University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Open Access Dissertations

5-2009

Correction Methods, Approximate Biases, and Inference for Misclassified Data

Meng-Shiou Shieh
University of Massachusetts Amherst, msshieh@math.umass.edu

CORRECTION METHODS, APPROXIMATE BIASES, AND INFERENCE FOR MISCLASSIFIED DATA

A Dissertation Presented

by

MENG-SHIOU SHIEH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2009

Department of Mathematics and Statistics

CORRECTION METHODS, APPROXIMATE BIASES, AND INFERENCE FOR MISCLASSIFIED DATA

A Dissertation Presented

by

MENG-SHIOU SHIEH

Approved as to style and content by:

John Staudenmayer, Chair

John Buonaccorsi, Member

Erin Conlon, Member

Andrea Foulkes, Member

George Avrunin, Department Head
Mathematics and Statistics

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Professor John Staudenmayer who has had infinite patience with me. His encouragement has been the backbone to support me to finish my dissertation. Whenever I had questions, he always responded quickly, and answered with a more intuitive idea and gave me a deeper understanding. When I made a serious mistake, John always responded in a humorous way and which made it not such a big deal and showed that I can learn from my mistakes. John not only taught me the knowledge of statistics, he also taught me how to think more intuitively about statistics. I will miss those times that John popped in at my door and his Red Sox talk.

I also want to thank Professor John Buonaccorsi who has a lot of knowledge in measurement error and who always answered my questions under his heavy duty schedule and gave me more knowledge than what I needed. John, thanks for always having an open door for me.

I would like to thank Professor Michael Lavine for letting me participate in projects. Joining projects allowed me to gain more knowledge and pushed me onto the right track to finish my work. Thanks also to Professor Anna Liu, Ben Letcher of the Silvio O. Conte Anadromous Fish Research Center, and to all my project partners: David, Sydne, Hugo and Enrique.

I would like to thank Professor Erin Conlon and Professor Andrea Foulkes, who at the last minute stepped in to be my committee members. I would also like to thank Professor H.K. Hsieh who constantly reminded me to finish my dissertation.

I should say a big thanks to the UMass library. The fast interlibrary loan service made my research work smooth, not delayed by missing material. I would also like to

thank the staff of RCF, and Professor Hans Johnson for letting me use abacus. Without that facility, my simulation would still be running.

I would like to thank the many people that I encountered who answered my little technical problems in LaTeXor other things. I would also like to thank the administrative staff in the Department of Mathematics and Statistics, who are so kind and answer my questions or direct me what to do.

Finally, I would like to thank my family. My husband, Peter, has been so supportive of me going back to school to get my degree. He takes care of children, and fixed up our dinner when our dog Venus ate our chicken. Peter, thanks for tolerating me when I am stressed out and helping me with LaTeX. Thanks to my children, Sonia, Conlan and Kanya. Thanks for understanding that Mommy had to study and sometimes could not participate in your activities. I would like to thank my parents Jing-Chun and Pao-Chin who supported me to come to America to study and waited for me for such long time to get this degree. Also I want to say thanks to my parents-in-law, Mary and Garry Brown. When I came to America, my goal was to get a Ph.D. degree, and their support helped me to achieve this goal.

I would like to dedicate my dissertation to my parents, and parents-in-law, Jing-Chun, Pao-Chin, Mary and Garry.

ABSTRACT

CORRECTION METHODS, APPROXIMATE BIASES, AND INFERENCE FOR MISCLASSIFIED DATA

May 2009

Meng-Shiou Shieh, B.A., Fu Jen Catholic University

M.A.,Syracuse University

M.A., University of Massachusetts, Amherst

Ph.D., University of Massachusetts, Amherst

Directed by Professor John Staudenmayer

When categorical data are misplaced into the wrong category, we say the data is affected by misclassification. This is common for data collection. It is well-known that naive estimators of category probabilities and coefficients for regression that ignore misclassification can be biased. In this dissertation, we develop methods to provide improved estimators and confidence intervals for a proportion when only a misclassified proxy is observed, and provide improved estimators and confidence intervals for regression coefficients when only misclassified covariates are observed.

