Spring 2020

# Nonparametric Misclassification Simulation and Extrapolation Method and Its Application

Congjian Liu

NONPARAMETRIC MISCLASSIFICATION SIMULATION AND EXTRAPOLATION METHOD AND ITS APPLICATION

by

CONGJIAN LIU

(Under the Direction of Lili Yu)

ABSTRACT

The misclassification simulation extrapolation (MC-SIMEX) method proposed by Küchenho et al. is a general method of handling categorical data with measurement error. It consists of two steps, the simulation and extrapolation steps. In the simulation step, it simulates observations with varying degrees of measurement error. Then parameter estimators for varying degrees of measurement error are obtained based on these observations. In the extrapolation step, it uses a parametric extrapolation function to obtain the parameter estimators for data with no measurement error. However, as shown in many studies, the parameter estimators are still biased as a result of the parametric extrapolation function used in the MC-SIMEX method. Therefore, we propose a nonparametric MC-SIMEX method in which we use a nonparametric extrapolation function. It uses the fractional polynomial method with cross-validation to choose the appropriate fractional polynomial terms. An example is provided based on data from the National Health and Nutrition Examination Survey.

INDEX WORDS: Misclassification error, MC-SIMEX, Cross-validation, Fractional polynomial, Logistic regression, Parameter estimator

# NONPARAMETRIC MISCLASSIFICATION SIMULATION AND EXTRAPOLATION METHOD AND ITS APPLICATION

by

CONGJIAN LIU

B.A., Sichuan University, P.R. China, 2007

M.S., Georgia Southern University, 2015

A Dissertation Submitted to the Graduate Faculty of Georgia Southern University

in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PUBLIC HEALTH

# NONPARAMETRIC MISCLASSIFICATION SIMULATION AND EXTRAPOLATION METHOD AND ITS APPLICATION

by

CONGJIAN LIU

| | |
|---|---|
| Major Professor: | Lili Yu |
| Committee: | Jingjing Yin |
| | Jun Liu |

Electronic Version Approved:
May 2020

# ACKNOWLEDGMENTS

I would like to thank my wife Yisong, my girls Coraline and Christina for their support through my study period. I would like to thank Dr. Lili Yu and Dr. Jingjing Yin from the Department of Biostatistics at the Jiann Ping Hsu College of Public Health, Dr. Jun Liu from the Department of Enterprise Systems and Analytics at Parker College of Business, and other professors of the Department of Biostatistics at the Jiann Ping Hsu College of Public Health for providing their time and effort in making this dissertation a success.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

# LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION

Classical measurement error refers to the truth being measured with additive error (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006). Random error and systematic error are two types of error (Taylor, 1997). In a set of measurements, systematic measurement errors (also called bias) are consistent, repeatable errors (Cimbala, 2009). In this dissertation, the focus is on systematic measurement error.

Nondifferential and differential are two types of systematic measurement error (Carroll et al., 2006). Let *W* denote the measurement error covariate, which means *W* is the observed covariate with measurement error and without any correction. Let *X* denote the (unobserved) true and gold standard covariate. Let *Y* denote the binary response variable. Let *Z* denote another covariate without measurement error in the model. The definition of nondifferential measurement error is that *W* does not depend on the response *Y* (Carroll et al., 2006). For example, in the case of diet, nondifferential measurement error can occur when instead of measuring a participant's long-term diet *X*, the measured *W* was each participant's diet in the previous 24 hours (Carroll et al., 2006). Otherwise, the measurement error is differential. Namely, *W* provides additional information about *Y*. In the study of the previous example, since the response variable *Y*, the diagnosis of cancer, is obtained first, a subject may change his or her diet after diagnosis. Thus, each participant's diet in the previous 24 hours, *W*, is correlated with cancer outcome *Y* (Carroll et al., 2006).

For a discrete variable, we refer the measurement error as misclassification. Measurement error proverbially exists in real data. For example, Armstrong has shown in a study of the relationship between lung cancer (*Y*) and the distance from a residence to a coke oven (*X*), that misclassification occurred since migration—some subjects in the study had moved their home during the follow-up period (Armstrong B. G., 1998). Millner et al. pointed out that in a national suicide (*Y*) study, respondents may not clearly understand the specific behavior (suicidal behavior, *Y*) in question. In addition, there are many more subtle steps in the process of attempting suicide that are omitted when using single-item assessment. Hence, covariate single-item assessment (*X*) owns misclassification (Millner, Lee, & Nock, 2015).

Measurement error leads to bias in estimated regression coefficients for statistical models; causes a loss of power, sometimes profound; and makes graphical model analysis difficult (Carroll et al., 2006). Several statistical methods have been proposed to correct estimation bias. Regression calibration is one of the popular methods in the misclassification literature used to correct the bias caused by misclassification (Armstrong B. , 1985; Carroll & Stefanski, 1990; Fraser & Stram, 2001; Bang et al., 2013). In this method, the value of the true covariate $X$ is estimated by regressing $X$ on the naive covariate $W$ (Rosner, Willett, & Spiegelman, 1989; Carroll et al., 2006). Based on the regression calibration method, Spiegelman et al. studied the effect of misclassification by combining the regression calibration estimator and an estimator from the validation data, called the pooled estimator method (Spiegelman, Carroll, & Kipnis, 2001). Cole et al. suggested using multiple imputation as a correction method (Cole, Chu, & Greenland, 2006). Researchers fit a logistic regression model between the true covariate $X$ and the naive covariate $W$ in the validation data. Thus, researchers can replace the naive covariate in the nonvalidation data by the estimated probability from the model (Rubin, 1976; Carroll et al., 2006). The corrected score estimator was proposed by Zucker and Spiegelman under survival analysis by using a corrected score function to estimate the parameters and standard errors (Zucker & Spiegelman, 2008). Simulation and extrapolation (SIMEX) method is another statistical approach that can correct the bias created by measurement error in the continuous variable(s) (Cook & Stefanski, 1994).

To deal with the effect of bias caused by misclassified discrete covariates, MC-SIMEX was developed from SIMEX by using a parametric extrapolation function and misclassification rates (sensitivity and specificity) (Küchenhoff, Mwalili, & Lesaffre, 2006). Just as the SIMEX method, the MC-SIMEX method is a simulation-based method that makes efficient use of sensitivity and specificity to produce bias-corrected estimates.

The estimated regression coefficients are still, however, biased by using the MC-SIMEX method. The estimation bias of misclassification has been shown in the logistic regression by Küchenhoff et al., the log-normal accelerated failure time model (AFT model) by Slate and Bandyopadhyay, and the log-logistic

AFT model by Sevilimedu (Küchenhoff et al., 2006; Slate & Bandyopadhyay, 2009; Sevilimedu, 2017). We notice that the bias may be caused by the parametric extrapolation function used in MC-SIMEX. Said function may not approximate the true function plate.

Therefore, in this dissertation, we modify the MC-SIMEX method by proposing a nonparametric MC-SIMEX method. We use a nonparametric extrapolation function, which is estimated by the fractional polynomial method with cross-validation. The simulation shows that it corrects the estimation bias well.

This dissertation is organized as follows. In Chapter 2, we provide a literature review. In Chapter 3, we introduce the model formulation, discuss the bias of ignoring measurement error in covariates, and describe the nonparametric MC-SIMEX method. To assess the performance of the new method and the impact of ignoring error in covariates on the estimation of the regression parameters, simulation studies are conducted in Chapter 4. An example is presented in Chapter 5 to illustrate the proposed method, followed by a general discussion in the last chapter.

# CHAPTER 2

# LITERATURE REVIEW

In this chapter, we will review the methods that are used to handle the measurement error in data sets. At the end of the chapter, we will introduce the fractional polynomial method, which is used in the nonparametric MC-SIMEX method.

## Regression calibration

Regression calibration (RC) is a standard method of dealing with measurement errors and their effects (Bang et al., 2013). It estimates the true covariates $X$ by regressing $X$ against the naive covariate $W$ in validation data. Then the $X$ in the nonvalidation data is replaced by the estimated values of $X$ (Bang et al., 2013; Carroll et al., 1990). The standard error of the estimate is calculated by the bootstrapping or sandwich methods (Carroll et al., 2006).

Agogo et al. used the RC method to adjust for the attenuation caused by measurement error in dietary intake in a single-replicate study design (Agogo et al., 2014). Rosner et al. applied the regression calibration method to study the effect of measurement error in fat, calories, or alcohol intake in logistic regression (Rosner, Spiegelman, & Willett, 1990). They also suggested the RC method to correct bias of relative risk estimates caused by measurement error of exposure that is independent of disease status (Rosner, Willett, & Spiegelman, 1989).

The advantage of the RC method is that it is convenient and highly popular for discrete data and nonnormal data (Sevilimedu, 2017). However, the limitation of the RC method is that it only deals with nondifferential measurement error.

## Pooled estimator

Spiegelman et al. developed the pooled estimator based on the regression calibration method by combining the regression calibration estimator and an estimator from the validation data (Spiegelman et al., 2001). When the validation sample is large, the efficiency is increased compared to the regression

calibration estimator (Bang et al., 2013). Therefore, choosing an appropriately large validation sample is important when using the pooled estimator method (Spiegelman et al., 2001). Bang et al. applied the pooled estimation method to survival analysis (Bang et al., 2013). Because the pooled estimator is based on the RC estimator, the assumptions of the RC method are also required here. The limitation of the method is that large validation data sets are not always available.

## Multiple imputation

Multiple imputation (MI) is suggested by Cole et al. to deal with the measurement error problem (Cole et al., 2006). When applying a multiple imputation procedure, researcher fits a logistic regression model between the $X$ and the naive covariate $W$ in the validation data. The naive covariate in the nonvalidation data is then replaced with the corrected value by using the estimated probability from the model aforementioned (Bang et al., 2013).

Cole et al. proposed using the MI method to correct the bias of the estimated hazard ratios for end-stage renal disease (Cole et al., 2006). Edwards et al. pointed out that the estimated bias in the model could be corrected by using the MI method to strengthen results from observational studies (Edwards et al., 2015).

However, the MI method has many limitations. First, the correct specification of the model is critical to its successful performance. The second disadvantage is that for survival analysis, the data set with censored outcomes is more difficult to implement. Many researchers noted this problem in their work (Qi, Wang, & He, 2010; White I. R., 2006). Finally, the performance of the MI method depends on the sample size or the proportion validated (Cole et al., 2006).

## Corrected score

The corrected score (CS) estimator was proposed by Zucker and Spiegelman for use under survival analysis by using a corrected score function to estimate parameters and standard errors (Zucker et al., 2008). The corrected score function, which equals the true score function in expected value, is used for the estimation of the parameters. The standard errors are calculated by the bootstrap method or sandwich method (Carroll et al., 2006; Bang et al., 2013; Sevilimedu, 2017).

Akazawa et al. applied the CS method to logistic regression to adjust the estimated offset the bias from misclassification (Akazawa, Kinukawa, & Nakamura, 1998). Zucker and Spiegelman corrected the estimated bias of misclassification in a covariate in Cox model (Zucker et al., 2008).

An advantage of this method is that it can handle models with both continuous and discrete covariates (Akazawa et al., 1998). The limitation of the method is that measurement error distribution must be known (Chen, Hanfelt, & Huang, 2015). Also, numerical problems occur when applying the CS method at the situation that the count of subjects at risk get smaller as time progresses in survival analysis (Sevilimedu, 2017).

## Simulation and extrapolation

The simulation and extrapolation (SIMEX) method was created by Cook and Stefanski to address measurement error of continuous variable (Cook et al., 1994). It assumes that the effect of measurement error on an estimator can be determined experimentally via simulations (Cook et al., 1994). The SIMEX method consists of two steps: the simulation and the extrapolation steps. In the simulation step, researchers add varying degrees of additional measurement error to the data to simulate observations. Researchers then obtain the parameter estimators for varying degrees of measurement error based on these observations. In the extrapolation step, researchers use a parametric extrapolation function to obtain the parameter estimators for data with no measurement error.

Pina-Sánchez applied the SIMEX method to correct recall errors in duration data (Pina-Sánchez, 2016). Hardin et al. suggested using the SIMEX method to correct the effect of measurement error on recall measurements recorded for calories of saturated fat intake (Hardin, Arnold, Schmiediche, & Carroll, 2003). Lederer and Küchenhoff employed the SIMEX method to address the effect of measurement error on dust in chronic bronchitis and dust concentration of the Deutsche Forschungsgemeinschaft study (Lederer & Küchenhoff, 2006). However, this method can only apply to continuous variables.

## Misclassification simulation and extrapolation

The misclassification simulation and extrapolation (MC-SIMEX) method was developed by Küchenhoff et al. from SIMEX to correct the effect of misclassified discrete covariates (Küchenhoff et al., 2006). The SIMEX and MC-SIMEX methods use consistent processes, including simulating observations in the simulation step. In addition, researchers obtain the parameter estimators for varying degrees of measurement error. Then they obtain the parameter estimators for data with no measurement error in the extrapolation step.

Küchenhoff et al. applied the method to address the effect of bias in children's probability of developing caries (Küchenhoff et al., 2006). Slate and Bandyopadhyay applied the MC-SIMEX method to study the effect of misclassification within periodontal outcomes in the log-normal AFT model (Slate et al., 2009). Sevilimedu proposed a modified MC-SIMEX method in the log-logistic AFT model and applied it in a prospective study of dietary fat intake and risk of breast cancer (Sevilimedu, 2017).

However, the parameter estimators are still biased due to the parametric extrapolation function used in the MC-SIMEX method.

## Fractional polynomial method

The fractional polynomial (FP) method was first introduced by Royston and Altman for continuous covariates and was expanded to categorized covariates by Sauerbrei and Royston to determine if the value of $p$ in $x^p$ yields the best model for the data (Royston & Altman, 1994; Sauerbrei & Royston, 1999; Hosmer, Lemeshow, & May, 2008). Fractional polynomials are an extended family of curves, whose power terms are restricted to a small, predefined set of values (Royston et al., 1994). The powers are selected so that conventional polynomials are a subset of the family. When using this method, a more complex model should be retained only when there is enough evidence that it is better than a simpler one (Nikolaeva, Bhatnagar, & Ghose, 2015).

Royston and Sauerbrei applied the FP method to model continuous risk variables (Sauerbrei et al., 1999). Mayer et al. employed the FP method to estimate the half-life periods in nonlinear data (Mayer,

Keller, Syrovets, & Wittau, 2013). Zhang introduced the FP method to continuous covariates in the German Breast Cancer Study Group (GBSG) database (Zhang, 2016).

However, the FP method is not suitable for small samples (Nikolaeva et al., 2015). Another disadvantage is that lack of flexibility may lead to a poor fit of the models (Sauerbrei et al., 1999). We will introduce the fractional polynomial method in detail in the next chapter.

# CHAPTER 3

# METHODOLOGY

This dissertation aims to correct the bias of misclassification in logistic regression by applying the nonparametric MC-SIMEX method. In the following sections, we will introduce the logistic regression model, the fractional polynomial method, the original MC-SIMEX method, and the nonparametric MC-SIMEX method in details.

## Logistic regression model

The logistic regression model is written as the regression model of the log of odds over covariates; i.e.,

$$logit(p_i) = log\frac{p_i}{1-p_i} = X_i'\beta, \tag{3.1}$$

where $p_i$ is the probability of the event that $Y_i = 1$, $X_i$ represents binary covariates and $\beta$ represents regression coefficients. We set $X_0 = 1$ corresponding to intercept coefficient $\beta_0$.

Solving for the $p_i$ in equation (3.1) gives the following function,

$$p_i = \frac{exp(X_i'\beta)}{1+exp(X_i'\beta)} \tag{3.2}$$

Note that $Y_i$ follows the Bernoulli distribution with parameter $p_i$ taking the following form (Rohatgi & Saleh, 2000),

$$Pr\{Y_i = y_i\} = p_i^{y_i}(1-p_i)^{1-y_i}, \quad y_i = 0,1. \tag{3.3}$$

*Likelihood function of logistic regression model*

Based on equations (3.1) and (3.3), the likelihood function is as follows,

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}. \quad (3.4)$$

Next, the log-likelihood turns products into sums and gives the log-likelihood function of the logistic regression model,

$$logL(\beta) = \sum_{i=1}^{n} \{ y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \}. \quad (3.5)$$

Based on equation (3.2), the aforementioned equation could be rewritten in following form,

$$logL(\beta) = \sum_{i=1}^{n} \{ y_i log \left( \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \right) + (1 - y_i) log \left( 1 - \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \right) \}. \quad (3.6)$$

Multiple Studies showed that the maximum likelihood estimation (MLE) of full log-likelihood of logistic regression model does not have a closed form (Rohatgi et al., 2000; Hogg, Craig, & McKean, 2005; Czepiel, 2002; Hilbe, 2017). However, Hogg et al. showed that the MLE exists in general and is unique (Hogg et al., 2005). Numerical methods, such as the Newton-Raphson method is widely used to obtain the MLE of log-likelihood of the logistic regression model (Rohatgi et al., 2000; Hogg et al., 2005; Czepiel, 2002; Hilbe, 2017). In most calculus textbooks, we can find the description and application of the Newton-Raphson method to obtain the MLE. Suppose $\beta_0$ is an initial guess at the solution and $\beta_1$ is the next guess, which is the horizontal intercept of the tangent line to the curve $logL'(\beta)$ at the point $(\beta^0, logL'(\beta^0))$, where $logL'(\beta)$ is the first derivative of function (3.6). Thus, $logL''(\beta)$ represents the second derivative of function (3.6). Then,

$$\beta^1 = \beta^0 - \frac{logL'(\beta^0)}{logL''(\beta^0)}. \quad (3.7)$$

The aforementioned process is repeated until convergence (Hogg et al., 2005).

## Fractional polynomial method

The fractional polynomial (FP) method is aimed to determine the value of $p$ in $x^p$ so that the fitted model yields the best model for the data (Hosmer et al., 2008). In theory, the value of $p$ could be any real

number. Royston and Altman proposed that a search through a set $\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ with possible transformation could avoid the complexity of the estimated problem (Royston et al., 1994).

