$u \sim t(5)$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
50	$\beta_0 = 1.0$	0.967	0.188	0.191	0.918	0.969	0.184	0.186	0.921
	$\beta_x = 1.2$	1.185	0.132	0.134	0.920	1.187	0.141	0.143	0.922
	$\beta_z = 2.0$	1.964	0.139	0.144	0.933	1.968	0.146	0.148	0.941
200	$\beta_0 = 1.0$	0.981	0.150	0.151	0.931	0.984	0.146	0.147	0.936
	$\beta_x = 1.2$	1.193	0.074	0.075	0.934	1.194	0.071	0.072	0.939
	$\beta_z = 2.0$	1.978	0.091	0.092	0.936	1.981	0.087	0.088	0.940
500	$\beta_0 = 1.0$	0.986	0.102	0.103	0.939	0.988	0.098	0.099	0.944
	$\beta_x = 1.2$	1.196	0.047	0.048	0.941	1.197	0.045	0.046	0.944
	$\beta_z = 2.0$	1.985	0.058	0.060	0.942	1.987	0.055	0.056	0.945

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.14: Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 2.

n	Parameter	$e \sim t(5)$				$u \sim t(5)$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
50	$\beta_0 = 1.0$	0.972	0.214	0.216	0.913	0.975	0.205	0.206	0.918
	$\beta_x = 1.2$	1.189	0.156	0.157	0.918	1.191	0.148	0.149	0.922
	$\beta_z = 2.0$	1.967	0.188	0.191	0.922	1.970	0.178	0.180	0.928
200	$\beta_0 = 1.0$	0.984	0.162	0.163	0.930	0.986	0.156	0.157	0.936
	$\beta_x = 1.2$	1.193	0.078	0.079	0.936	1.194	0.074	0.075	0.940
	$\beta_z = 2.0$	1.975	0.096	0.097	0.936	1.978	0.091	0.092	0.940
500	$\beta_0 = 1.0$	0.989	0.108	0.109	0.939	0.990	0.103	0.104	0.943
	$\beta_x = 1.2$	1.196	0.051	0.052	0.941	1.197	0.048	0.049	0.944
	$\beta_z = 2.0$	1.985	0.061	0.062	0.941	1.987	0.058	0.059	0.944

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.15: Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 3.

n	Parameter	$e \sim t(5)$				$u \sim t(5)$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
50	$\beta_0 = 1.0$	1.016	0.247	0.248	0.916	1.018	0.238	0.239	0.920
	$\beta_x = 1.2$	1.191	0.171	0.172	0.918	1.193	0.164	0.165	0.922
	$\beta_{z1} = 2.0$	1.984	0.233	0.234	0.939	1.986	0.225	0.226	0.943
	$\beta_{z2} = 3.2$	3.209	0.306	0.307	0.940	3.212	0.294	0.295	0.944
200	$\beta_0 = 1.0$	1.012	0.126	0.127	0.936	1.014	0.121	0.122	0.940
	$\beta_x = 1.2$	1.196	0.134	0.134	0.938	1.197	0.129	0.129	0.941
	$\beta_{z1} = 2.0$	1.993	0.134	0.135	0.948	1.995	0.129	0.130	0.951
	$\beta_{z2} = 3.2$	3.188	0.170	0.171	0.944	3.190	0.163	0.164	0.947
500	$\beta_0 = 1.0$	1.008	0.082	0.083	0.944	1.009	0.079	0.080	0.947
	$\beta_x = 1.2$	1.199	0.084	0.084	0.943	1.199	0.081	0.081	0.945
	$\beta_{z1} = 2.0$	1.998	0.084	0.084	0.951	1.999	0.081	0.081	0.953
	$\beta_{z2} = 3.2$	3.192	0.108	0.109	0.947	3.193	0.104	0.105	0.950

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.16: Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 4.