Following the introduction and literature review , we develop two estimators for a proportion , one which reduces the bias, and one with smaller mean square error. Then we will give two methods to find a confidence interval for a proportion, one using optimization techniques, and the other one using Fieller's method. After that, we will focus on developing methods to find corrected estimators for coefficients of regression with misclassified covariates, with or without perfectly measured covariates, and with a known estimated misclassification/reclassification model. These correction methods use the score function approach, regression calibration and a mixture model. We also use Fieller's method to find a confidence interval for the slope of simple regression with with misclassified binary covariates. Finally, we use simulation to demonstrate the performance of our proposed methods.

TABLE OF CONTENTS

CHAPTER

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

We begin with a brief general introduction to the problems of interest and terminology. A description of the contents of the thesis and a more complete literature review follow.

1.1 Introduction

In studies involving categorical data, there is always the possibility that the data may be affected by misclassification, which occurs when an observed category does not match the true category. This is common for data collected from surveys, or error-prone measurements. A specific example of misclassified data comes from a case control study of prescription of antibiotics during pregnancy and subsequent occurrence of Sudden Infant Death Syndrome (SIDS) (Greenland, 1988, 2008). This data includes a main study data among women from whom only interview data were examined, and a seperate data set from a validation study from medical records. The interview data is subject to misclassification.

A second motivating example concerns the estimation of how physically active a person is. A metabolic equivalent (MET) is a measure of a person's physical activity level at a given point in time. It is defined as the ratio of a person's metabolic rate to her resting metabolic rate, where the resting metabolic rate is defined as consuming 3.5 mL O^2/ kg of body weight / minute. It is believed that the amount of time spent at > 3 METs (moderate activity) has important health implications (Pate et al, 1995), but it is difficult to measure the fraction of time someone spends above 3 METs accurately, precisely, and cheaply outside a lab or without burdensome equipment (Sirard and Pate, 2001). One method to address the problem involves affixing an accelerometer that records evidence of motion on a dense time scale to a person's hip; other methods, such as surveys, calorimetry, and doubly labeled water, are reviewed in Levine (2005), for instance. Accelerometer data, known as counts, can then be processed in one of a number of ways to estimate a person's energy expenditure over time, and this is an ongoing area of research (Pober et al, 2006). One simple and widely used processing method relates the total accelerometer counts in a minute to the average METs in the minute with linear regression (Freedson, Melanson, and Sirard, 1998). That relation then can define cutpoints to classify each minute into two categories ≤ 3 METs (sedentary or light activity) > 3 METs (at least moderately active). These binary data can be subject to misclassification.

It is well-known that naive estimators of category probabilities and coefficients for regression that ignore misclassification can be biased. Suppose $X_i, i = 1, \ldots, n$ are i.i.d. discrete random variables, each with K categories. Let $\pi_j = P(X_i = j)$, and $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_{K-1})^{\mathsf{T}}$. Instead of observing X_i, we observe W_i. Let $\lambda_j = P(W_i = \lambda_j)$ and $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_{K-1})^{\mathsf{T}}$. The relationship between X_i and W_i can be defined in one of two ways: through a misclassification model, $\mathbf{P}$, or a reclassification model, $\mathbf{Q}$. The misclassification model is $\mathbf{P} = (\theta_{lm})_{l=0,\ldots,K-1,m=0,\ldots,K-1}$, where $\theta_{lm} = P(W_i = l | X_i = m)$, and $\boldsymbol{\lambda} = \mathbf{P}\boldsymbol{\pi}$. The reclassification model is $\mathbf{Q} = (\gamma_{lm})_{l=0,\ldots,K-1,m=0,\ldots,K-1}$, where $\gamma_{lm} = P(X_i = l | W_i = m)$, and $\boldsymbol{\pi} = \mathbf{Q}\boldsymbol{\lambda}$. Throughout, we use zero-based indexing for matrices, so that a $K \times K$ matrix has elements $\begin{pmatrix} a_{00} & \cdots & a_{0(k-1)} \\ \ldots & \ddots & \ldots \\ a_{(K-1)0} & \cdots & a_{(K-1)(K-1)} \end{pmatrix}$. This is so we can use notation like θ_{00} to indicate the values of X, W etc. Reclassification models are analogous to a Berkson model in general measurement error models (a general reference for measurement error terms is Carroll et al., 2006), and misclassification models are analogous to a classical model. In either case, knoowledge of the misclassification or reclassification model can help reduce the bias in estimators of $\boldsymbol{\pi}$ or regression coefficients when W_i is observed. Either of those models can be estimated from validation data.