*Basic notation and formula*

Let $J$ denote the terms of power $p$. In most of studies, $J = 1$ or 2 is the possible value (Hosmer et al., 2008). When $J = 1$, there is one term of $x^p$ in the model, namely, $y \sim x^{p_1}, p_1 \in \mathcal{P}$, and called the FP1 model. When $J = 2$, there are two terms of $x^p$ in the model, namely, $y \sim x^{p_1} + x^{p_2}, p_1 \in \mathcal{P}, p_2 \in \mathcal{P}$, and called the FP2 model. Hence there are 8 FP1 models and 36 FP2 models. Note that there are two conventions in the transformation (Hosmer et al., 2008). First when $p = 0$ in the fitted model, $x^0$ is replaced by $\ln(x)$. Second, for models that involve repeated powers such as $(p_1, p_2) = (2, 2)$, the second term is multiplied by $\ln(x)$. Namely, the fitted model is $y \sim x^{p_1} + x^{p_2} * \ln(x), p_1 \in \mathcal{P}, p_2 \in \mathcal{P}$.

In the method, the best model is the one with the largest log partial likelihood (Hosmer et al., 2008). Let L(0) denote the log partial likelihood of the null model where *x* is not in the model. Let L(1) denote the log partial likelihood of the linear model; that is, $x^p = x$. Let L($p_1$) denote the largest log partial likelihood of FP1 models. And L($p_1, p_2$) denotes the largest log partial likelihood of FP2 models. Note that Royston et al. pointed out that each term in the FP model contributes approximately 2 degrees of freedom (df) to the model, one for the coefficient and one for the power (Royston et al., 1994; Royston, Ambler, & Sauerbrei, 1999). Hence, to compare the linear model to the best FP1 model,

$$G(1, p_1) = -2[L(1) - L(p_1)], \tag{3.8}$$

is approximately distributed as chi-square with 1 df under the null hypothesis of linearity (Hosmer et al., 2008). Thus, the function,

$$G(p_1, (p_1, p_2)) = -2[L(p_1) - L(p_1, p_2)] \tag{3.9}$$

is the test comparing the best FP1 model to the best FP2 model with chi-square and 2 df under the null hypothesis that the second FP function term is equal to 0. Similarly, the test comparing the linear and the best FP2 model is approximately distributed as chi-square with 3 df.

*Model fitting*

For FP1 models, choosing $p_1 \in \mathcal{P}$, we can fit 8 FP1 models to data by setting the model $y \sim x^{p_1}$. The power $p_1$ in the model with the largest log partial likelihood is selected for the next step.

For FP2 models, choosing $p_1 \in \mathcal{P}$ and $p_2 \in \mathcal{P}$, 36 FP2 models can be obtained to data by setting the model $y \sim x^{p_1} + x^{p_2}$. The power combination $(p_1, p_2)$ in the model with the largest log partial likelihood is selected for the next step.

*Closed test*

A closed test could be used to find the best FP model (Hosmer et al., 2008). In the closed test procedure, we begin by comparing the linear model to the best of the FP2 models by performing the test,

$$G\big(1, (p_1, p_2)\big) = -2[L(1) - L(p_1, p_2)], \tag{3.10}$$

with chi-square and 3 df. If this test result is not significant, then we stop here and the best model we choose is the linear model. If the test results are significant, then we compare the best of FP1 models to the best of FP2 models via $G\big(p_1, (p_1, p_2)\big)$. If the test results are significant, then we use the best model of the FP2 models, otherwise we use the best model of the FP1 models as the best fractional polynomial model.

Original MC-SIMEX method

This section is based on Küchenhoff et al.'s (Küchenhoff et al., 2006) and Stefanski and Cook's (Cook & Stefanski, 1994)work. A misclassification matrix Π is defined with components (Küchenhoff et al., 2006),

$$\Pi = \begin{bmatrix} \pi_{00} & 1 - \pi_{11} \\ 1 - \pi_{00} & \pi_{11} \end{bmatrix}, \tag{3.11}$$

where

$$\pi_{11} = Sensitivity\ (SE) = P(W = 1 | X = 1),$$
$$\pi_{00} = Specificity(SP) = P(W = 0 | X = 0).$$

Hence Π gives the probabilities of misclassification. Note that Π is a *K*K* matrix where *K* is the number of possible outcomes for *X*. In addition, Π is known or can be estimated from validation data (Küchenhoff et al., 2006).

The parameter β in equation (3.1) is the parameter of interest. Let $\hat{\beta}$ denote the naive estimation of β. The proof for the existence of $\hat{\beta}$ and its estimation is given in the works of White (White H. , 1982). Because the estimate of $\hat{\beta}$ depends on Π, we denote it as $\hat{\beta}(\Pi)$. In addition, note that the estimator with no misclassification is $\hat{\beta}(I_{k*k})$., where $I_{k*k}$ is the identity matrix.

In the MC-SIMEX method, the authors define the function (Küchenhoff et al., 2006),

$$\lambda \to \hat{\beta}(\Pi^{\lambda}), \tag{3.12}$$

indicating that $\hat{\beta}\left(\Pi^{\lambda}\right)$ is a function of $\lambda$, where $\lambda \geq 0$. The spectral decomposition shows

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

where $\lambda$ is the diagonal matrix of eigenvalues of $\Pi$, and E is the corresponding eigenvector of Π (Küchenhoff et al., 2006). Based on function (3.12), if *W* has a relationship to *X* as a result of the misclassification matrix Π , $W^*$ has a relationship to *W* as a result of the misclassification matrix $\Pi^{\lambda}$, then $W^*$ has a relationship to *X* as a result of the misclassification matrix $\Pi^{1+\lambda}$, assuming that the two misclassification mechanisms are independent. To make function (3.9) well defined, Gastwirth proved that $\det(\Pi) = \pi_{00} + \pi_{11} - 1 > 0$, i.e., $\pi_{00} > 0.5$ and $\pi_{11} > 0.5$, is necessary to ensure the existence of $\Pi^{\lambda}$ (Gastwirth, 1987).

The MC-SIMEX method consists of a simulation step and an extrapolation step which is explained in detail.

*Simulation step*

The simulation step simulates datasets with varying degrees of misclassification as a result of the misclassification matrix $\Pi^{\lambda}$.

For a fixed grid of values $\lambda_k \in (\lambda_1 \dots \dots \lambda_m), b = 1, \dots, B, W_i$, are simulated by,

$$W_{b,i}(\lambda_k) \coloneqq MC\left[\Pi^{\lambda_k}\right](W_i), \qquad i = 1 \dots, n; k = 1, \dots, m, \tag{3.13}$$

Namely, we can obtain $W_{b,i}(\lambda_k)$ by inflating the misclassification in $W_i$ by a factor $\lambda_k$. Hence, we can denote the naive estimator as,

$$\hat{\beta}_{na}(\lambda_k) := B^{-1} \sum_{b=1}^{B} [\, \hat{\beta}_{na}(Y_i, W_{b,i}(\lambda_k), Z_i)], \quad k = 1, \dots, m; \quad i = 1, \dots, n. \tag{3.14}$$

Namely, the mean value of the naive estimators over B bootstrap samples is the naive estimator for a particular $\lambda_k$.

The $\lambda_k$ values are chosen as $\lambda_k \in (0, 2]$ (Cook et al., 1994). In addition, a large value should be chosen for *B* so that the Monte Carlo error is negligible (Cook et al., 1994). Stefanski and Cook showed that the MC-SIMEX method performs well using *B* = 50 (Cook et al., 1994). After the development of the computational resource, we can use a larger value for B .

*Extrapolation step*

The corresponding parameter estimates produced with each degree of misclassification are extrapolated using a parametric function of the form (Küchenhoff et al., 2006).

$$\lambda \to \hat{\beta}\left(\Pi^{\lambda}\right) \approx D(1 + \lambda, \Gamma), \tag{3.15}$$

where *D* is the extrapolation function, and Γ is the vector of parameters for the extrapolation function. For example, $D(1 + \lambda, \Gamma) = \Gamma_0 + \Gamma_1(1 + \lambda) + \Gamma_2(1 + \lambda)^2$ is the quadratic extrapolation function. After Γ is estimated, we extrapolate $D(1 + \lambda, \Gamma)$ to a point on the y-axis where $1 + \lambda = 0$ to obtain the estimator $\hat{\beta}_{MCsimex}$. Namely,

$$\hat{\beta}_{MCsimex} = D(1 + \lambda, \Gamma), \lambda = -1. \tag{3.16}$$

*Variance estimation of estimator*

Within a single simulation with B bootstrap samples, a given misclassification matrix, and a fixed grid of values $\lambda_k, k = 1, \ldots, m$ values, the sample variance of the estimator $\hat{\beta}_{sim}(\lambda_k)$ could be calculated by the following formulations,

$$\hat{V}_{sim}(\lambda_k) \coloneqq B^{-1} \sum_{b=1}^{B} \{ \hat{\beta}_{na}[(Y_i, W_{b,i}(\lambda_k), Z_i)_{i=1}^{n}] - \hat{\beta}_{na}(\lambda_k)\}^2 \tag{3.17}$$

where $\hat{\beta}_{na}(\lambda_k)$ is in equation (3.14), $k = 1, \ldots, m$. Note that $V_{sim}(0) \coloneqq 0$ . In addition, $\hat{V}_{naive}(\hat{\beta}_{na}[(Y_i, W_{b,i}(\lambda_k), Z_i)_{i=1}^{n}])$ is the variance for each naive estimate, which is calculated based on the information matrix for each value of $\lambda_k$. Then, we have,

$$\hat{V}_{na}(\lambda_k) \coloneqq B^{-1} \sum_{b=1}^{B} \hat{V}_{naive}\,(\hat{\beta}_{na}[(Y_i, W_{b,i}^{*}(\lambda_k), Z_i)_{i=1}^{n}]) \tag{3.18}$$

Stefanski and Cook suggested that the variance estimator of the MC-SIMEX estimator ($V_{ST}$) is given by the extrapolation of the difference between the sample variance and the variance obtained through the information matrix (Cook et al., 1994). Namely,

$$V_{ST}(\lambda_k) = V_{na}(\lambda_k) - V_{sim}(\lambda_k). \tag{3.19}$$

and

$$V_{ST} = \lim_{\lambda \to -1} \big(V_{na}(\lambda) - V_{sim}(\lambda)\big). \tag{3.20}$$

nonparametric MC-SIMEX

Because the parameter estimators are still biased as a result of the parametric extrapolation function used in the MC-SIMEX, we propose a nonparametric MC-SIMEX method to correct the estimated bias in the logistic regression model.

*Simulation step*

In the simulation step, the nonparametric MC-SIMEX method has the same procedure as the original MC-SIMEX method, except we use m = 100 because we need to estimate the extrapolation function

by the FP method with the cross-validation process. In addition, some power coefficients in $\mathcal{P}$ require the value of the base to be positive, such as $\lambda^{\frac{1}{2}}$. However, based on the original MC-SIMEX method, we extrapolate the extrapolation function to the point on the Y-axis where $\lambda = -1$. Hence, we replace $\lambda$ in the original MC-SIMEX method by $\exp(\lambda)$. Küchenhoff et al. pointed out the exponential in $\lambda$ works very well in the extrapolation process (Küchenhoff et al., 2006).

*Extrapolation step*

We use the FP method to approximate the extrapolation function, which is,

$$\lambda \to \hat{\beta}\left(\Pi^{\lambda}\right) \approx D(1+\lambda, \Gamma).$$

Since $\lambda$ is replaced by $exp(\lambda)$, aforementioned function (which is the same as function 3.15) could be rewritten as,

$$exp(\lambda) \to \hat{\beta}\left(\Pi^{exp(\lambda)}\right) \approx D(exp(1+\lambda), \Gamma), \tag{3.21}$$

where *D* is the extrapolation function, and Γ is the vector of parameters for the extrapolation function. Thus, based on the function (3.16), the estimator $\hat{\beta}_{nonpMC-SIMEX}$ is given by,

$$\hat{\beta}_{nonpMC-SIMEX} = D(exp(1+\lambda), \Gamma). \tag{3.22}$$

For example, if the best nonparametric extrapolation function is the quadratic extrapolation function, then $D(exp(1+\lambda), \Gamma) = \Gamma_0 + \Gamma_1(exp(1+\lambda))^{p_1} + \Gamma_2(exp(1+\lambda))^{p_2}, p_1 = 1, p_2 = 2$.Based on the FP method, there are 8 FP1 models and 36 FP2 models included in the method.

To choose the best power coefficient(s) in the nonparametric extrapolation function among the power coefficient(s) in the null, linear, FP1, and FP2 models, we use the cross-validation process. We separate data into 5 equally sized folds (also called K-fold cross-validation where K = 5) and consider one as test data, other 4 as training data. Then the training data are used to obtain the parameters Γ as a result of the *D* function. Thus, with Γ, *D* function, and $exp(\lambda)$ in the test data, we can obtain the predicted value of

$\hat{\beta}$ called $\hat{\beta}_{pred}$ as a result of $D(exp(\lambda), \Gamma)$. Let $\hat{\beta}_{test}$ denote the corresponding $\hat{\beta}$ to $exp(\lambda)$ in the test data. Then, the squared prediction error (SPE) is,

$$SPE = (\hat{\beta}_{pred} - \hat{\beta}_{test})^2. \tag{3.23}$$

Repeat the aforementioned steps until each fold used as the test data. Let $SPE_i, i = 1, \dots, K$ denote the squared prediction error for each test data. Let the mean squared prediction error (MPE) denote the average squared prediction error of all test data. In other words,

$$MPE = K^{-1} \sum_{i=1}^{K} (SPE_i), i = 1, \dots, K. \tag{3.24}$$

We choose MPE as an indication of the performance of each model. Let MPE(0) denote the MPE of the null model where $exp(\lambda)$ is not in the model. Let MPE(1) denote the MPE of the linear model; that is, $exp^p(\lambda) = exp(\lambda)$. Let MPE($p_1$) denote the minimum MPE of the 8 FP1 models, and MPE($p_1, p_2$) denote the minimum MPE of the 36 FP2 models. Thus, the best fractional polynomial model is the minimum of MPE(0), MPE(1), MPE($p_1$) and MPE($p_1, p_2$). We use the selected power coefficient(s) in the best fractional polynomial model in the nonparametric extrapolation function.

Based on the preceding results, the estimator $\Gamma_{ext}$ is obtained by least squares on $[exp(1 + \lambda), \hat{\beta}_{na}(\lambda_k)]_{k=0}^{m}$ with fitting a nonparametric model $D(exp(1 + \lambda), \Gamma)$. We estimate the estimator $\hat{\beta}_{nonpMC-SIMEX}$ as a result of the model $D(exp(1 + \lambda), \Gamma_{ext})$. That is,

$$\hat{\beta}_{nonpMC-SIMEX} = D(exp(1 + \lambda), \Gamma_{ext}). \tag{3.25}$$

*Variance estimation of estimator*

The estimation of the variance in the method is similar to the original MC-SIMEX method. In the process, the extrapolation function is replaced by the nonparametric extrapolation function, and $\lambda$ is replaced by $exp(\lambda)$ in extrapolation function. It can be otherwise expressed as,

$$V_{ST}(\lambda_k) = V_{na}(exp(\lambda_k)) - V_{sim}(exp(\lambda_k)). \tag{3.26}$$

and

$$V_{ST} = \lim_{\lambda \to -1} \left[ V_{na}\left(exp(\lambda)\right) - V_{sim}\left(exp(\lambda)\right) \right]. \tag{3.27}$$

We use three different extrapolation functions to extrapolate the variance estimation.

1. We use the quadratic extrapolation function to extrapolate the variance estimation, which is the same as the original MC-SIMEX method (Küchenhoff et al., 2006; Sevilimedu, 2017). Note that in this situation, we use $\lambda$ to estimate the variance. Hence $\Gamma_{ext.Q}$ is obtained as a result of the model $V_{ST}(\lambda_k) \sim \lambda_k + (\lambda_k)^2$. Let Q.VAR denote the variance estimation from the quadratic extrapolation function,

$$\text{Q.VAR} = \Gamma_{ext.Q.0} - \Gamma_{ext.Q.1} + \Gamma_{ext.Q.2}. \tag{3.28}$$

2. We use the extrapolation function for $\hat{\beta}_{nonpMC-SIMEX}$ as the extrapolation function for the variance estimation. That is,

$$\begin{aligned} VAR &= D\left(\exp(-1), \Gamma_{ext.fp}\right) \\ &= \Gamma_{ext.fp.0} + \Gamma_{ext.fp.1}(\exp(-1))^{p_1} + \Gamma_{ext.fp.2}(\exp(-1))^{p_2}, \end{aligned} \tag{3.29}$$

where $\Gamma_{ext.fp}$ is obtained as a result of the nonparametric extrapolation function. Note that if the nonparametric extrapolation function is the best of the FP1 models, there is no $p_2$ term in the function (3.29).

3. To choose the most appropriate curve for the variance data, we use the same procedure (FP method with cross-validation) as described in Chapter 3 for choosing a new nonparametric extrapolation function for the variance estimation. Let CV.VAR denote the variance estimation from the new nonparametric extrapolation function, which is called $D_{var}$ and where $\Gamma_{ext.cv}$ are the parameters from the $D_{var}$. Hence,

$$\begin{aligned} CV.VAR &= D_{var}(\exp(-1), \Gamma_{ext.cv}) \\ &= \Gamma_{ext.cv.0} + \Gamma_{ext.cv.1}(\exp(-1))^{p_1} + \Gamma_{ext.cv.2}(\exp(-1))^{p_2}. \end{aligned} \tag{3.30}$$

Note that if $D_{var}$ is the best of the FP1 models, there is only $p_1$ term in the function (3.30).

# CHAPTER 4

# SIMULATION STUDY

A simulation study is conducted to evaluate the performance of the nonparametric MC-SIMEX method in a logistic model with differential and nondifferential misclassification error on predictor. In section 4.1, we describe the simulation processes. In section 4.2, we describe the algorithms used to estimate parameters. Section 4.3 describes the results of the performance of the original and nonparametric MC-SIMEX methods, followed by the conclusions in section 4.4.

## An overview of simulation methods

We do the following steps to achieve the simulation work for differential misclassification error:

Step 1: Assign the values for regression parameters ($\beta_0$ and $\beta_1$). Define the number of iteration B, the sample size (n), the values of misclassification matrix ($\pi_{000}, \pi_{011}, \pi_{100}, \pi_{111}$) and the number of bootstrap simulation (M) for the differential misclassification error.