n	Parameter	$e \sim t(5)$				$u \sim t(5)$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
50	$\beta_0 = 1.0$	0.905	0.228	0.247	0.913	0.908	0.220	0.237	0.918
	$\beta_x = 1.2$	1.170	0.220	0.222	0.914	1.172	0.212	0.214	0.918
	$\beta_{z1} = 2.0$	2.001	0.235	0.235	0.918	2.003	0.227	0.227	0.922
	$\beta_{z2} = 3.2$	3.236	0.238	0.242	0.919	3.238	0.229	0.233	0.923
200	$\beta_0 = 1.0$	0.995	0.178	0.178	0.928	0.997	0.171	0.171	0.932
	$\beta_x = 1.2$	1.191	0.118	0.118	0.931	1.192	0.113	0.113	0.935
	$\beta_{z1} = 2.0$	1.993	0.135	0.136	0.934	1.995	0.129	0.129	0.938
	$\beta_{z2} = 3.2$	3.198	0.126	0.126	0.935	3.200	0.121	0.121	0.939
500	$\beta_0 = 1.0$	1.007	0.123	0.123	0.936	1.009	0.118	0.118	0.939
	$\beta_x = 1.2$	1.198	0.108	0.108	0.938	1.198	0.103	0.103	0.940
	$\beta_{z1} = 2.0$	2.000	0.073	0.073	0.939	2.001	0.070	0.070	0.942
	$\beta_{z2} = 3.2$	3.201	0.117	0.117	0.939	3.202	0.112	0.112	0.942

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Table A.17: Simulation results for different sample sizes under heavy tailed measurement error and model error for Experiment 5.

n	Parameter	$e \sim t(5)$				$u \sim t(5)$			
		Est	MCSD	RMSE	CP	Est	MCSD	RMSE	CP
50	$\beta_0 = 1.0$	1.008	0.192	0.192	0.931	1.010	0.184	0.184	0.935
	$\beta_x = 1.2$	1.178	0.137	0.139	0.936	1.180	0.131	0.132	0.939
	$\beta_z = 2.0$	1.946	0.159	0.168	0.933	1.948	0.152	0.160	0.937
200	$\beta_0 = 1.0$	1.008	0.168	0.168	0.939	1.009	0.161	0.161	0.942
	$\beta_x = 1.2$	1.186	0.124	0.125	0.941	1.187	0.119	0.120	0.944
	$\beta_z = 2.0$	1.952	0.151	0.156	0.940	1.954	0.145	0.150	0.943
500	$\beta_0 = 1.0$	1.009	0.156	0.156	0.942	1.010	0.149	0.149	0.944
	$\beta_x = 1.2$	1.179	0.089	0.091	0.943	1.180	0.085	0.087	0.945
	$\beta_z = 2.0$	2.006	0.132	0.132	0.944	2.008	0.126	0.126	0.946

Note: MCSD denotes the Monte Carlo standard deviation of the parameter estimates across simulation replications.

Tables A13 to A15 show the results for the IV method when the normality assumption is violated for measurement error e and latent covariate model error u . Across all five experiments, the IV estimator produced estimates closer to the true values. Specifically, for higher sample sizes it shows lower RMSE values and coverage fairly closer to the nominal level, but compared to the scenario where e and u are normally distributed, the performance is low. For example, when we consider the results for Experiment 4, when the errors are normally distributed, according to Table 4.7, for the coefficient β_x under $n = 500$ the RMSE value is 0.098 and the coverage probability is 0.941. For the same scenario, when the normality assumption is violated according to Table A16, for e and u , the RMSE values and coverage probabilities are 0.118 and 0.931, and 0.113 and 0.935, respectively. Moreover, when looking through all five experiments, it can be observed that the violation of normality in the measurement error e affects more adversely compared to the case where the violation in normality of the latent covariate model error u .

Index

acknowledgment, [i](#)

dedication, [ii](#)

introduction, [7](#)

thesis regulations, [i](#), [ii](#)

 acknowledgment, [i](#)

 dedication, [ii](#)