Most correction methods for misclassification in research require auxiliary data or some knowledge about misclassification or reclassification matrices. One exception is Li et al. (2004), who assume the surrogate has a Poisson distribution. In that case, the mean and variance of a Poisson distribution are the same, and the true parameters are recoverable without additional data or known misclassification/reclassification model.

There are four common types of auxiliary data to adjust for the bias due to misclassification. They are internal/external validation data (X is observable directly), replicated values (replicates of W are available) and instrumental variables (another available S is observable in addition to W). External validation data are independent observations from the main study, but we have to make sure the external validation data are transportable across different study populations. Internal validation data is also known as two stage or doubling sampling (Tenenbein, 1970). Through repeated measures of the surrogate we can recover the misclassification probabilities if there is no identifiability problem (Harper, 1964, Hui and Walter, 1980, White et al., 2001). Quade et al. (1980), Walter and Irwig (1988) present the expectation-maximization (EM) algorithm to recover misclassification probabilities using replicated data. Walter and Irwig (1988) also review how to use replicated data to recover misclassification probabilities in various designs. Data collection practicalities sometimes determine whether a reclassification or misclassification validation sample can be collected. Reclassification based models are often more efficient.

Let Y be the response variable. The measurement error model of W given X is non-differential if the distribution of Y given (X, W) is the same as the distribution of Y given X, where W is the observed value for X. Otherwise, it is called differential mismeasurement. Non-differential mismeasurement means the misclassification probabilities do not depend on Y. For example, Greenland (1998) studies the association of antibiotic use in mothers during pregnancy (X) and sudden infant death syndrome (SIDS) which is Y. The observed data (W) is self-reported by the mother. If $P(W|X, Y)$ does not depend on Y, the misclassification is non-differential. Otherwise, it is differential.

The odds ratio and relative risk are very important in epidemiology studies. If an event E has probability $P(E)$, the odds of the event is $P(E)/(1 - P(E))$. In general if two events E_1 and E_2 have respective probabilities $P(E_1)$ and $P(E_2)$, the odds ratio comparing E_1 with E_2 is the ratio of the odds of E_1 to the odds of E_2. In case-control studies, let $Y(= 0, 1)$ denote disease status and $X(= 0, 1)$ denote exposure status. The odds ratio is

$$\frac{P(Y = 1|X = 1)(1 - P(Y = 1|X = 0))}{P(Y = 1|X = 0)(1 - P(Y = 1|Y = 0))}.$$

The relative risk is defined as

$$\frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}.$$

If we observe W instead of X, then the naive estimators of odds ratio and relative risk are biased, and we would need some information about misclassification probabilities to do correction.

1.2 Purpose of Thesis Contents

The purposes of this dissertation are:

1. To provide improved estimators and confidence intervals for $P(X_i = j)$ when only a misclassified proxy for X_i is observed, and

2. to provide improved estimators and confidence intervals for regression coefficients when only misclassified covariates are observed.

The main focus of this dissertation is to account for the effects of an estimated misclassification model.

Following this introduction, we review literature on correcting for misclassification when the observed data are misclassified and when misclassified data are used as covariates in regression models. Chapter 2 will consider the case when $\boldsymbol{\pi}$ is of interest. We will discuss existing estimators of $\boldsymbol{\pi}$ and introduce some estimators which are not unbiased but have smaller mean square errors. In Chapter 3, we will develop confidence intervals for π in the case when $K = 2$. In Chapter 4, we will focus on regression models with misclassified covariates. We present a correction method when one of $\mathbf{P}$ or $\mathbf{Q}$ is known, and explore the approximate bias of that method when $\mathbf{P}$ or $\mathbf{Q}$ is estimated from external data. In Chapter 5, we use simulation to evaluate the performance of the methods described above, and in Chapter 6 we will present our conclusions.