Step 2: Generate n random values for the true binary covariate *X* which follows Bernoulli distribution with the probability 0.5. [randomly select n samples for the true binary covariate X from Bernoulli distribution with probability of 0.5.]

Step 3: Generate true binary response *Y* which follows Bernoulli distribution according to the probability:

$$P(Y = 1) = 1/(1 + exp(-\beta_0 - \beta_X X)). \tag{4.1}$$

Step 4: Generate naive covariate *W* by the misclassification operation (see equation 3.13) with the *misclass* function in R and misclassification matrix for differential misclassification error:

$$\Pi_{diff} = \begin{bmatrix} \pi_{000} & \pi_{001} & 0 & 0 \\ \pi_{010} & \pi_{011} & 0 & 0 \\ 0 & 0 & \pi_{100} & \pi_{101} \\ 0 & 0 & \pi_{110} & \pi_{111} \end{bmatrix}. \tag{4.2}$$

Step 5: Fit the logistic model $y \sim x$ using *glm* procedure in R. Obtain the $\hat{\beta}_{true}$ and $\widehat{VAR}_{true}$ from the true model.

Step 6: Fit the logistic model $y \sim W$ using *glm* procedure in R. Obtain naive estimators $\hat{\beta}_{naive}$ and $\widehat{VAR}_{naive}$ for differential misclassification error from the naive model.

Step 7: Repeating step 4 to generate the differential misclassification $W^*$ at $\lambda$ = (0.5, 1, 1.5, 2) respectively, where the $\lambda$ is the power of the misclassification matrix: $\Pi^{\lambda}$.

Step 8: At each level of $\Pi^{\lambda}$, fit a logistic model with function $y \sim W^*$. Obtain the $\hat{\beta}_{MCsimex}$ and variance for each model at each $\lambda$ level.

Step 9: Run B times of iterations for steps 7-8 in each simulation.

Step 10: Use $\lambda$ to classify all the results of logistic models at each $\lambda$ level.

Step 11: Based on section 3.3.3, variance estimation is obtained.

Step 12: Build the models $\beta \sim \lambda$ and $V_{ST} \sim \lambda$ with the naive estimators ($\hat{\beta}_{naive}$ and $\widehat{VAR}_{naive}$) using the quadratic extrapolation function. Extrapolate the fitted models to the point on the Y-axis where $\lambda = -1$. The values on the Y-axis are the $\hat{\beta}_{MCsimex}$ and its estimated variance $\widehat{VAR}_{MCsimex}$. By the end of step 12, we finish the simulation steps of original MC-SIMEX method.

Step 13: Repeat steps 4-11 with $\lambda \in (0.01,2]$ with gap 0.02. Create the dataset called extrapolation data with average of B $\hat{\beta}_{MCsimex}$'s for each $\lambda$, $\hat{V}_{ST}$'s for each $\lambda$ and naive estimators.

Step 14: Using 5-fold cross-validation method be described in section 3.4.2 we reorganize the extrapolation data.

Step 15: Based on fractional polynomial method described in section 3.2 and fractional polynomial process steps described in section 3.4.2, extract the extrapolation function power coefficients for each simulation run.

Step 16: Using the extrapolation function from step 15, build the model $\beta \sim e^{\lambda}$ and $V_{ST} \sim e^{\lambda}$. Extrapolate the fitted model to the point on the Y-axis where $\lambda = -1$. The values on the Y-axis are the $\hat{\beta}_{nonp}$.

Step 17: Using the quadratic extrapolation function to estimate the variance of $\hat{\beta}_{nonp}$, named $\widehat{VAR}_{nonp.Q}$. Note that here the extrapolation function is $V_{ST} \sim \lambda$.

Step 18: Using the extrapolation function from step 15 to estimate the variance of $\hat{\beta}_{nonp}$, named $\widehat{VAR}_{nonp}$.

Step 19: [Estimate the power coefficients of extrapolation function for variance using fractional polynomial process from the cross-validation data set from step 14.] Using the cross-validation data sets from step 14, we do the fractional polynomial process to obtain the power coefficients of extrapolation function for variance.

Step 20: Based on the extrapolation function from step 19, estimate the variance of $\hat{\beta}_{nonp}$, named $\widehat{VAR}_{nonp.CV}$. By the end of step 20, we finish one bootstrap simulation run.

Step 21: Repeat steps 1-20 until a set of M Monte Carlo run are completed. We will compare the performance of new method in different setting.

Step 22: The final estimates are average of all Monte Carlo run of $\hat{\beta}_{true}$, $\hat{\beta}_{naive}$, $\hat{\beta}_{MCsimex}$ and $\hat{\beta}_{nonp}$ along with their corresponding mean, SE, RMSE, estimated variances , empirical variances and coverage.

For the nondifferential misclassification error, the difference of simulation work is following:

In step 1, we need to assign the $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ values for the nondifferential misclassification error.

In step 4, we generate $W$ by the misclassification operation with the misclassification matrix for nondifferential misclassification error:

$$\Pi_{non} = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}. \tag{4.3}$$

Other steps are similar for the nondifferential misclassification error simulation work.

Data simulation and estimation of parameters

In this section, we will compare the performance of true model, naive model, original and nonparametric MC-SIMEX method in different situations. We conduct M = 300 Monte Carlo run in each situation. We consider following parameter settings:

The sample size $n \in (200,500,1000)$;

The true values of $\beta$ : $(\beta_0, \beta_1) = (0,1)$ and $(\beta_0, \beta_1) = (0, -log2)$;

The differential misclassification matrices are $(\pi_{000}, \pi_{011}, \pi_{100}, \pi_{111}) = (0.9,0.7,0.7,0.8)$ and $(\pi_{000}, \pi_{011}, \pi_{100}, \pi_{111}) = (0.8,0.8,0.75,0.75)$;

and the nondifferential misclassification matrices are $(\pi_{00}, \pi_{11}) = (0.9,0.7)$ and $(\pi_{00}, \pi_{11}) = (0.8,0.8)$.

In other words, the first differential misclassification matrix setting is:

$$\Pi_{diff.1} = \begin{bmatrix} 0.9 & 0.3 & 0 & 0 \\ 0.1 & 0.7 & 0 & 0 \\ 0 & 0 & 0.7 & 0.2 \\ 0 & 0 & 0.3 & 0.8 \end{bmatrix}, \tag{4.4}$$

and the second differential misclassification matrix setting is:

$$\Pi_{diff.2} = \begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.75 & 0.25 \\ 0 & 0 & 0.25 & 0.75 \end{bmatrix}. \tag{4.5}$$

The first nondifferential misclassification matrix setting is:

$$\Pi_{non.1} = \begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix}, \tag{4.6}$$

and the second nondifferential misclassification matrix setting is:

$$\Pi_{non.2} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}. \tag{4.7}$$

We use number of bootstrap samples $B \in (50,100,300,500)$ for nonparametric MC-SIMEX method. A simple logistic regression model $Y \sim X$ is considered in this simulation study. The binary covariate, $X$, is generated from a Bernoulli distribution with the probability 0.5. The binary response variable Y follows Bernoulli distribution with the probability:

$$P(Y_i = 1) = 1/(1 + exp(-\beta_0 - \beta_1 X_i)). \tag{4.8}$$

For each simulation run, true estimator, naive estimator, MC-SIMEX estimator, nonparametric MC-SIMEX estimator and the corresponding mean, standard error (SE), estimated variances, empirical variances are obtained. The estimated values of $\hat{\beta}$ and their corresponding performance are obtained as average of results of each Monte Carlo run. They are defined as following:

$$\hat{\beta}_{true} = \frac{1}{M}\sum_{i=1}^{M} \hat{\beta}_{true_i},$$

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

$$\hat{\beta}_{MCsimex} = \frac{1}{M}\sum_{i=1}^{M} \hat{\beta}_{MCsimex_i},$$

$$\hat{\beta}_{nonp} = \frac{1}{M}\sum_{i=1}^{M} \hat{\beta}_{nonp_i},$$

$$\hat{\beta}_{nonp} = \hat{\beta}_{nonp.CV} = \hat{\beta}_{nonp.Q}.$$

The biases are defined as following:

$$\widehat{bias}_{true} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{true_i} - \beta_1),$$

$$\widehat{bias}_{naive} = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\beta}_{naive_i} - \beta_1\right),$$

$$\widehat{bias}_{MCsimex} = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\beta}_{MCsimex_i} - \beta_1\right),$$

$$\widehat{bias}_{nonp} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{nonp_i} - \beta_1),$$

$$\widehat{bias}_{nonp} = \widehat{bias}_{nonp.CV} = \widehat{bias}_{nonp.Q}.$$

The empirical variances (emp) are defined as following:

$$\widehat{emp}_{true} = \frac{1}{M}\sum_{i=1}^{M}\widehat{bias}^2_{true_i},$$

$$\widehat{emp}_{naive} = \frac{1}{M}\sum_{i=1}^{M}\widehat{bias}^2_{naive_i},$$

$$\widehat{emp}_{MCsimex} = \frac{1}{M}\sum_{i=1}^{M}\widehat{bias}^2_{MCsimex_i},$$

$$\widehat{emp}_{nonp} = \frac{1}{M}\sum_{i=1}^{M}\widehat{bias}^2_{nonp_i},$$

$$\widehat{emp}_{nonp} = \widehat{emp}_{nonp.CV} = \widehat{emp}_{nonp.Q}.$$

The RMSE are defined as following:

$$\widehat{RMSE}_{true} = sqrt\left(\widehat{emp}_{true} + \widehat{bias}^2_{true}\right),$$

$$\widehat{RMSE}_{naive} = sqrt\left(\widehat{emp}_{naive} + \widehat{bias}^2_{naive}\right),$$

$$\widehat{RMSE}_{MCsimex} = sqrt\left(\widehat{emp}_{MCsimex} + \widehat{bias}^2_{MCsimex}\right),$$

$$\widehat{RMSE}_{nonp} = sqrt\left(\widehat{emp}_{nonp} + \widehat{bias}^2_{nonp}\right),$$

$$\widehat{RMSE}_{nonp} = \widehat{RMSE}_{nonp.CV} = \widehat{RMSE}_{nonp}.Q.$$

The 95%CI for each Monte Carlo run is estimated using the following formula:

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

The coverage probability is assessed according to the percentage of the occurrences of 95%CI including the value of the true parameter.

Results of the performance of the methods

The simulation results table is presented as following table 1.

*Table 1: Form of results table.*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| True | . | . | . | . | . | . | . |
| Naïve | . | . | . | . | . | . | . |
| Mcsimex.Q | . | . | . | . | . | . | . |
| NONP | . | . | . | . | . | . | . |
| NONP.CV.Var | . | . | . | . | . | . | . |
| NONP.Q.Var | . | . | . | . | . | . | . |

In "Methods" column, the "True" indicate the true model (see section 3.1). The "Naive" stands for the naive model (see section 3.1). The "MCsimex.Q" denote the original MC-SIMEX method with quadratic extrapolation function. The "NONP" means the nonparametric MC-SIMEX method with same extrapolation function for $\hat{\beta}_1$ and variance. The "NONP.CV.Var" represent the nonparametric MC-SIMEX method using cross-validation process for $\hat{\beta}_1$ and variance separately. And the "NONP.Q.Var" implies the nonparametric MC-SIMEX method using quadratic extrapolation function. The "Mean" column presents the average value of $\hat{\beta}_1$ of all Monte Carlo run.

The performance of the nonparametric MC-SIMEX estimator is evaluated for different combinations of sample size, number of bootstrap samples, true $\beta$s, differential misclassification matrix.

*Results for differential misclassification error*

For differential misclassification error, tables 2 - 13 show that the nonparametric MC-SIMEX estimator consistently performs better than the original MC-SIMEX method with quadratic extrapolation function in all combinations. The magnitude of the bias associated with the nonparametric MC-SIMEX estimator is always lower than that of the original MC-SIMEX method with quadratic extrapolation function across all levels of combinations. With regard to the coverage probabilities, the nonparametric MC-SIMEX estimator is shown to perform satisfactorily and consistent with the original MC-SIMEX method with quadratic extrapolation function across all levels of combinations.

*Results for nondifferential misclassification error*

Tables 14 - 25 show that the nonparametric MC-SIMEX estimator consistently performs better than the original MC-SIMEX method with quadratic extrapolation function in all combinations with nondifferential misclassification error. Compare to the estimator of original MC-SIMEX method with quadratic extrapolation function, the nonparametric MC-SIMEX estimator has lower bias across all levels of combinations. With regard to the coverage probabilities, the nonparametric MC-SIMEX estimator is shown to perform satisfactorily and consistent with the original MC-SIMEX method with quadratic extrapolation function across all levels of combinations.

## Conclusion

The simulation work in this chapter shows that the original MC-SIMEX method only performs better in some situations. For example, when sample size is small (n = 200). In other words, the robustness of original MC-SIMEX method relies on small sample size. However, the nonparametric MC-SIMEX method proposed in this dissertation presents stronger robustness across all levels of parameter settings. Another proof of strong robustness is that compare to the original MC-SIMEX method, the biases of estimator in our approach are closer to the bias of true model. In the simulation results tables, the empirical

variance of nonparametric MC-SIMEX method always larger than the original MC-SIMEX method, since we applied the fractional polynomial process in the method. The number of bootstrap samples B is another influencing factor for our approach. When B is large enough ($B \geq 100$), the results of our method are always better than the original method. Hence, for the nonparametric MC-SIMEX method, B value should as large as possible. We proposed three ways to approximate the variance function in our method. The $\widehat{VAR}_{nonp}$ and $\widehat{VAR}_{nonp.CV}$ are reliable and valid estimators of variance, $\widehat{VAR}_{nonp}.Q$ didn't work very well in the simulation. In addition, with two kinds of misclassification error (nondifferential and differential), the nonparametric MC-SIMEX method works fine. Finally, some researchers reported that the increased bias in the MC-SIMEX estimates in the robustness test (Slate et al., 2009; Sevilimedu, 2017). For the nonparametric MC-SIMEX method, we haven't found this phenomenon.

*Table 2: Simulation results based on 300 simulations each with sample size = 1000 and Π_(diff.1): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.002 | 0.135 | 0.131 | 0.018 | 0.017 | 0.002 | 0.96 |
| Naive | 1.201 | 0.138 | 0.313 | 0.019 | 0.058 | 0.201 | 0.70 |
| McSimex.Q | 1.038 | 0.198 | 0.215 | 0.039 | 0.045 | 0.038 | 0.93 |
| NONP | 1.017 | 0.216 | 0.237 | 0.048 | 0.056 | 0.017 | 0.91 |
| NONP.CV.Var | 1.017 | 0.228 | 0.237 | 0.054 | 0.056 | 0.017 | 0.91 |
| NONP.Q.Var | 1.017 | 0.196 | 0.237 | 0.039 | 0.056 | 0.017 | 0.90 |
| B = 300 | | | | | | | |
| TRUE | 0.999 | 0.135 | 0.139 | 0.018 | 0.019 | -0.001 | 0.93 |
| Naive | 1.196 | 0.138 | 0.307 | 0.019 | 0.056 | 0.196 | 0.73 |
| McSimex.Q | 1.029 | 0.198 | 0.216 | 0.039 | 0.046 | 0.029 | 0.94 |
| NONP | 1.006 | 0.218 | 0.233 | 0.049 | 0.054 | 0.006 | 0.93 |
| NONP.CV.Var | 1.006 | 0.218 | 0.233 | 0.051 | 0.054 | 0.006 | 0.88 |
| NONP.Q.Var | 1.006 | 0.196 | 0.233 | 0.039 | 0.054 | 0.006 | 0.92 |
| B = 100 | | | | | | | |
| TRUE | 1.002 | 0.135 | 0.131 | 0.018 | 0.017 | 0.002 | 0.95 |
| Naive | 1.193 | 0.138 | 0.301 | 0.019 | 0.053 | 0.193 | 0.73 |
| McSimex.Q | 1.025 | 0.196 | 0.210 | 0.039 | 0.044 | 0.025 | 0.94 |
| NONP | 1.007 | 0.209 | 0.255 | 0.045 | 0.065 | 0.007 | 0.90 |
| NONP.CV.Var | 1.007 | 0.218 | 0.255 | 0.051 | 0.065 | 0.007 | 0.90 |
| NONP.Q.Var | 1.007 | 0.195 | 0.255 | 0.038 | 0.065 | 0.007 | 0.88 |
| B = 50 | | | | | | | |
| TRUE | 0.997 | 0.135 | 0.131 | 0.018 | 0.017 | -0.003 | 0.97 |
| Naive | 1.192 | 0.138 | 0.302 | 0.019 | 0.054 | 0.192 | 0.72 |
| McSimex.Q | 1.022 | 0.197 | 0.219 | 0.039 | 0.048 | 0.022 | 0.92 |
| NONP | 1.020 | 0.206 | 0.318 | 0.043 | 0.101 | 0.020 | 0.86 |
| NONP.CV.Var | 1.020 | 0.225 | 0.318 | 0.053 | 0.101 | 0.020 | 0.87 |
| NONP.Q.Var | 1.020 | 0.195 | 0.318 | 0.038 | 0.101 | 0.020 | 0.84 |

*Table 3: Simulation results based on 300 simulations each with sample size = 500 and Π_(diff.1): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 0.998 | 0.191 | 0.186 | 0.037 | 0.034 | -0.002 | 0.96 |
| Naive | 1.197 | 0.196 | 0.338 | 0.038 | 0.075 | 0.197 | 0.82 |
| McSimex.Q | 1.029 | 0.282 | 0.309 | 0.079 | 0.095 | 0.029 | 0.95 |
| NONP | 0.999 | 0.307 | 0.343 | 0.099 | 0.117 | -0.001 | 0.90 |
| NONP.CV.Var | 0.999 | 0.314 | 0.343 | 0.105 | 0.117 | -0.001 | 0.90 |
| NONP.Q.Var | 0.999 | 0.278 | 0.343 | 0.078 | 0.117 | -0.001 | 0.88 |
| B = 300 | | | | | | | |
| TRUE | 1.005 | 0.192 | 0.178 | 0.037 | 0.032 | 0.005 | 0.97 |
| Naive | 1.194 | 0.195 | 0.346 | 0.038 | 0.082 | 0.194 | 0.82 |
| McSimex.Q | 1.025 | 0.281 | 0.336 | 0.079 | 0.112 | 0.025 | 0.91 |
| NONP | 1.011 | 0.302 | 0.390 | 0.096 | 0.152 | 0.011 | 0.87 |
| NONP.CV.Var | 1.011 | 0.317 | 0.390 | 0.106 | 0.152 | 0.011 | 0.87 |
| NONP.Q.Var | 1.011 | 0.278 | 0.390 | 0.077 | 0.152 | 0.011 | 0.84 |
| B = 100 | | | | | | | |
| TRUE | 1.008 | 0.192 | 0.202 | 0.037 | 0.041 | 0.008 | 0.95 |
| Naive | 1.209 | 0.196 | 0.359 | 0.038 | 0.085 | 0.209 | 0.82 |
| McSimex.Q | 1.050 | 0.280 | 0.333 | 0.079 | 0.108 | 0.050 | 0.97 |
| NONP | 1.043 | 0.288 | 0.388 | 0.086 | 0.149 | 0.043 | 0.84 |
| NONP.CV.Var | 1.043 | 0.309 | 0.388 | 0.099 | 0.149 | 0.043 | 0.88 |
| NONP.Q.Var | 1.043 | 0.278 | 0.388 | 0.078 | 0.149 | 0.043 | 0.86 |
| B = 50 | | | | | | | |
| TRUE | 0.991 | 0.191 | 0.198 | 0.037 | 0.039 | -0.009 | 0.95 |
| Naive | 1.186 | 0.195 | 0.331 | 0.038 | 0.075 | 0.186 | 0.84 |
| McSimex.Q | 1.013 | 0.278 | 0.324 | 0.078 | 0.105 | 0.013 | 0.90 |
| NONP | 1.014 | 0.294 | 0.431 | 0.091 | 0.185 | 0.014 | 0.83 |
| NONP.CV.Var | 1.014 | 0.331 | 0.431 | 0.111 | 0.185 | 0.014 | 0.87 |
| NONP.Q.Var | 1.014 | 0.276 | 0.431 | 0.076 | 0.185 | 0.014 | 0.80 |

*Table 4: Simulation results based on 300 simulations each with sample size = 200 and Π_(diff.1): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.025 | 0.305 | 0.288 | 0.093 | 0.082 | 0.025 | 0.96 |
| Naive | 1.208 | 0.311 | 0.431 | 0.097 | 0.143 | 0.208 | 0.90 |
| McSimex.Q | 1.037 | 0.451 | 0.500 | 0.204 | 0.249 | 0.037 | 0.92 |
| NONP | 1.021 | 0.489 | 0.557 | 0.251 | 0.310 | 0.021 | 0.90 |
| NONP.CV.Var | 1.021 | 0.518 | 0.557 | 0.281 | 0.310 | 0.021 | 0.91 |
| NONP.Q.Var | 1.021 | 0.444 | 0.557 | 0.198 | 0.310 | 0.021 | 0.89 |
| B = 300 | | | | | | | |
| TRUE | 1.050 | 0.306 | 0.337 | 0.094 | 0.111 | 0.050 | 0.92 |
| Naive | 1.214 | 0.312 | 0.435 | 0.098 | 0.143 | 0.214 | 0.90 |
| McSimex.Q | 1.044 | 0.451 | 0.500 | 0.204 | 0.248 | 0.044 | 0.92 |
| NONP | 1.025 | 0.488 | 0.552 | 0.252 | 0.304 | 0.025 | 0.89 |
| NONP.CV.Var | 1.025 | 0.498 | 0.552 | 0.264 | 0.304 | 0.025 | 0.89 |
| NONP.Q.Var | 1.025 | 0.446 | 0.552 | 0.199 | 0.304 | 0.025 | 0.89 |
| B = 100 | | | | | | | |
| TRUE | 1.026 | 0.306 | 0.328 | 0.094 | 0.107 | 0.026 | 0.94 |
| Naive | 1.242 | 0.314 | 0.473 | 0.099 | 0.165 | 0.242 | 0.90 |
| McSimex.Q | 1.089 | 0.453 | 0.539 | 0.207 | 0.282 | 0.089 | 0.92 |
| NONP | 1.074 | 0.482 | 0.629 | 0.249 | 0.390 | 0.074 | 0.84 |
| NONP.CV.Var | 1.074 | 0.508 | 0.629 | 0.271 | 0.390 | 0.074 | 0.88 |
| NONP.Q.Var | 1.074 | 0.447 | 0.629 | 0.201 | 0.390 | 0.074 | 0.86 |
| B = 50 | | | | | | | |
| TRUE | 0.998 | 0.304 | 0.317 | 0.093 | 0.100 | -0.002 | 0.95 |
| Naive | 1.197 | 0.311 | 0.425 | 0.097 | 0.142 | 0.197 | 0.91 |
| McSimex.Q | 1.023 | 0.441 | 0.516 | 0.198 | 0.266 | 0.023 | 0.92 |
| NONP | 1.037 | 0.454 | 0.669 | 0.218 | 0.446 | 0.037 | 0.79 |
| NONP.CV.Var | 1.037 | 0.513 | 0.669 | 0.270 | 0.446 | 0.037 | 0.88 |
| NONP.Q.Var | 1.037 | 0.441 | 0.669 | 0.195 | 0.446 | 0.037 | 0.83 |

*Table 5: Simulation results based on 300 simulations each with sample size = 1000 and Π_(diff.1): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.715 | 0.131 | 0.132 | 0.017 | 0.017 | -0.022 | 0.95 |
| Naive | 0.221 | 0.129 | 1.299 | 0.017 | 0.852 | 0.914 | 0.00 |
| McSimex.Q | -0.519 | 0.188 | 0.325 | 0.035 | 0.075 | 0.174 | 0.84 |
| NONP | -0.702 | 0.228 | 0.252 | 0.054 | 0.063 | -0.008 | 0.89 |
| NONP.CV.Var | -0.702 | 0.224 | 0.252 | 0.053 | 0.063 | -0.008 | 0.90 |
| NONP.Q.Var | -0.702 | 0.185 | 0.252 | 0.034 | 0.063 | -0.008 | 0.83 |
| B = 300 | | | | | | | |
| TRUE | -0.689 | 0.131 | 0.135 | 0.017 | 0.018 | 0.004 | 0.95 |
| Naive | 0.239 | 0.129 | 1.326 | 0.017 | 0.888 | 0.933 | 0.00 |
| McSimex.Q | -0.488 | 0.188 | 0.365 | 0.036 | 0.091 | 0.205 | 0.77 |
| NONP | -0.668 | 0.226 | 0.274 | 0.053 | 0.074 | 0.026 | 0.85 |
| NONP.CV.Var | -0.668 | 0.215 | 0.274 | 0.050 | 0.074 | 0.026 | 0.82 |
| NONP.Q.Var | -0.668 | 0.185 | 0.274 | 0.034 | 0.074 | 0.026 | 0.80 |
| B = 100 | | | | | | | |
| TRUE | -0.700 | 0.131 | 0.138 | 0.017 | 0.019 | -0.006 | 0.94 |
| Naive | 0.224 | 0.129 | 1.304 | 0.017 | 0.859 | 0.917 | 0.00 |
| McSimex.Q | -0.511 | 0.187 | 0.338 | 0.035 | 0.081 | 0.182 | 0.77 |
| NONP | -0.692 | 0.216 | 0.272 | 0.048 | 0.074 | 0.001 | 0.86 |
| NONP.CV.Var | -0.692 | 0.215 | 0.272 | 0.049 | 0.074 | 0.001 | 0.84 |
| NONP.Q.Var | -0.692 | 0.185 | 0.272 | 0.034 | 0.074 | 0.001 | 0.81 |
| B = 50 | | | | | | | |
| TRUE | -0.691 | 0.131 | 0.128 | 0.017 | 0.016 | 0.002 | 0.94 |
| Naive | 0.235 | 0.129 | 1.319 | 0.017 | 0.877 | 0.928 | 0.00 |
| McSimex.Q | -0.496 | 0.186 | 0.349 | 0.035 | 0.083 | 0.197 | 0.81 |
| NONP | -0.651 | 0.206 | 0.270 | 0.044 | 0.071 | 0.042 | 0.88 |
| NONP.CV.Var | -0.651 | 0.219 | 0.270 | 0.050 | 0.071 | 0.042 | 0.88 |
| NONP.Q.Var | -0.651 | 0.185 | 0.270 | 0.034 | 0.071 | 0.042 | 0.85 |

*Table 6: Simulation results based on 300 simulations each with sample size = 500 and Π_(diff.1): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.704 | 0.185 | 0.200 | 0.034 | 0.040 | -0.010 | 0.92 |
| Naive | 0.223 | 0.182 | 1.309 | 0.033 | 0.873 | 0.916 | 0.00 |
| McSimex.Q | -0.517 | 0.266 | 0.392 | 0.071 | 0.123 | 0.176 | 0.85 |
| NONP | -0.708 | 0.316 | 0.362 | 0.103 | 0.131 | -0.015 | 0.89 |
| NONP.CV.Var | -0.708 | 0.325 | 0.362 | 0.109 | 0.131 | -0.015 | 0.92 |
| NONP.Q.Var | -0.708 | 0.262 | 0.362 | 0.069 | 0.131 | -0.015 | 0.87 |
| B = 300 | | | | | | | |
| TRUE | -0.688 | 0.185 | 0.188 | 0.034 | 0.035 | 0.005 | 0.96 |
| Naive | 0.227 | 0.183 | 1.315 | 0.033 | 0.881 | 0.920 | 0.00 |
| McSimex.Q | -0.510 | 0.270 | 0.403 | 0.071 | 0.129 | 0.183 | 0.84 |
| NONP | -0.697 | 0.310 | 0.368 | 0.099 | 0.136 | -0.004 | 0.90 |
| NONP.CV.Var | -0.697 | 0.317 | 0.368 | 0.106 | 0.136 | -0.004 | 0.92 |
| NONP.Q.Var | -0.697 | 0.263 | 0.368 | 0.069 | 0.136 | -0.004 | 0.85 |
| B = 100 | | | | | | | |
| TRUE | -0.706 | 0.185 | 0.184 | 0.034 | 0.034 | -0.013 | 0.94 |
| Naive | 0.224 | 0.183 | 1.308 | 0.033 | 0.871 | 0.917 | 0.00 |
| McSimex.Q | -0.514 | 0.265 | 0.382 | 0.071 | 0.114 | 0.179 | 0.89 |
| NONP | -0.689 | 0.296 | 0.347 | 0.091 | 0.121 | 0.004 | 0.89 |
| NONP.CV.Var | -0.689 | 0.303 | 0.347 | 0.096 | 0.121 | 0.004 | 0.87 |
| NONP.Q.Var | -0.689 | 0.262 | 0.347 | 0.068 | 0.121 | 0.004 | 0.86 |
| B = 50 | | | | | | | |
| TRUE | -0.689 | 0.185 | 0.180 | 0.034 | 0.032 | 0.004 | 0.96 |
| Naive | 0.227 | 0.183 | 1.313 | 0.033 | 0.879 | 0.920 | 0.00 |
| McSimex.Q | -0.514 | 0.263 | 0.396 | 0.069 | 0.125 | 0.179 | 0.84 |
| NONP | -0.663 | 0.289 | 0.398 | 0.089 | 0.158 | 0.030 | 0.84 |
| NONP.CV.Var | -0.663 | 0.314 | 0.398 | 0.104 | 0.158 | 0.030 | 0.87 |
| NONP.Q.Var | -0.663 | 0.261 | 0.398 | 0.068 | 0.158 | 0.030 | 0.81 |

*Table 7: Simulation results based on 300 simulations each with sample size = 200 and Π_(diff.1): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.704 | 0.294 | 0.292 | 0.086 | 0.085 | -0.011 | 0.95 |
| Naive | 0.240 | 0.290 | 1.351 | 0.084 | 0.953 | 0.933 | 0.10 |
| McSimex.Q | -0.495 | 0.423 | 0.553 | 0.179 | 0.267 | 0.199 | 0.88 |
| NONP | -0.664 | 0.486 | 0.575 | 0.245 | 0.329 | 0.029 | 0.90 |
| NONP.CV.Var | -0.664 | 0.498 | 0.575 | 0.262 | 0.329 | 0.029 | 0.88 |
| NONP.Q.Var | -0.664 | 0.417 | 0.575 | 0.174 | 0.329 | 0.029 | 0.86 |
| B = 300 | | | | | | | |
| TRUE | -0.704 | 0.294 | 0.301 | 0.087 | 0.091 | -0.011 | 0.93 |
| Naive | 0.247 | 0.290 | 1.362 | 0.084 | 0.970 | 0.940 | 0.13 |
| McSimex.Q | -0.481 | 0.424 | 0.570 | 0.180 | 0.280 | 0.213 | 0.90 |
| NONP | -0.650 | 0.484 | 0.599 | 0.242 | 0.357 | 0.043 | 0.92 |
| NONP.CV.Var | -0.650 | 0.504 | 0.599 | 0.270 | 0.357 | 0.043 | 0.90 |
| NONP.Q.Var | -0.650 | 0.418 | 0.599 | 0.175 | 0.357 | 0.043 | 0.85 |
| B = 100 | | | | | | | |
| TRUE | -0.696 | 0.294 | 0.274 | 0.086 | 0.075 | -0.003 | 0.96 |
| Naive | 0.228 | 0.290 | 1.333 | 0.084 | 0.929 | 0.921 | 0.09 |
| McSimex.Q | -0.511 | 0.422 | 0.537 | 0.179 | 0.255 | 0.182 | 0.92 |
| NONP | -0.665 | 0.457 | 0.579 | 0.216 | 0.334 | 0.029 | 0.92 |
| NONP.CV.Var | -0.665 | 0.488 | 0.579 | 0.255 | 0.334 | 0.029 | 0.90 |
| NONP.Q.Var | -0.665 | 0.416 | 0.579 | 0.173 | 0.334 | 0.029 | 0.86 |
| B = 50 | | | | | | | |
| TRUE | -0.673 | 0.294 | 0.273 | 0.086 | 0.074 | 0.020 | 0.98 |
| Naive | 0.236 | 0.290 | 1.346 | 0.084 | 0.947 | 0.929 | 0.12 |
| McSimex.Q | -0.506 | 0.418 | 0.548 | 0.176 | 0.265 | 0.187 | 0.88 |
| NONP | -0.615 | 0.456 | 0.643 | 0.217 | 0.408 | 0.078 | 0.86 |
| NONP.CV.Var | -0.615 | 0.492 | 0.643 | 0.254 | 0.408 | 0.078 | 0.84 |
| NONP.Q.Var | -0.615 | 0.416 | 0.643 | 0.174 | 0.408 | 0.078 | 0.79 |

*Table 8: Simulation results based on 300 simulations each with sample size = 1000 and Π_(diff.2): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.003 | 0.135 | 0.149 | 0.018 | 0.022 | 0.003 | 0.90 |
| Naive | 0.564 | 0.132 | 0.633 | 0.017 | 0.210 | -0.437 | 0.11 |
| McSimex.Q | 0.917 | 0.194 | 0.259 | 0.038 | 0.060 | -0.083 | 0.87 |
| NONP | 1.030 | 0.234 | 0.276 | 0.056 | 0.075 | 0.030 | 0.93 |
| NONP.CV.Var | 1.030 | 0.236 | 0.276 | 0.058 | 0.075 | 0.030 | 0.93 |
| NONP.Q.Var | 1.030 | 0.190 | 0.276 | 0.036 | 0.075 | 0.030 | 0.83 |
| B = 300 | | | | | | | |
| TRUE | 0.995 | 0.135 | 0.136 | 0.018 | 0.018 | -0.005 | 0.97 |
| Naive | 0.536 | 0.132 | 0.668 | 0.017 | 0.231 | -0.465 | 0.04 |
| McSimex.Q | 0.869 | 0.193 | 0.274 | 0.037 | 0.058 | -0.131 | 0.88 |
| NONP | 0.980 | 0.238 | 0.246 | 0.058 | 0.060 | -0.020 | 0.93 |
| NONP.CV.Var | 0.980 | 0.225 | 0.246 | 0.054 | 0.060 | -0.020 | 0.89 |
| NONP.Q.Var | 0.980 | 0.191 | 0.246 | 0.036 | 0.060 | -0.020 | 0.84 |
| B = 100 | | | | | | | |
| TRUE | 1.008 | 0.135 | 0.139 | 0.018 | 0.019 | 0.008 | 0.96 |
| Naive | 0.549 | 0.132 | 0.650 | 0.017 | 0.219 | -0.451 | 0.06 |
| McSimex.Q | 0.893 | 0.194 | 0.256 | 0.038 | 0.054 | -0.107 | 0.91 |
| NONP | 0.991 | 0.231 | 0.277 | 0.055 | 0.077 | -0.009 | 0.89 |
| NONP.CV.Var | 0.991 | 0.220 | 0.277 | 0.051 | 0.077 | -0.009 | 0.86 |
| NONP.Q.Var | 0.991 | 0.190 | 0.277 | 0.036 | 0.077 | -0.009 | 0.83 |
| B = 50 | | | | | | | |
| TRUE | 1.024 | 0.135 | 0.146 | 0.018 | 0.021 | 0.024 | 0.95 |
| Naive | 0.565 | 0.132 | 0.630 | 0.017 | 0.207 | -0.435 | 0.09 |
| McSimex.Q | 0.919 | 0.192 | 0.249 | 0.037 | 0.056 | -0.081 | 0.87 |
| NONP | 1.001 | 0.225 | 0.312 | 0.052 | 0.097 | 0.001 | 0.85 |
| NONP.CV.Var | 1.001 | 0.227 | 0.312 | 0.055 | 0.097 | 0.001 | 0.83 |
| NONP.Q.Var | 1.001 | 0.190 | 0.312 | 0.036 | 0.097 | 0.001 | 0.79 |