1.3 Literature Review: Misclassified Categorical Data

The problem of misclassified categorical data has been considered for over 50 years. An early reference is Bross (1954) who discusses the biases caused by misclassification in binary data. Bross (1954) shows that when misclassification is ignored, the estimated difference between two proportions of interest of two different populations (e.g., case and control) is biased toward the null of no difference, the significance level is correct if both populations have the same misclassification probabilities (non-differential) and power is reduced. Tests about the difference are discussed further by Rubin et al. (1956), Katz (1979), and Zellen and Haitozsky (1991) for the binary case, and Mote and Anderson (1965) for the multinomial case. In these articles, it is shown that the power of the χ^2 test will decrease under misclassification (differential or non-differential), and the false positive probabilities ($P(W = 1|X = 0)$), the false positive probability in a differential misclassification model, for two populations plays an important role in how much the power reduces when π is less than 0.5. Reducing the false positive probability can improve the efficiency of the test if π is small. When π is big, the roles of false-positive and false-negative will switch. Schwartz (1985) also states the bias of the naive estimator and how the misclassification probabilities affect the coverage probability of conventional confidence intervals for misclassified binary data.

Kuha et al. (1998) give a concise summary of the development of correction methods for misclassified data in different epidemiological models. Quade et al. (1980) present the bias of a naive estimator and the bias of the estimator that treats estimated misclassification probabilities as known. Chen (1989) presents a review of methods for misclassified categorical data in epidemiology, and also shows that the usual misclassification models are a subclass of log-linear models. Tenenbein (1970) uses double sampling (i.e. internal validation data) to get a maximum likelihood estimator for the true proportion and the asymptotic variance for misclassified binomial data. Espeland and Hui (1987) demonstrate how to model misclassified data with validation data as an incomplete data problem using a log-linear model and estimate using the Fisher scoring algorithm. Barron (1977) uses the misclassification model, also known as the matrix method (Morrissey and Spiegelman, 1999) or indirect method (Marshall, 1990), to obtain unbiased estimators from misclassified 2×2 table data, and he uses the results to correct relative risk estimates. This work assumes the misclassification model is known. Marshall (1990) compares the relative efficiency of the direct method (also known as reclassification or inverse matrix method) with the indirect method and shows that the direct method is more efficient than the indirect method. Greenland (1988) derives the variance of corrected estimators when the misclassification model is estimated from external or internal validation data. In van den Hout and van der Heijden (2002), it is shown that under known misclassification probabilities, the maximum likelihood estimator (MLE) and moment estimator are the same if the moment estimator is in the interior of the parameter space.

Other articles related to misclassified data itself include Copeland et al. (1977) and Hofler (2005), who discuss the bias of relative risk for misclassified data when misclassification is non-differential or differential. Gladen and Rogan (1979) show that the power of statistical tests about relative risk is reduced when data are affected by misclassification. Morrissey and Spiegelman (1999) compare the matrix method and inverse matrix method to correct estimates of the odds-ratio of misclassified binary data. They conclude that the inverse method estimator performs better. Lyles (2002) points out that the inverse matrix estimator used in Morrissey and Spiegelman (1999) is the MLE under differential misclassification. Greenland (2008) shows that the matrix method estimators in Barron (1997) and Greenland (1998) are MLEs under the assumptions given by those authors. Selén (1986) uses a matrix method to correct estimates of group means for misclassified data and derives the variance of the corrected estimator.

1.4 Literature Review: Misclassified Covariates in Regression

For the regression model, Fuller (1987) and Carroll et al. (2006) are book length reviews of measurement error models, and Carroll (1998) has a summary for epidemiologists. Most of the literature that provides correction methods for regression models with mismeasured covariates focuses on continuous cases, and does not apply generally to categorical data. Cochran (1968) shows how to model binary misclassified data from a measurement error model perspective. This work also shows that, in some simple situations, binomial misclassified data can be modeled with a non-standard type of continuous measurement error.

It is well known that the coefficient estimators usually are inconsistent for regression models when discrete covariates are misclassified. Christopher and Kupper (1995) study the bias of the least squares estimator in multiple linear regression models with misclassified covariates, perfectly measured covariates and a known reclassification model.