*Table 9: Simulation results based on 300 simulations each with sample size = 500 and Π_(diff.2): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.012 | 0.192 | 0.190 | 0.037 | 0.036 | 0.012 | 0.96 |
| Naive | 0.569 | 0.187 | 0.640 | 0.035 | 0.223 | -0.431 | 0.36 |
| McSimex.Q | 0.928 | 0.275 | 0.339 | 0.076 | 0.110 | -0.072 | 0.90 |
| NONP | 1.041 | 0.338 | 0.392 | 0.117 | 0.152 | 0.041 | 0.90 |
| NONP.CV.Var | 1.041 | 0.334 | 0.392 | 0.116 | 0.152 | 0.041 | 0.88 |
| NONP.Q.Var | 1.041 | 0.271 | 0.392 | 0.073 | 0.152 | 0.041 | 0.81 |
| B = 300 | | | | | | | |
| TRUE | 1.002 | 0.191 | 0.190 | 0.037 | 0.036 | 0.002 | 0.94 |
| Naive | 0.546 | 0.187 | 0.668 | 0.035 | 0.241 | -0.454 | 0.31 |
| McSimex.Q | 0.887 | 0.274 | 0.347 | 0.075 | 0.107 | -0.113 | 0.90 |
| NONP | 0.989 | 0.327 | 0.383 | 0.110 | 0.147 | -0.011 | 0.90 |
| NONP.CV.Var | 0.989 | 0.330 | 0.383 | 0.115 | 0.147 | -0.011 | 0.89 |
| NONP.Q.Var | 0.989 | 0.270 | 0.383 | 0.073 | 0.147 | -0.011 | 0.85 |
| B = 100 | | | | | | | |
| TRUE | 1.004 | 0.191 | 0.188 | 0.037 | 0.035 | 0.004 | 0.95 |
| Naive | 0.545 | 0.186 | 0.670 | 0.035 | 0.242 | -0.455 | 0.31 |
| McSimex.Q | 0.888 | 0.273 | 0.344 | 0.075 | 0.106 | -0.112 | 0.90 |
| NONP | 0.955 | 0.314 | 0.400 | 0.102 | 0.158 | -0.045 | 0.87 |
| NONP.CV.Var | 0.955 | 0.314 | 0.400 | 0.104 | 0.158 | -0.045 | 0.85 |
| NONP.Q.Var | 0.955 | 0.269 | 0.400 | 0.072 | 0.158 | -0.045 | 0.81 |
| B = 50 | | | | | | | |
| TRUE | 1.000 | 0.191 | 0.194 | 0.037 | 0.037 | 0.000 | 0.96 |
| Naive | 0.548 | 0.187 | 0.667 | 0.035 | 0.240 | -0.452 | 0.34 |
| McSimex.Q | 0.900 | 0.271 | 0.348 | 0.074 | 0.111 | -0.100 | 0.91 |
| NONP | 0.960 | 0.302 | 0.405 | 0.095 | 0.162 | 0.040 | 0.87 |
| NONP.CV.Var | 0.960 | 0.326 | 0.405 | 0.111 | 0.162 | 0.040 | 0.88 |
| NONP.Q.Var | 0.960 | 0.268 | 0.405 | 0.072 | 0.162 | 0.040 | 0.82 |

*Table 10: Simulation results based on 300 simulations each with sample size = 200 and Π_(diff.2): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 0.97 | 0.304 | 0.311 | 0.093 | 0.096 | -0.030 | 0.95 |
| Naive | 0.550 | 0.297 | 0.711 | 0.088 | 0.303 | -0.450 | 0.66 |
| McSimex.Q | 0.889 | 0.438 | 0.543 | 0.192 | 0.283 | -0.111 | 0.89 |
| NONP | 0.995 | 0.506 | 0.608 | 0.266 | 0.370 | -0.005 | 0.88 |
| NONP.CV.Var | 0.995 | 0.536 | 0.608 | 0.301 | 0.370 | -0.005 | 0.88 |
| NONP.Q.Var | 0.995 | 0.430 | 0.608 | 0.185 | 0.370 | -0.005 | 0.84 |
| B = 300 | | | | | | | |
| TRUE | 0.993 | 0.305 | 0.320 | 0.093 | 0.102 | -0.007 | 0.94 |
| Naive | 0.575 | 0.297 | 0.670 | 0.088 | 0.268 | -0.425 | 0.72 |
| McSimex.Q | 0.938 | 0.438 | 0.499 | 0.192 | 0.245 | -0.062 | 0.93 |
| NONP | 1.035 | 0.504 | 0.590 | 0.263 | 0.347 | 0.035 | 0.93 |
| NONP.CV.Var | 1.035 | 0.515 | 0.590 | 0.281 | 0.347 | 0.035 | 0.90 |
| NONP.Q.Var | 1.035 | 0.431 | 0.590 | 0.186 | 0.347 | 0.035 | 0.87 |
| B = 100 | | | | | | | |
| TRUE | 0.978 | 0.304 | 0.322 | 0.093 | 0.103 | -0.022 | 0.94 |
| Naive | 0.544 | 0.297 | 0.715 | 0.088 | 0.303 | -0.456 | 0.67 |
| McSimex.Q | 0.888 | 0.435 | 0.529 | 0.190 | 0.268 | -0.112 | 0.88 |
| NONP | 0.924 | 0.483 | 0.670 | 0.245 | 0.444 | -0.076 | 0.85 |
| NONP.CV.Var | 0.924 | 0.516 | 0.670 | 0.281 | 0.444 | -0.076 | 0.88 |
| NONP.Q.Var | 0.924 | 0.429 | 0.670 | 0.184 | 0.444 | -0.076 | 0.82 |
| B = 50 | | | | | | | |
| TRUE | 1.038 | 0.304 | 0.309 | 0.093 | 0.094 | 0.038 | 0.93 |
| Naive | 0.567 | 0.297 | 0.684 | 0.088 | 0.281 | -0.433 | 0.68 |
| McSimex.Q | 0.921 | 0.432 | 0.524 | 0.189 | 0.269 | -0.079 | 0.90 |
| NONP | 0.958 | 0.460 | 0.657 | 0.222 | 0.429 | -0.042 | 0.86 |
| NONP.CV.Var | 0.958 | 0.511 | 0.657 | 0.275 | 0.429 | -0.042 | 0.85 |
| NONP.Q.Var | 0.958 | 0.427 | 0.657 | 0.183 | 0.429 | -0.042 | 0.84 |

*Table 11: Simulation results based on 300 simulations each with sample size = 1000 and Π_(diff.2): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.69 | 0.130 | 0.136 | 0.017 | 0.019 | 0.004 | 0.94 |
| Naive | -0.375 | 0.129 | 0.470 | 0.017 | 0.120 | 0.318 | 0.32 |
| McSimex.Q | -0.623 | 0.190 | 0.251 | 0.036 | 0.058 | 0.071 | 0.88 |
| NONP | -0.701 | 0.230 | 0.266 | 0.054 | 0.070 | -0.008 | 0.91 |
| NONP.CV.Var | -0.701 | 0.237 | 0.266 | 0.058 | 0.070 | -0.008 | 0.90 |
| NONP.Q.Var | -0.701 | 0.188 | 0.266 | 0.035 | 0.070 | -0.008 | 0.85 |
| B = 300 | | | | | | | |
| TRUE | -0.710 | 0.131 | 0.123 | 0.017 | 0.015 | -0.017 | 0.96 |
| Naive | -0.378 | 0.129 | 0.464 | 0.017 | 0.116 | 0.315 | 0.33 |
| McSimex.Q | -0.626 | 0.191 | 0.239 | 0.037 | 0.053 | 0.067 | 0.91 |
| NONP | -0.709 | 0.231 | 0.270 | 0.055 | 0.072 | -0.016 | 0.87 |
| NONP.CV.Var | -0.709 | 0.232 | 0.270 | 0.056 | 0.072 | -0.016 | 0.86 |
| NONP.Q.Var | -0.709 | 0.188 | 0.270 | 0.035 | 0.072 | -0.016 | 0.84 |
| B = 100 | | | | | | | |
| TRUE | -0.707 | 0.131 | 0.128 | 0.017 | 0.016 | -0.014 | 0.96 |
| Naive | -0.388 | 0.129 | 0.449 | 0.017 | 0.108 | 0.305 | 0.32 |
| McSimex.Q | -0.642 | 0.189 | 0.221 | 0.036 | 0.046 | 0.051 | 0.91 |
| NONP | -0.705 | 0.227 | 0.271 | 0.053 | 0.073 | -0.012 | 0.90 |
| NONP.CV.Var | -0.705 | 0.226 | 0.271 | 0.054 | 0.073 | -0.012 | 0.85 |
| NONP.Q.Var | -0.705 | 0.187 | 0.271 | 0.035 | 0.073 | -0.012 | 0.80 |
| B = 50 | | | | | | | |
| TRUE | -0.705 | 0.131 | 0.134 | 0.017 | 0.018 | -0.011 | 0.95 |
| Naive | -0.374 | 0.129 | 0.470 | 0.017 | 0.119 | 0.320 | 0.30 |
| McSimex.Q | -0.621 | 0.190 | 0.245 | 0.036 | 0.055 | 0.072 | 0.85 |
| NONP | -0.667 | 0.217 | 0.313 | 0.049 | 0.097 | 0.026 | 0.82 |
| NONP.CV.Var | -0.667 | 0.225 | 0.313 | 0.053 | 0.097 | 0.026 | 0.84 |
| NONP.Q.Var | -0.667 | 0.187 | 0.313 | 0.035 | 0.097 | 0.026 | 0.78 |

*Table 12: Simulation results based on 300 simulations each with sample size = 500 and Π_(diff.2): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.694 | 0.185 | 0.189 | 0.034 | 0.036 | -0.001 | 0.94 |
| Naive | -0.369 | 0.183 | 0.499 | 0.033 | 0.144 | 0.325 | 0.58 |
| McSimex.Q | -0.612 | 0.271 | 0.348 | 0.073 | 0.115 | 0.081 | 0.90 |
| NONP | -0.683 | 0.317 | 0.393 | 0.104 | 0.154 | 0.011 | 0.86 |
| NONP.CV.Var | -0.683 | 0.334 | 0.393 | 0.115 | 0.154 | 0.011 | 0.88 |
| NONP.Q.Var | -0.683 | 0.266 | 0.393 | 0.071 | 0.154 | 0.011 | 0.82 |
| B = 300 | | | | | | | |
| TRUE | -0.693 | 0.185 | 0.183 | 0.034 | 0.034 | 0.000 | 0.94 |
| Naive | -0.378 | 0.183 | 0.482 | 0.033 | 0.133 | 0.316 | 0.55 |
| McSimex.Q | -0.626 | 0.271 | 0.324 | 0.073 | 0.100 | 0.067 | 0.90 |
| NONP | -0.696 | 0.324 | 0.372 | 0.109 | 0.138 | -0.003 | 0.90 |
| NONP.CV.Var | -0.696 | 0.326 | 0.372 | 0.112 | 0.138 | -0.003 | 0.89 |
| NONP.Q.Var | -0.696 | 0.266 | 0.372 | 0.071 | 0.138 | -0.003 | 0.86 |
| B = 100 | | | | | | | |
| TRUE | -0.687 | 0.185 | 0.177 | 0.034 | 0.031 | 0.006 | 0.95 |
| Naive | -0.370 | 0.183 | 0.493 | 0.033 | 0.138 | 0.323 | 0.58 |
| McSimex.Q | -0.615 | 0.270 | 0.327 | 0.073 | 0.101 | 0.078 | 0.90 |
| NONP | -0.654 | 0.308 | 0.394 | 0.100 | 0.154 | 0.039 | 0.88 |
| NONP.CV.Var | -0.654 | 0.318 | 0.394 | 0.107 | 0.154 | 0.039 | 0.88 |
| NONP.Q.Var | -0.654 | 0.265 | 0.394 | 0.070 | 0.154 | 0.039 | 0.82 |
| B = 50 | | | | | | | |
| TRUE | -0.685 | 0.185 | 0.188 | 0.034 | 0.035 | 0.008 | 0.95 |
| Naive | -0.377 | 0.183 | 0.482 | 0.033 | 0.132 | 0.316 | 0.61 |
| McSimex.Q | -0.629 | 0.270 | 0.329 | 0.074 | 0.104 | 0.064 | 0.89 |
| NONP | -0.645 | 0.299 | 0.474 | 0.094 | 0.223 | 0.048 | 0.84 |
| NONP.CV.Var | -0.645 | 0.321 | 0.474 | 0.108 | 0.223 | 0.048 | 0.83 |
| NONP.Q.Var | -0.645 | 0.266 | 0.474 | 0.071 | 0.223 | 0.048 | 0.78 |

*Table 13: Simulation results based on 300 simulations each with sample size = 200 and Π_(diff.2): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.721 | 0.295 | 0.293 | 0.087 | 0.085 | -0.028 | 0.95 |
| Naive | -0.357 | 0.290 | 0.555 | 0.084 | 0.196 | 0.336 | 0.77 |
| McSimex.Q | -0.594 | 0.431 | 0.501 | 0.186 | 0.241 | 0.099 | 0.92 |
| NONP | -0.657 | 0.496 | 0.565 | 0.259 | 0.318 | 0.036 | 0.88 |
| NONP.CV.Var | -0.657 | 0.532 | 0.565 | 0.295 | 0.318 | 0.036 | 0.93 |
| NONP.Q.Var | -0.657 | 0.423 | 0.565 | 0.179 | 0.318 | 0.036 | 0.85 |
| B = 300 | | | | | | | |
| TRUE | -0.706 | 0.294 | 0.292 | 0.087 | 0.085 | -0.013 | 0.95 |
| Naive | -0.348 | 0.290 | 0.567 | 0.084 | 0.202 | 0.345 | 0.78 |
| McSimex.Q | -0.577 | 0.431 | 0.507 | 0.186 | 0.243 | 0.116 | 0.91 |
| NONP | -0.643 | 0.483 | 0.571 | 0.247 | 0.323 | 0.051 | 0.88 |
| NONP.CV.Var | -0.643 | 0.524 | 0.571 | 0.287 | 0.323 | 0.051 | 0.91 |
| NONP.Q.Var | -0.643 | 0.424 | 0.571 | 0.179 | 0.323 | 0.051 | 0.85 |
| B = 100 | | | | | | | |
| TRUE | -0.686 | 0.294 | 0.295 | 0.087 | 0.087 | 0.008 | 0.97 |
| Naive | -0.353 | 0.290 | 0.563 | 0.084 | 0.202 | 0.340 | 0.79 |
| McSimex.Q | -0.584 | 0.430 | 0.513 | 0.186 | 0.251 | 0.109 | 0.90 |
| NONP | -0.638 | 0.473 | 0.607 | 0.238 | 0.366 | 0.055 | 0.86 |
| NONP.CV.Var | -0.638 | 0.514 | 0.607 | 0.279 | 0.366 | 0.055 | 0.89 |
| NONP.Q.Var | -0.638 | 0.422 | 0.607 | 0.179 | 0.366 | 0.055 | 0.82 |
| B = 50 | | | | | | | |
| TRUE | -0.675 | 0.294 | 0.285 | 0.086 | 0.081 | 0.018 | 0.96 |
| Naive | -0.380 | 0.291 | 0.533 | 0.084 | 0.186 | 0.313 | 0.77 |
| McSimex.Q | -0.634 | 0.425 | 0.515 | 0.182 | 0.262 | 0.059 | 0.90 |
| NONP | -0.661 | 0.457 | 0.692 | 0.228 | 0.478 | 0.032 | 0.78 |
| NONP.CV.Var | -0.661 | 0.519 | 0.692 | 0.282 | 0.478 | 0.032 | 0.87 |
| NONP.Q.Var | -0.661 | 0.422 | 0.692 | 0.179 | 0.478 | 0.032 | 0.80 |

*Table 14: Simulation results based on 300 simulations each with sample size = 1000 and Π_(nondiff.1): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.011 | 0.135 | 0.129 | 0.018 | 0.017 | 0.011 | 0.96 |
| Naive | 0.632 | 0.137 | 0.539 | 0.019 | 0.155 | -0.368 | 0.21 |
| McSimex.Q | 0.940 | 0.191 | 0.231 | 0.036 | 0.050 | -0.060 | 0.89 |
| NONP | 0.976 | 0.215 | 0.230 | 0.047 | 0.052 | -0.024 | 0.91 |
| NONP.CV.Var | 0.976 | 0.219 | 0.230 | 0.048 | 0.052 | -0.024 | 0.91 |
| NONP.Q.Var | 0.976 | 0.190 | 0.230 | 0.036 | 0.052 | -0.02 | 0.87 |
| B = 300 | | | | | | | |
| TRUE | 0.988 | 0.135 | 0.132 | 0.018 | 0.017 | -0.012 | 0.95 |
| Naive | 0.615 | 0.137 | 0.561 | 0.019 | 0.166 | -0.385 | 0.17 |
| McSimex.Q | 0.915 | 0.191 | 0.237 | 0.036 | 0.049 | -0.085 | 0.91 |
| NONP | 0.950 | 0.215 | 0.227 | 0.047 | 0.049 | -0.050 | 0.93 |
| NONP.CV.Var | 0.950 | 0.220 | 0.227 | 0.049 | 0.049 | -0.050 | 0.93 |
| NONP.Q.Var | 0.950 | 0.189 | 0.227 | 0.036 | 0.049 | -0.050 | 0.91 |
| B = 100 | | | | | | | |
| TRUE | 1.010 | 0.135 | 0.134 | 0.018 | 0.018 | 0.010 | 0.96 |
| Naive | 0.625 | 0.137 | 0.547 | 0.019 | 0.158 | -0.375 | 0.26 |
| McSimex.Q | 0.929 | 0.191 | 0.223 | 0.037 | 0.045 | -0.071 | 0.92 |
| NONP | 0.967 | 0.213 | 0.215 | 0.046 | 0.045 | -0.033 | 0.95 |
| NONP.CV.Var | 0.967 | 0.212 | 0.215 | 0.044 | 0.045 | -0.033 | 0.91 |
| NONP.Q.Var | 0.967 | 0.190 | 0.215 | 0.036 | 0.045 | -0.033 | 0.91 |
| B = 50 | | | | | | | |
| TRUE | 1.004 | 0.135 | 0.140 | 0.018 | 0.020 | 0.004 | 0.94 |
| Naive | 0.627 | 0.137 | 0.546 | 0.019 | 0.159 | -0.373 | 0.23 |
| McSimex.Q | 0.930 | 0.190 | 0.236 | 0.036 | 0.051 | -0.070 | 0.90 |
| NONP | 0.975 | 0.208 | 0.254 | 0.044 | 0.064 | -0.025 | 0.90 |
| NONP.CV.Var | 0.975 | 0.215 | 0.254 | 0.046 | 0.064 | -0.025 | 0.89 |
| NONP.Q.Var | 0.975 | 0.189 | 0.254 | 0.036 | 0.064 | -0.025 | 0.87 |

*Table 15: Simulation results based on 300 simulations each with sample size = 500 and Π_(nondiff.1): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.028 | 0.192 | 0.196 | 0.037 | 0.038 | 0.028 | 0.95 |
| Naive | 0.652 | 0.195 | 0.526 | 0.038 | 0.156 | -0.348 | 0.57 |
| McSimex.Q | 0.967 | 0.272 | 0.285 | 0.074 | 0.080 | -0.033 | 0.94 |
| NONP | 0.999 | 0.309 | 0.301 | 0.097 | 0.091 | -0.001 | 0.94 |
| NONP.CV.Var | 0.999 | 0.309 | 0.301 | 0.095 | 0.091 | -0.001 | 0.95 |
| NONP.Q.Var | 0.999 | 0.270 | 0.301 | 0.073 | 0.091 | 0.00 | 0.91 |
| B = 300 | | | | | | | |
| TRUE | 0.999 | 0.191 | 0.191 | 0.037 | 0.037 | -0.001 | 0.94 |
| Naive | 0.624 | 0.194 | 0.566 | 0.038 | 0.179 | -0.376 | 0.49 |
| McSimex.Q | 0.927 | 0.271 | 0.309 | 0.074 | 0.090 | -0.073 | 0.92 |
| NONP | 0.971 | 0.299 | 0.312 | 0.091 | 0.097 | -0.029 | 0.92 |
| NONP.CV.Var | 0.971 | 0.314 | 0.312 | 0.098 | 0.097 | -0.029 | 0.94 |
| NONP.Q.Var | 0.971 | 0.268 | 0.312 | 0.072 | 0.097 | -0.029 | 0.90 |
| B = 100 | | | | | | | |
| TRUE | 1.001 | 0.192 | 0.182 | 0.037 | 0.033 | 0.001 | 0.96 |
| Naive | 0.619 | 0.194 | 0.568 | 0.038 | 0.178 | -0.381 | 0.47 |
| McSimex.Q | 0.916 | 0.268 | 0.296 | 0.072 | 0.081 | -0.084 | 0.93 |
| NONP | 0.971 | 0.290 | 0.310 | 0.086 | 0.095 | -0.029 | 0.92 |
| NONP.CV.Var | 0.971 | 0.303 | 0.310 | 0.090 | 0.095 | -0.029 | 0.91 |
| NONP.Q.Var | 0.971 | 0.269 | 0.310 | 0.072 | 0.095 | -0.029 | 0.91 |
| B = 50 | | | | | | | |
| TRUE | 1.002 | 0.191 | 0.191 | 0.037 | 0.036 | 0.002 | 0.95 |
| Naive | 0.631 | 0.194 | 0.561 | 0.038 | 0.179 | -0.369 | 0.52 |
| McSimex.Q | 0.937 | 0.266 | 0.336 | 0.071 | 0.109 | -0.063 | 0.87 |
| NONP | 0.998 | 0.290 | 0.386 | 0.086 | 0.149 | -0.002 | 0.86 |
| NONP.CV.Var | 0.998 | 0.308 | 0.386 | 0.093 | 0.149 | -0.002 | 0.89 |
| NONP.Q.Var | 0.998 | 0.268 | 0.386 | 0.072 | 0.149 | -0.002 | 0.84 |

*Table 16: Simulation results based on 300 simulations each with sample size = 200 and Π_(nondiff.1): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.02 | 0.305 | 0.306 | 0.093 | 0.093 | 0.020 | 0.94 |
| Naive | 0.634 | 0.310 | 0.601 | 0.096 | 0.227 | -0.366 | 0.77 |
| McSimex.Q | 0.939 | 0.433 | 0.467 | 0.188 | 0.214 | -0.061 | 0.92 |
| NONP | 0.968 | 0.487 | 0.488 | 0.244 | 0.237 | -0.032 | 0.92 |
| NONP.CV.Var | 0.968 | 0.500 | 0.488 | 0.251 | 0.237 | -0.032 | 0.95 |
| NONP.Q.Var | 0.968 | 0.431 | 0.488 | 0.186 | 0.237 | -0.03 | 0.9 |
| B = 300 | | | | | | | |
| TRUE | 1.043 | 0.305 | 0.325 | 0.093 | 0.104 | 0.043 | 0.93 |
| Naive | 0.657 | 0.309 | 0.586 | 0.096 | 0.225 | -0.343 | 0.75 |
| McSimex.Q | 0.972 | 0.434 | 0.492 | 0.189 | 0.241 | -0.028 | 0.91 |
| NONP | 1.012 | 0.483 | 0.536 | 0.239 | 0.287 | 0.012 | 0.90 |
| NONP.CV.Var | 1.012 | 0.495 | 0.536 | 0.244 | 0.287 | 0.012 | 0.91 |
| NONP.Q.Var | 1.012 | 0.429 | 0.536 | 0.185 | 0.287 | 0.012 | 0.88 |
| B = 100 | | | | | | | |
| TRUE | 0.991 | 0.304 | 0.317 | 0.092 | 0.100 | -0.009 | 0.93 |
| Naive | 0.638 | 0.309 | 0.601 | 0.096 | 0.230 | -0.362 | 0.75 |
| McSimex.Q | 0.944 | 0.429 | 0.487 | 0.186 | 0.234 | -0.056 | 0.92 |
| NONP | 1.001 | 0.461 | 0.565 | 0.215 | 0.320 | 0.001 | 0.89 |
| NONP.CV.Var | 1.001 | 0.493 | 0.565 | 0.234 | 0.320 | 0.001 | 0.89 |
| NONP.Q.Var | 1.001 | 0.428 | 0.565 | 0.184 | 0.320 | 0.001 | 0.88 |
| B = 50 | | | | | | | |
| TRUE | 1.023 | 0.305 | 0.318 | 0.093 | 0.101 | 0.023 | 0.95 |
| Naive | 0.655 | 0.309 | 0.590 | 0.096 | 0.229 | -0.345 | 0.75 |
| McSimex.Q | 0.958 | 0.429 | 0.501 | 0.187 | 0.250 | -0.042 | 0.91 |
| NONP | 1.021 | 0.456 | 0.633 | 0.213 | 0.400 | 0.021 | 0.87 |
| NONP.CV.Var | 1.021 | 0.486 | 0.633 | 0.237 | 0.400 | 0.021 | 0.87 |
| NONP.Q.Var | 1.021 | 0.427 | 0.633 | 0.183 | 0.400 | 0.021 | 0.84 |

*Table 17: Simulation results based on 300 simulations each with sample size = 1000 and Π_(nondiff.1): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.691 | 0.131 | 0.139 | 0.017 | 0.019 | 0.002 | 0.93 |
| Naive | -0.446 | 0.133 | 0.373 | 0.018 | 0.078 | 0.247 | 0.49 |
| McSimex.Q | -0.661 | 0.185 | 0.200 | 0.034 | 0.039 | 0.032 | 0.93 |
| NONP | -0.685 | 0.207 | 0.208 | 0.044 | 0.043 | 0.008 | 0.93 |
| NONP.CV.Var | -0.685 | 0.210 | 0.208 | 0.043 | 0.043 | 0.008 | 0.94 |
| NONP.Q.Var | -0.685 | 0.184 | 0.208 | 0.034 | 0.043 | 0.01 | 0.92 |
| B = 300 | | | | | | | |
| TRUE | -0.693 | 0.131 | 0.132 | 0.017 | 0.017 | 0.000 | 0.95 |
| Naive | -0.430 | 0.133 | 0.399 | 0.018 | 0.090 | 0.263 | 0.49 |
| McSimex.Q | -0.638 | 0.185 | 0.229 | 0.034 | 0.050 | 0.055 | 0.87 |
| NONP | -0.663 | 0.201 | 0.227 | 0.041 | 0.051 | 0.030 | 0.89 |
| NONP.CV.Var | -0.663 | 0.209 | 0.227 | 0.043 | 0.051 | 0.030 | 0.92 |
| NONP.Q.Var | -0.663 | 0.183 | 0.227 | 0.034 | 0.051 | 0.030 | 0.88 |
| B = 100 | | | | | | | |
| TRUE | -0.686 | 0.131 | 0.136 | 0.017 | 0.019 | 0.007 | 0.94 |
| Naive | -0.429 | 0.133 | 0.394 | 0.018 | 0.086 | 0.264 | 0.55 |
| McSimex.Q | -0.637 | 0.182 | 0.206 | 0.033 | 0.039 | 0.056 | 0.91 |
| NONP | -0.664 | 0.200 | 0.217 | 0.041 | 0.046 | 0.029 | 0.92 |
| NONP.CV.Var | -0.664 | 0.207 | 0.217 | 0.041 | 0.046 | 0.029 | 0.92 |
| NONP.Q.Var | -0.664 | 0.183 | 0.217 | 0.034 | 0.046 | 0.029 | 0.91 |
| B = 50 | | | | | | | |
| TRUE | -0.701 | 0.131 | 0.136 | 0.017 | 0.018 | -0.008 | 0.93 |
| Naive | -0.435 | 0.133 | 0.388 | 0.018 | 0.084 | 0.258 | 0.50 |
| McSimex.Q | -0.649 | 0.185 | 0.213 | 0.035 | 0.043 | 0.044 | 0.91 |
| NONP | -0.678 | 0.198 | 0.280 | 0.041 | 0.078 | 0.015 | 0.86 |
| NONP.CV.Var | -0.678 | 0.207 | 0.280 | 0.043 | 0.078 | 0.015 | 0.85 |
| NONP.Q.Var | -0.678 | 0.182 | 0.280 | 0.033 | 0.078 | 0.015 | 0.82 |

*Table 18: Simulation results based on 300 simulations each with sample size = 500 and Π_(nondiff.1): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.701 | 0.185 | 0.188 | 0.034 | 0.035 | -0.008 | 0.95 |
| Naive | -0.437 | 0.189 | 0.406 | 0.036 | 0.099 | 0.256 | 0.76 |
| McSimex.Q | -0.647 | 0.263 | 0.280 | 0.069 | 0.076 | 0.046 | 0.94 |
| NONP | -0.673 | 0.292 | 0.295 | 0.086 | 0.087 | 0.020 | 0.92 |
| NONP.CV.Var | -0.673 | 0.200 | 0.295 | 0.089 | 0.087 | 0.020 | 0.94 |
| NONP.Q.Var | -0.673 | 0.260 | 0.295 | 0.068 | 0.087 | 0.02 | 0.92 |
| B = 300 | | | | | | | |
| TRUE | -0.709 | 0.185 | 0.180 | 0.034 | 0.032 | -0.016 | 0.96 |
| Naive | -0.444 | 0.189 | 0.394 | 0.036 | 0.093 | 0.250 | 0.74 |
| McSimex.Q | -0.657 | 0.262 | 0.266 | 0.069 | 0.070 | 0.036 | 0.95 |
| NONP | -0.692 | 0.284 | 0.277 | 0.082 | 0.077 | 0.001 | 0.94 |
| NONP.CV.Var | -0.692 | 0.298 | 0.277 | 0.087 | 0.077 | 0.001 | 0.95 |
| NONP.Q.Var | -0.692 | 0.260 | 0.277 | 0.068 | 0.077 | 0.001 | 0.94 |
| B = 100 | | | | | | | |
| TRUE | -0.684 | 0.185 | 0.185 | 0.034 | 0.034 | 0.009 | 0.95 |
| Naive | -0.419 | 0.189 | 0.428 | 0.036 | 0.108 | 0.274 | 0.69 |
| McSimex.Q | -0.624 | 0.260 | 0.293 | 0.068 | 0.081 | 0.069 | 0.92 |
| NONP | -0.657 | 0.279 | 0.317 | 0.079 | 0.099 | 0.036 | 0.92 |
| NONP.CV.Var | -0.657 | 0.289 | 0.317 | 0.081 | 0.099 | 0.036 | 0.92 |
| NONP.Q.Var | -0.657 | 0.260 | 0.317 | 0.068 | 0.099 | 0.036 | 0.90 |
| B = 50 | | | | | | | |
| TRUE | -0.689 | 0.185 | 0.188 | 0.034 | 0.035 | 0.004 | 0.95 |
| Naive | -0.428 | 0.189 | 0.421 | 0.036 | 0.107 | 0.265 | 0.69 |
| McSimex.Q | -0.639 | 0.259 | 0.302 | 0.068 | 0.088 | 0.054 | 0.91 |
| NONP | -0.664 | 0.275 | 0.374 | 0.077 | 0.139 | 0.029 | 0.87 |
| NONP.CV.Var | -0.664 | 0.296 | 0.374 | 0.086 | 0.139 | 0.029 | 0.88 |
| NONP.Q.Var | -0.664 | 0.259 | 0.374 | 0.067 | 0.139 | 0.029 | 0.86 |

*Table 19: Simulation results based on 300 simulations each with sample size = 200 and Π_(nondiff.1): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.695 | 0.294 | 0.305 | 0.087 | 0.093 | -0.002 | 0.94 |
| Naive | -0.458 | 0.301 | 0.459 | 0.090 | 0.155 | 0.235 | 0.84 |
| McSimex.Q | -0.676 | 0.419 | 0.474 | 0.176 | 0.225 | 0.017 | 0.91 |
| NONP | -0.702 | 0.457 | 0.501 | 0.214 | 0.251 | -0.008 | 0.89 |
| NONP.CV.Var | -0.702 | 0.473 | 0.501 | 0.222 | 0.251 | -0.008 | 0.93 |
| NONP.Q.Var | -0.702 | 0.416 | 0.501 | 0.173 | 0.251 | -0.01 | 0.89 |
| B = 300 | | | | | | | |
| TRUE | -0.698 | 0.294 | 0.200 | 0.087 | 0.090 | -0.004 | 0.94 |
| Naive | -0.435 | 0.200 | 0.478 | 0.090 | 0.161 | 0.258 | 0.85 |
| McSimex.Q | -0.638 | 0.419 | 0.465 | 0.176 | 0.213 | 0.055 | 0.92 |
| NONP | -0.663 | 0.459 | 0.500 | 0.212 | 0.249 | 0.030 | 0.92 |
| NONP.CV.Var | -0.663 | 0.483 | 0.500 | 0.228 | 0.249 | 0.030 | 0.95 |
| NONP.Q.Var | -0.663 | 0.414 | 0.500 | 0.172 | 0.249 | 0.030 | 0.89 |
| B = 100 | | | | | | | |
| TRUE | -0.718 | 0.294 | 0.277 | 0.087 | 0.076 | -0.025 | 0.97 |
| Naive | -0.463 | 0.301 | 0.439 | 0.091 | 0.140 | 0.230 | 0.89 |
| McSimex.Q | -0.691 | 0.414 | 0.447 | 0.173 | 0.200 | 0.002 | 0.93 |
| NONP | -0.721 | 0.442 | 0.532 | 0.201 | 0.282 | -0.027 | 0.88 |
| NONP.CV.Var | -0.721 | 0.471 | 0.532 | 0.209 | 0.282 | -0.027 | 0.90 |
| NONP.Q.Var | -0.721 | 0.415 | 0.532 | 0.173 | 0.282 | -0.027 | 0.88 |
| B = 50 | | | | | | | |
| TRUE | -0.683 | 0.294 | 0.296 | 0.086 | 0.087 | 0.010 | 0.95 |
| Naive | -0.450 | 0.200 | 0.457 | 0.090 | 0.149 | 0.243 | 0.85 |
| McSimex.Q | -0.658 | 0.418 | 0.460 | 0.177 | 0.211 | 0.036 | 0.91 |
| NONP | -0.670 | 0.435 | 0.576 | 0.194 | 0.331 | 0.023 | 0.85 |
| NONP.CV.Var | -0.670 | 0.467 | 0.576 | 0.210 | 0.331 | 0.023 | 0.88 |
| NONP.Q.Var | -0.670 | 0.414 | 0.576 | 0.172 | 0.331 | 0.023 | 0.86 |

*Table 20: Simulation results based on 300 simulations each with sample size = 1000 and Π_(nondiff.2): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.01 | 0.135 | 0.131 | 0.018 | 0.017 | 0.010 | 0.97 |
| Naive | 0.584 | 0.132 | 0.605 | 0.017 | 0.193 | -0.416 | 0.13 |
| McSimex.Q | 0.912 | 0.189 | 0.255 | 0.036 | 0.058 | -0.088 | 0.87 |
| NONP | 0.992 | 0.210 | 0.254 | 0.045 | 0.064 | -0.008 | 0.88 |
| NONP.CV.Var | 0.992 | 0.218 | 0.254 | 0.055 | 0.064 | -0.008 | 0.88 |
| NONP.Q.Var | 0.992 | 0.187 | 0.254 | 0.035 | 0.064 | -0.01 | 0.86 |
| B = 300 | | | | | | | |
| TRUE | 1.003 | 0.135 | 0.132 | 0.018 | 0.017 | 0.003 | 0.95 |
| Naive | 0.600 | 0.132 | 0.580 | 0.017 | 0.176 | -0.400 | 0.12 |
| McSimex.Q | 0.938 | 0.189 | 0.214 | 0.036 | 0.042 | -0.062 | 0.91 |
| NONP | 1.013 | 0.208 | 0.229 | 0.045 | 0.052 | 0.013 | 0.92 |
| NONP.CV.Var | 1.013 | 0.209 | 0.229 | 0.047 | 0.052 | 0.013 | 0.88 |
| NONP.Q.Var | 1.013 | 0.187 | 0.229 | 0.035 | 0.052 | 0.013 | 0.88 |
| B = 100 | | | | | | | |
| TRUE | 1.014 | 0.135 | 0.136 | 0.018 | 0.018 | 0.014 | 0.94 |
| Naive | 0.599 | 0.132 | 0.584 | 0.017 | 0.180 | -0.401 | 0.15 |
| McSimex.Q | 0.937 | 0.188 | 0.235 | 0.036 | 0.051 | -0.063 | 0.90 |
| NONP | 1.002 | 0.204 | 0.265 | 0.043 | 0.070 | 0.002 | 0.85 |
| NONP.CV.Var | 1.002 | 0.210 | 0.265 | 0.046 | 0.070 | 0.002 | 0.84 |
| NONP.Q.Var | 1.002 | 0.187 | 0.265 | 0.035 | 0.070 | 0.002 | 0.84 |
| B = 50 | | | | | | | |
| TRUE | 0.994 | 0.135 | 0.139 | 0.018 | 0.019 | -0.006 | 0.94 |
| Naive | 0.586 | 0.132 | 0.600 | 0.017 | 0.188 | -0.414 | 0.13 |
| McSimex.Q | 0.914 | 0.187 | 0.241 | 0.035 | 0.051 | -0.086 | 0.89 |
| NONP | 0.956 | 0.198 | 0.279 | 0.040 | 0.076 | -0.044 | 0.85 |
| NONP.CV.Var | 0.956 | 0.211 | 0.279 | 0.046 | 0.076 | -0.044 | 0.85 |
| NONP.Q.Var | 0.956 | 0.186 | 0.279 | 0.035 | 0.076 | -0.044 | 0.82 |

*Table 21: Simulation results based on 300 simulations each with sample size = 500 and Π_(nondiff.2): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.009 | 0.192 | 0.180 | 0.037 | 0.032 | 0.009 | 0.97 |
| Naive | 0.619 | 0.187 | 0.567 | 0.035 | 0.177 | -0.381 | 0.47 |
| McSimex.Q | 0.969 | 0.270 | 0.287 | 0.073 | 0.081 | -0.031 | 0.92 |
| NONP | 1.046 | 0.297 | 0.330 | 0.091 | 0.107 | 0.046 | 0.92 |
| NONP.CV.Var | 1.046 | 0.308 | 0.330 | 0.099 | 0.107 | 0.046 | 0.91 |
| NONP.Q.Var | 1.046 | 0.266 | 0.330 | 0.071 | 0.107 | 0.05 | 0.87 |
| B = 300 | | | | | | | |
| TRUE | 1.004 | 0.191 | 0.194 | 0.037 | 0.037 | 0.004 | 0.95 |
| Naive | 0.583 | 0.187 | 0.614 | 0.035 | 0.203 | -0.417 | 0.40 |
| McSimex.Q | 0.911 | 0.268 | 0.298 | 0.072 | 0.081 | -0.089 | 0.95 |
| NONP | 0.974 | 0.287 | 0.313 | 0.085 | 0.097 | -0.026 | 0.93 |
| NONP.CV.Var | 0.974 | 0.297 | 0.313 | 0.093 | 0.097 | -0.026 | 0.89 |
| NONP.Q.Var | 0.974 | 0.265 | 0.313 | 0.070 | 0.097 | -0.026 | 0.91 |
| B = 100 | | | | | | | |
| TRUE | 1.004 | 0.192 | 0.190 | 0.037 | 0.036 | 0.004 | 0.95 |
| Naive | 0.604 | 0.187 | 0.595 | 0.035 | 0.197 | -0.396 | 0.45 |
| McSimex.Q | 0.944 | 0.267 | 0.324 | 0.072 | 0.102 | -0.056 | 0.89 |
| NONP | 1.015 | 0.282 | 0.368 | 0.082 | 0.135 | 0.015 | 0.88 |
| NONP.CV.Var | 1.015 | 0.301 | 0.368 | 0.096 | 0.135 | 0.015 | 0.86 |
| NONP.Q.Var | 1.015 | 0.265 | 0.368 | 0.070 | 0.135 | 0.015 | 0.85 |
| B = 50 | | | | | | | |
| TRUE | 1.017 | 0.192 | 0.201 | 0.037 | 0.040 | 0.017 | 0.95 |
| Naive | 0.610 | 0.187 | 0.585 | 0.035 | 0.190 | -0.390 | 0.44 |
| McSimex.Q | 0.957 | 0.267 | 0.317 | 0.072 | 0.099 | -0.043 | 0.88 |
| NONP | 1.001 | 0.281 | 0.385 | 0.081 | 0.148 | 0.001 | 0.84 |
| NONP.CV.Var | 1.001 | 0.303 | 0.385 | 0.096 | 0.148 | 0.001 | 0.85 |
| NONP.Q.Var | 1.001 | 0.264 | 0.385 | 0.070 | 0.148 | 0.001 | 0.81 |

*Table 22:Simulation results based on 300 simulations each with sample size = 200 and Π_(nondiff.2): The true logistic regression coefficients were (0, 1).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | 1.02 | 0.305 | 0.295 | 0.093 | 0.087 | 0.020 | 0.97 |
| Naive | 0.590 | 0.297 | 0.649 | 0.088 | 0.253 | -0.410 | 0.72 |
| McSimex.Q | 0.922 | 0.427 | 0.468 | 0.182 | 0.213 | -0.078 | 0.91 |
| NONP | 0.994 | 0.456 | 0.529 | 0.217 | 0.280 | -0.006 | 0.92 |
| NONP.CV.Var | 0.994 | 0.493 | 0.529 | 0.254 | 0.280 | -0.006 | 0.88 |
| NONP.Q.Var | 0.994 | 0.422 | 0.529 | 0.178 | 0.280 | -0.01 | 0.88 |
| B = 300 | | | | | | | |
| TRUE | 0.985 | 0.304 | 0.299 | 0.093 | 0.089 | -0.015 | 0.96 |
| Naive | 0.570 | 0.297 | 0.679 | 0.088 | 0.276 | -0.430 | 0.68 |
| McSimex.Q | 0.895 | 0.426 | 0.498 | 0.182 | 0.237 | -0.105 | 0.91 |
| NONP | 0.924 | 0.447 | 0.572 | 0.206 | 0.321 | -0.076 | 0.88 |
| NONP.CV.Var | 0.924 | 0.497 | 0.572 | 0.254 | 0.321 | -0.076 | 0.91 |
| NONP.Q.Var | 0.924 | 0.422 | 0.572 | 0.178 | 0.321 | -0.076 | 0.85 |
| B = 100 | | | | | | | |
| TRUE | 1.020 | 0.305 | 0.315 | 0.093 | 0.099 | 0.020 | 0.95 |
| Naive | 0.598 | 0.297 | 0.640 | 0.088 | 0.247 | -0.402 | 0.71 |
| McSimex.Q | 0.932 | 0.428 | 0.477 | 0.184 | 0.223 | -0.068 | 0.93 |
| NONP | 0.968 | 0.457 | 0.582 | 0.216 | 0.338 | -0.032 | 0.90 |
| NONP.CV.Var | 0.968 | 0.485 | 0.582 | 0.244 | 0.338 | -0.032 | 0.90 |
| NONP.Q.Var | 0.968 | 0.423 | 0.582 | 0.179 | 0.338 | -0.03 | 0.88 |
| B = 50 | | | | | | | |
| TRUE | 1.008 | 0.304 | 0.200 | 0.093 | 0.090 | 0.008 | 0.95 |
| Naive | 0.599 | 0.296 | 0.641 | 0.088 | 0.250 | -0.401 | 0.69 |
| McSimex.Q | 0.932 | 0.425 | 0.490 | 0.182 | 0.236 | -0.068 | 0.92 |
| NONP | 0.972 | 0.444 | 0.615 | 0.204 | 0.377 | -0.028 | 0.87 |
| NONP.CV.Var | 0.972 | 0.474 | 0.615 | 0.231 | 0.377 | -0.028 | 0.89 |
| NONP.Q.Var | 0.972 | 0.421 | 0.615 | 0.177 | 0.377 | -0.028 | 0.87 |

*Table 23: Simulation results based on 300 simulations each with sample size = 1000 and Π_(nondiff.2): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.702 | 0.131 | 0.130 | 0.017 | 0.017 | -0.009 | 0.95 |
| Naive | -0.423 | 0.129 | 0.400 | 0.017 | 0.087 | 0.270 | 0.45 |
| McSimex.Q | -0.660 | 0.185 | 0.192 | 0.034 | 0.036 | 0.033 | 0.95 |
| NONP | -0.703 | 0.194 | 0.207 | 0.038 | 0.043 | -0.010 | 0.94 |
| NONP.CV.Var | -0.703 | 0.209 | 0.207 | 0.045 | 0.043 | -0.010 | 0.92 |
| NONP.Q.Var | -0.703 | 0.183 | 0.207 | 0.033 | 0.043 | -0.01 | 0.91 |
| B = 300 | | | | | | | |
| TRUE | -0.690 | 0.131 | 0.141 | 0.017 | 0.020 | 0.003 | 0.93 |
| Naive | -0.405 | 0.129 | 0.426 | 0.017 | 0.099 | 0.288 | 0.38 |
| McSimex.Q | -0.631 | 0.185 | 0.214 | 0.034 | 0.042 | 0.062 | 0.93 |
| NONP | -0.677 | 0.193 | 0.217 | 0.038 | 0.047 | 0.016 | 0.91 |
| NONP.CV.Var | -0.677 | 0.206 | 0.217 | 0.044 | 0.047 | 0.016 | 0.90 |
| NONP.Q.Var | -0.677 | 0.183 | 0.217 | 0.033 | 0.047 | 0.016 | 0.90 |
| B = 100 | | | | | | | |
| TRUE | -0.701 | 0.131 | 0.134 | 0.017 | 0.018 | -0.008 | 0.95 |
| Naive | -0.409 | 0.129 | 0.423 | 0.017 | 0.098 | 0.284 | 0.40 |
| McSimex.Q | -0.639 | 0.186 | 0.222 | 0.035 | 0.046 | 0.054 | 0.92 |
| NONP | -0.668 | 0.191 | 0.272 | 0.036 | 0.074 | 0.025 | 0.87 |
| NONP.CV.Var | -0.668 | 0.200 | 0.272 | 0.042 | 0.074 | 0.025 | 0.85 |
| NONP.Q.Var | -0.668 | 0.183 | 0.272 | 0.033 | 0.074 | 0.03 | 0.84 |
| B = 50 | | | | | | | |
| TRUE | -0.684 | 0.131 | 0.131 | 0.017 | 0.017 | 0.009 | 0.95 |
| Naive | -0.419 | 0.129 | 0.411 | 0.017 | 0.094 | 0.274 | 0.45 |
| McSimex.Q | -0.656 | 0.181 | 0.220 | 0.033 | 0.047 | 0.037 | 0.89 |
| NONP | -0.685 | 0.189 | 0.268 | 0.037 | 0.071 | 0.008 | 0.83 |
| NONP.CV.Var | -0.685 | 0.207 | 0.268 | 0.044 | 0.071 | 0.008 | 0.84 |
| NONP.Q.Var | -0.685 | 0.182 | 0.268 | 0.033 | 0.071 | 0.008 | 0.81 |

*Table 24: Simulation results based on 300 simulations each with sample size = 500 and Π_(nondiff.2): The true logistic regression coefficients were(0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.686 | 0.185 | 0.190 | 0.034 | 0.036 | 0.007 | 0.96 |
| Naive | -0.418 | 0.183 | 0.436 | 0.034 | 0.114 | 0.276 | 0.66 |
| McSimex.Q | -0.652 | 0.262 | 0.311 | 0.069 | 0.095 | 0.041 | 0.90 |
| NONP | -0.696 | 0.274 | 0.337 | 0.077 | 0.113 | -0.003 | 0.88 |
| NONP.CV.Var | -0.696 | 0.200 | 0.337 | 0.092 | 0.113 | -0.003 | 0.90 |
| NONP.Q.Var | -0.696 | 0.259 | 0.337 | 0.067 | 0.113 | 0.00 | 0.86 |
| B = 300 | | | | | | | |
| TRUE | -0.678 | 0.185 | 0.194 | 0.034 | 0.038 | 0.015 | 0.94 |
| Naive | -0.420 | 0.183 | 0.431 | 0.034 | 0.112 | 0.273 | 0.66 |
| McSimex.Q | -0.656 | 0.261 | 0.308 | 0.068 | 0.093 | 0.037 | 0.92 |
| NONP | -0.692 | 0.271 | 0.349 | 0.076 | 0.122 | 0.001 | 0.90 |
| NONP.CV.Var | -0.692 | 0.299 | 0.349 | 0.093 | 0.122 | 0.001 | 0.89 |
| NONP.Q.Var | -0.692 | 0.259 | 0.349 | 0.067 | 0.122 | 0.001 | 0.85 |
| B = 100 | | | | | | | |
| TRUE | -0.697 | 0.185 | 0.187 | 0.034 | 0.035 | -0.004 | 0.94 |
| Naive | -0.407 | 0.183 | 0.445 | 0.034 | 0.116 | 0.286 | 0.65 |
| McSimex.Q | -0.636 | 0.261 | 0.305 | 0.068 | 0.090 | 0.057 | 0.89 |
| NONP | -0.652 | 0.271 | 0.346 | 0.076 | 0.118 | 0.041 | 0.89 |
| NONP.CV.Var | -0.652 | 0.283 | 0.346 | 0.082 | 0.118 | 0.041 | 0.85 |
| NONP.Q.Var | -0.652 | 0.259 | 0.346 | 0.067 | 0.118 | 0.04 | 0.87 |
| B = 50 | | | | | | | |
| TRUE | -0.698 | 0.185 | 0.185 | 0.034 | 0.034 | -0.004 | 0.94 |
| Naive | -0.411 | 0.183 | 0.434 | 0.033 | 0.109 | 0.282 | 0.65 |
| McSimex.Q | -0.640 | 0.260 | 0.288 | 0.068 | 0.080 | 0.053 | 0.93 |
| NONP | -0.662 | 0.271 | 0.342 | 0.076 | 0.116 | 0.031 | 0.90 |
| NONP.CV.Var | -0.662 | 0.291 | 0.342 | 0.088 | 0.116 | 0.031 | 0.90 |
| NONP.Q.Var | -0.662 | 0.259 | 0.342 | 0.067 | 0.116 | 0.031 | 0.88 |

*Table 25: Simulation results based on 300 simulations each with sample size = 200 and Π_(nondiff.2): The true logistic regression coefficients were (0, -log2).*

| Methods | Mean | SE | RMSE | Estimated Variance | Empirical Variance | Bias | Coverage |
|---|---|---|---|---|---|---|---|
| B = 500 | | | | | | | |
| TRUE | -0.7 | 0.294 | 0.298 | 0.086 | 0.089 | -0.007 | 0.94 |
| Naive | -0.431 | 0.291 | 0.473 | 0.085 | 0.155 | 0.262 | 0.87 |
| McSimex.Q | -0.673 | 0.416 | 0.461 | 0.174 | 0.212 | 0.021 | 0.91 |
| NONP | -0.712 | 0.435 | 0.513 | 0.195 | 0.262 | -0.019 | 0.88 |
| NONP.CV.Var | -0.712 | 0.485 | 0.513 | 0.241 | 0.262 | -0.019 | 0.91 |
| NONP.Q.Var | -0.712 | 0.413 | 0.513 | 0.170 | 0.262 | -0.02 | 0.88 |
| B = 300 | | | | | | | |
| TRUE | -0.688 | 0.294 | 0.283 | 0.087 | 0.080 | 0.005 | 0.95 |
| Naive | -0.421 | 0.291 | 0.481 | 0.085 | 0.157 | 0.272 | 0.85 |
| McSimex.Q | -0.658 | 0.418 | 0.456 | 0.175 | 0.206 | 0.035 | 0.92 |
| NONP | -0.685 | 0.434 | 0.511 | 0.193 | 0.261 | 0.008 | 0.90 |
| NONP.CV.Var | -0.685 | 0.473 | 0.511 | 0.232 | 0.261 | 0.008 | 0.91 |
| NONP.Q.Var | -0.685 | 0.413 | 0.511 | 0.170 | 0.261 | 0.008 | 0.89 |
| B = 100 | | | | | | | |
| TRUE | -0.723 | 0.295 | 0.200 | 0.087 | 0.089 | -0.030 | 0.95 |
| Naive | -0.404 | 0.291 | 0.504 | 0.085 | 0.170 | 0.289 | 0.81 |
| McSimex.Q | -0.630 | 0.413 | 0.473 | 0.171 | 0.220 | 0.063 | 0.91 |
| NONP | -0.637 | 0.437 | 0.566 | 0.196 | 0.318 | 0.056 | 0.84 |
| NONP.CV.Var | -0.637 | 0.467 | 0.566 | 0.229 | 0.318 | 0.056 | 0.87 |
| NONP.Q.Var | -0.637 | 0.411 | 0.566 | 0.169 | 0.318 | 0.06 | 0.86 |
| B = 50 | | | | | | | |
| TRUE | -0.689 | 0.294 | 0.268 | 0.086 | 0.072 | 0.004 | 0.97 |
| Naive | -0.417 | 0.291 | 0.480 | 0.085 | 0.154 | 0.276 | 0.83 |
| McSimex.Q | -0.641 | 0.416 | 0.455 | 0.174 | 0.204 | 0.052 | 0.93 |
| NONP | -0.660 | 0.414 | 0.577 | 0.181 | 0.331 | 0.033 | 0.83 |
| NONP.CV.Var | -0.660 | 0.461 | 0.577 | 0.221 | 0.331 | 0.033 | 0.87 |
| NONP.Q.Var | -0.660 | 0.411 | 0.577 | 0.169 | 0.331 | 0.033 | 0.88 |

# CHAPTER 5

# APPLICATION TO NHANES DATA

## Introduction

The National Health and Nutrition Examination Survey (NHANES) is a program of the Centers for Disease Control and Prevention (CDC) which focus on a variety of health and nutrition measurements. The NHANES program began in the 1960s and after 1999 it became a continuous, annual (CDC, 2017a). It uses a multistage sampling design to select participants from the United States. Approximately 5,000 randomly selected, confidential and voluntary residents across the United States have the opportunity to participate in the latest NHANES (CDC, 2017a). This program is unique because it consists of two parts: interviews and laboratory examinations. Hence, we can discover the accuracy of participants self-reported data by comparing the interview data and laboratory examination data.

We use the data from the NHANES to show how our new method corrects the bias produced by the MC-SIMEX estimator in a logistic regression in an epidemiological study. We are interested in the association between obesity (exposure) and diabetes (outcome). In the NHANES data, we included male aged 50 to 60 participating in NHANES in 2010. The variables that are provided in this dataset include: SEQN - respondent sequence number, RIDAGEYR - age at screening adjudicated (in years), mobesity - obesity status as measured BMI determined on 30 lb./in^2 (CDC, 2017b), srobesity - obesity status as self-reported BMI determined on 30 lb./in^2 (CDC, 2017b), diabetes - self-reported diabetes's status, HBP - participant has been told by a doctor or other health professional that he/she had hypertension, also called high blood pressure. For simplicity, there is no distinction between type 1 and type 2 diabetes, and we assume that diabetes's status and HBP status were reported with no misclassification.

## Misclassification matrix

The obesity status was categorized into two classes, no (0) and yes (1). When individuals' BMI was great than and equal to 30 lb./in^2, it was considered as obesity. When individuals' BMI was less than 30 lb./in^2, it was considered as non-obesity. The BMI value calculated from self-reported height and weight

was considered as “self-reported BMI”, hence the obesity status assessed by “self-reported BMI” was considered as a naive covariate (misclassified exposure, *W*). The BMI value calculated from laboratory examines height and weight was considered as “measured BMI”, hence the obesity status assessed by “measured BMI” was considered as a true covariate (true exposure, *X*).

The misclassification matrix Π is estimated using the validation data since we have both measured and self-reported height and weight for each individual. Table 26 and 27 are the frequency table of the obesity by diabetes's status.

*Table 26: Table of srobesity by mobesity when diabetes = 1*

| | mobesity | |
|---|---|---|
| srobesity | 1 | 0 |
| 1 | 46 | 5 |
| 0 | 8 | 25 |

*Table 27: Table of srobesity by mobesity when diabetes = 0*

| | mobesity | |
|---|---|---|
| srobesity | 1 | 0 |
| 1 | 119 | 19 |
| 0 | 33 | 281 |

We use tables 26 and 27 to estimate the misclassification matrix Π,

$$\Pi = \begin{bmatrix} 0.94 & 0.22 & 0 & 0 \\ 0.06 & 0.78 & 0 & 0 \\ 0 & 0 & 0.83 & 0.15 \\ 0 & 0 & 0.17 & 0.85 \end{bmatrix}. \tag{5.1}$$

This is obvious differential misclassification matrix; hence the misclassification error of obesity status is the differential misclassification error based on the diabetes's status.

## Analysis

We categorize the obesity status into two categories: individual was considered as obesity when their BMI was great than and equal to 30. And individual was considered as non-obesity when their BMI was less than 30. Then we treat the obesity status due to "self-reported BMI" as naive covariate ($W$) and the obesity status due to "measured BMI" as the true covariate ($X$).

In this section, we will compare three methods: naive method, original MC-SIMEX method and nonparametric MC-SIMEX method with three ways of variance estimation approach.

## Results

In the dataset, there were a total of 536 observations. The mean age among all participants was 54.8 years with standard deviation 3.16. Out of 536 observations, 84 individuals had diabetes, 189 individuals had self-reported obesity status which equals 1, and 206 individuals had measured obesity status which equals 1. 215 participants have been told by a doctor or other health professional that he/she had hypertension (high blood pressure).

The results presented in this section are those of a logistic model, the model considered here adjusts for the obesity, age and hypertension:

$$logit(P_{diabetes} = 1) = \hat{\beta}_0 + \hat{\beta}_1 * obesity + \hat{\beta}_2 * RIDAGEYR + \hat{\beta}_3 * HBP. \quad (5.2)$$

With the misclassification matrix Π, following table shows the results:

*Table 28: Estimation results of estimator based on NHANES data with B = 200*

| Methods | (SE) | 95%CI |
|---|---|---|
| TRUE | 1.004(0.259) | (0.496, 1.512) |
| Naive | 1.068(0.253) | (0.572, 1.564) |
| McSimex.Q | 1.054(0.340) | (0.388, 1.719) |
| NONP | 1.001(0.336) | (0.342, 1.660) |
| NONP.CV.Var | 1.001(0.397) | (0.223, 1.779) |
| NONP.Q.Var | 1.001(0.336) | (0.342, 1.660) |

Again, the true method means the laboratory measured covariate was used in the GLM process in R, which means covariate "obesity" has been replaced by "mobesity" in the model 5.2. Similar to the true method, the naive method means the naïve (self-reported) covariate was used in the GLM process, which means covariate "obesity" has been replaced by "srobesity" in the model 5.2. 95%CI means 95% confidence interval of $\hat{\beta}_1$. The plots of 95% confidence intervals for all methods are shown in figure 1. Since overlaps of confidence intervals of estimators among all methods occur, the estimators are not significantly different from each other. Based on table 28 and figure 1, for the $\hat{\beta}_1$, estimator of our new method is closer to the estimator of true method compare to the MC-SIMEX method. We used nonparametric process in the new method, hence the estimated variances of estimator are larger than those from original MC-SIMEX method. The results in the above table can be interpreted as follows: The nonparametric MC-SIMEX method indicates that after adjusting for age and hypertension, we expect to see about 172% (which is exp^(1.001)-1) increase in the odds of having diabetes for an individual with obesity than an individual without obesity. The MC-SIMEX method estimate that about 187% (which equals exp^(1.054)-1) increase in the odds of having diabetes for an individual with obesity than an individual without obesity. Meanwhile, based on the true model, about 173% (which is exp^(1.004)-1) increasing was estimated. When we use the naive model, about 191% (which is exp^(1.068)-1) increasing was estimated.

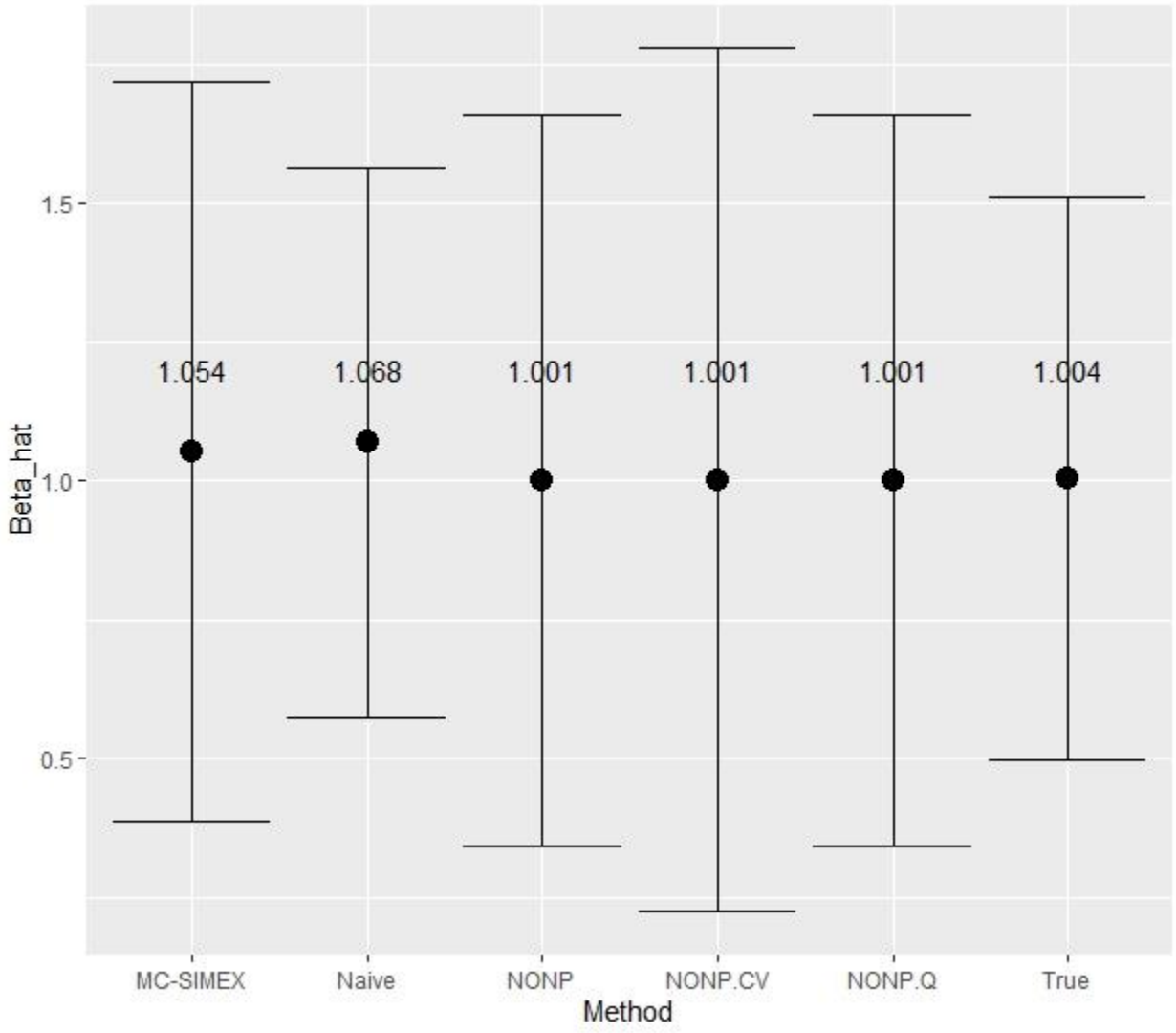


*Figure 1: Plot of the MCsimex, naive, new method and true method with their 95% confidence intervals.*

*The x-axis represents the type of estimator and the y-axis represents the estimator values*

# CHAPTER 6

# DISCUSSION AND CONCLUSION

Misclassification error is pervasive in medical and epidemiology research. There are a lot of literature on it, but we found that there are still some areas unexplored. One of them is to improve the accuracy of MC-SIMEX method estimator. In this dissertation, we aim to reduce the bias of MC-SIMEX method estimator by studying the effect of misclassification in logistic model. We found that the original MC-SIMEX method extrapolation function cannot approximate the true function in some situation. We proposed nonparametric MC-SIMEX method which use the fractional polynomial method to approximate the extrapolation function.

The simulation shows that the bias of MC-SIMEX method estimator is visible. It also showed that the improved MC-SIMEX method, nonparametric MC-SIMEX method, is a reasonable, general approximation with the same assumption but less biased estimator was created. The nonparametric MC-SIMEX method with fractional polynomial process and cross-validation process works very well in all considered setting in this dissertation. In addition, the results are consistence both for differential and nondifferential misclassification error on predictor.

This dissertation has certain limitations. First, in this dissertation, we only focus on a simple logistic model with two confounding predictors and a misclassified binary variable. This simple setting may not suffice. Second, we didn't consider missing observations in the new proposed method. However, there is few such perfect figures as in our study. Third, we only focus on misclassified predictor in this dissertation. But, misclassified binary outcome and misclassified multilevel variable are also common in statistical analysis. Finally, determining the most appropriate approach way of estimated variance is another challenge.

This dissertation opens up possibilities for future research in biostatistics. First, misclassified multilevel variable can be further explored. Second, unknown distribution of data is another challenge in the future. Third, considering that survival analysis is also a large category in statistics, expanding our new

methods to survival analysis with the censor data can be further proved. Finally, development of the new method in the model where multiple binary variables have misclassification errors could be evaluated.

# REFERENCES


Agogo, G. O., van der Voet, H., van't Veer, P., Ferrari, P., Leenders, M., Muller, D. C., . . . Boshuizen, H. (2014). Use of two-part regression calibration model to correct for measurement error in episodically consumed foods in a single-replicate study design: EPIC case study. *PLoS One, 9*(11), e113160. doi:10.1371/journal.pone.0113160

Akazawa, K., Kinukawa, N., & Nakamura, T. (1998). A note on the corrected score function adjusting for misclassification. *Journal of the Japan Statistical Society, 28*. doi:10.14490/jjss1995.28.115

Armstrong, B. (1985). Measurement error in the generalised linear model. *Communications in Statistics - Simulation and Computation, 14*(3), 529-544. doi:10.1080/03610918508812457

Armstrong, B. G. (1998). Effect of Measurement Error on Epidemiological Studies of Environmental and Occupational Exposures. *Occupational and Environmental Medicine, 55*(10), 651-656.

Bang, H., Chiu, Y.-L., Kaufman, J. S., Patel, M. D., Heiss, G., & Rose, K. M. (2013). Bias Correction Methods for Misclassified Covariates in the Cox Model: comparison of five correction methods by simulation and data analysis. *Journal of Statistical Theory and Practice, 7*, 381–400. doi:10.1080/15598608.2013.772830

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* CRC Press.

Carroll, R., & Stefanski, L. (1990). Approximate Quasi-likelihood Estimation in Models With Surrogate Predictors. *Journal of the American Statistical Association, 85*, 652-663.

CDC. (2017a, 9 15). *About the National Health and Nutrition Examination Survey*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

CDC. (2017b, 8 29). *About Adult BMI*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Chen, J., Hanfelt, J. J., & Huang, Y. (2015). A Simple Corrected Score for Logistic Regression with Errors-in-Covariates. *Communications in Statistics - Theory and Methods, 44*, 2024-2036. doi:10.1080/03610926.2013.773350

Cimbala, J. M. (2009). *Errors and Calibration*. Retrieved from The Pennsylvania State University, Department of Mechanical Engineering: https://www.me.psu.edu/cimbala/me345/Lectures/Errors_and_Calibration.pdf

Cole, S. R., Chu, H., & Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology, 35*, 1074-1081. doi:10.1093/ije/dyl097

Cook, J. R., & Stefanski, L. A. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association, 89*, 1314-1328. Retrieved from http://www.jstor.org/stable/2290994

Czepiel, S. A. (2002). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation.* Retrieved from Variables and Observations: https://czep.net/stat/mlelr.pdf

Edwards, J. K., Cole, S. R., Westreich, D., Crane, H., Eron, J. J., Mathews, W. C., . . . CNICS(2015). (2015). Multiple imputation to account for measurement error in marginal structural models. *Epidemiology (Cambridge, Mass.), 26*, 645–652. doi:10.1097/EDE.0000000000000330

Fraser, G., & Stram, D. (2001). Regression Calibration in Studies with Correlated Variables Measured with Error. *American Journal of Epidemiology, 154*, 836-844. doi:10.1093/aje/154.9.836

Gastwirth, J. L. (1987). The Statistical Precision of Medical Screening Procedures: Application to Polygraph and AIDS Antibodies Test Data. *Statistical Science, 2*(3), 213-222. Retrieved from http://www.jstor.org/stable/2245749

Hardin, J., Arnold, N., Schmiediche, H., & Carroll, R. (2003). The Simulation Extrapolation Method for Fitting Generalized Linear Models with Additive Measurement Error. *The Stata Journal, 3*, 1-12. doi:10.1177/1536867X0300300407

Hilbe, J. M. (2017). *Logistic Regression Models, 1st Edition.* Chapman and Hall/CRC.

Hogg, R. V., Craig, A. T., & McKean, J. W. (2005). *Introduction to Mathematical Statistics, 6th Edition.* Pearson.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition.* John Wiley \& Sons, Inc.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition.* John Wiley \& Sons, Inc.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics, 62*, 85-96. doi:10.1111/j.1541-0420.2005.00396.x

Lederer, W., & Küchenhoff, H. (2006). A short introduction to the SIMEX and MCSIMEX. *R news, 6*, 26-31.

Mayer, B., Keller, F., Syrovets, T., & Wittau, M. (2013). Estimation of half-life periods in nonlinear data with fractional polynomials. *Statistical Methods in Medical Research, 25*(5), 1791-1803. doi:10.1177/0962280213502403

Millner, A., Lee, M., & Nock, M. (2015). Single-Item Measurement of Suicidal Behaviors: Validity and Consequences of Misclassification. *PLOS ONE, 10*(10), e0141606. doi:10.1371/journal.pone.0141606

Nikolaeva, R., Bhatnagar, A., & Ghose, S. (2015). Exploring Curvilinearity Through Fractional Polynomials in Management Research. *Organizational Research Methods, 18*, 738-760. doi:10.1177/1094428115584006

Nikolaeva, R., Bhatnagar, A., & Ghose, S. (2015). Exploring Curvilinearity Through Fractional Polynomials in Management Research. *Organizational Research Methods, 18*(4), 738-760. doi:10.1177/1094428115584006

Pina-Sánchez, J. (2016). Adjustment of Recall Errors in Duration Data Using SIMEX. *Metodološki Zvezki -- Advances in Methodology and Statistics, 13*.

Qi, L., Wang, Y.-F., & He, Y. (2010). A Comparison of Multiple Imputation and Fully Augmented Weighted Estimators for Cox Regression with Missing Covariates. *Statistics in medicine, 29*, 2592-2604. doi:10.1002/sim.4016

Rohatgi, V. K., & Saleh, A. K. (2000). *An Introduction to Probability and Statistics, Second Edition.* John Wiley & Sons, Inc.

Rosner, B., Spiegelman, D., & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology, 132*, 734-745. doi:10.1093/oxfordjournals.aje.a115715

Rosner, B., Willett, W. C., & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine, 8*, 1051-1069. doi:10.1002/sim.4780080905

Royston, P., & Altman, D. G. (1994). Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 43*, 429-467. Retrieved from http://www.jstor.org/stable/2986270

Royston, P., Ambler, G., & Sauerbrei, W. (1999). The Use of Fractional Polynomials in Multivariable Regression Modelling. *International Journal of Epidemiology, 28*, 964-974.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592. doi:10.1093/biomet/63.3.581

Sauerbrei, W., & Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 162*, 71-94. doi:10.1111/1467-985X.00122

Sevilimedu, V. (2017). *Application of the Misclassification Simulation Extrapolation (Mc-Simex) Procedure to Log-Logistic Accelerated Failure Time (AFT) Models In Survival Analysis.* Retrieved from Georgia Southern University, Electronic Theses and Dissertation: https://digitalcommons.georgiasouthern.edu/etd/1659/

Slate, E. H., & Bandyopadhyay, D. (2009). An investigation of the MC-SIMEX method with application to measurement error in periodontal outcomes. *Statistics in medicine, 28*, 3523–3538. doi:10.1002/sim.3656

Spiegelman, D., Carroll, R. J., & Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in medicine, 20*(1), 139-160.

Taylor, J. R. (1997). *Introduction To Error Analysis: The Study of Uncertainties in Physical Measurements 2nd Edition.* University Science Books.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica, 50*(1), 1-25. Retrieved from http://www.jstor.org/stable/1912526

White, I. R. (2006). Commentary: Dealing with measurement error: multiple imputation or regression calibration? *International Journal of Epidemiology, 35*, 1081-1082. doi:10.1093/ije/dyl139

Zhang, Z. (2016). Multivariable fractional polynomial method for regression model. *Annals of translational medicine, 4*(9), 174. doi:10.21037/atm.2016.05.01

Zucker, D. M., & Spiegelman, D. (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in medicine, 27*, 1911–1933. doi:0.1002/sim.3159