They also explore the impact on certain test statistics and show that misclassification will cause the power of such tests to be reduced. In the situation of continuous mismeasured coavariates, it is known that the naive coefficient estimates corresponding to the perfectly measured covariates will be unbiased if the mismeasured covariates and the perfectly measured covariates are uncorrelated (Carroll et al., 1985). Buonaccorsi et al. (2005) prove this is also true for misclassified covariates. Davidov et al. (2003) study the effect of misclassification on the parameters of a logistic regression with misclassified binary covariates and Veierød and Laake (2001) derive the bias for Poisson regression with misclassified and perfectly measured covariates.

Common correction methods for regression include the method of moments, likelihood methods, regression calibration, simulation extrapolation (SIMEX), estimating equation approaches and Bayesian approaches. Reade-Christopher and Kupper (1991) study logistic and log-linear regression with misclassified covariates. They use known or estimated reclassification models and use maximum likelihood to get a naive estimator, then follow the method of moments to perform the correction. Spiegelman et al. (2000) present likelihood-based computational strategies for logistic regression with both covariate measurement error and reclassification models on one or more covariates. Linear regression with misclassified covariates and a known misclassification model is considered by van den Hout and Kooiman (2006). They use the idea in Spiegelman et al. (2000) and implement the EM algorithm to find corrected estimators.

When it is hard to find the maximum of a likelihood that involves many parameters, a pseudo likelihood method can be used. Gong and Samanjego (1981) define pseudo maximum likelihood estimation and get the asymptotic distribution of pseudo MLE. A pseudo method estimates some parameters from validation data first, then treats those parameters as known and finds the maximum likelihood estimators for the remaining parameters. Parke (1986) has a simpler expression for the asymptotic variance of a pseudo MLE. Liu and Liang (1991) use quasi-likelihood scores and the pseudo approach for generalized linear models with only categorical covariates, misclassified or not, and non-differential misclassification. They derive the variance for the corrected estimator that accounts for the variation due to estimation of the parameters in the misclassification model. They estimate the misclassification parameters from replicate data and discuss how many replicates are needed to reach a desired efficiency.

Rosner et al. (1989) apply regression calibration, data imputation and likelihood approximation methods to logistic regression with a mismeasured covariate. Frost and Thompson (2000) compare moment-based and regression methods to correct the correction factor (the inverse of the correction factor is the attenuating factor or reliability ratio (Carroll et al. 2006)) of a simple regression slope with a mismeasured covariate, and the simulations show that the moment-based method performed better. White et al. (2001) demonstrate how to use replicated data to correct using regression calibration in the regression model with measurement error in binary and continuous covariates.

Nakamura (1990) proposes a corrected score approach. This work develops a score function whose conditional expectation given the response and true covariates is the usual log likelihood based on the response and the unknown true covariates. This article also includes a proof that the solution of a corrected score function is a consistent estimator under some regularity conditions. An unbiased score function is a score function whose expectation is zero at the true parameter, and it is not necessarily based on the likelihood function. We should note that an unbiased score function is not necessarily a corrected score function. Akazawa et al. (1998), prove that a corrected score function always exists for a regression model with misclassified covariates, but it does not necessarily exist in the case of mismeasured continuous covariates. The existence of a corrected score function assumes the misclassification matrix is known. Recently, Zucker and Speigelman (2008) apply the idea of Akazawa et al. (1998) to a hazard model with misclassified covariates. Stefanski and Carroll (1987) study conditional score estimators for the generalized linear model, and they obtain an unbiased score function by conditioning on sufficient statistics. Buonaccorsi (1996) uses a modified estimating equation approach which can be applied when the measurement error variances and covariances differ across units. The measurement error variances and covariances for this approach can be known or estimated.

Cook and Stefanski (1994) propose the idea of the simulation-extrapolation (SIMEX) correction method for the measurement error model. The method adds more measurement error to the mismeasured variables (covariates or response), then gets regression parameters corresponding to the extra error, studies the trend of the parameters, and then extrapolates this trend back to get a SIMEX estimator of the parameter of interest. Küchenoff et al (2006) develops an innovative SIMEX approach for misclassified data (MC-SIMEX). They assume that the misclassification matrix $\mathbf{P}$ is known or can be estimated from validation data. The way they add extra errors to the misclassified data is interesting: in each simulation step ξ, they construct a new misclassification matrix $\mathbf{P}^{\xi} = \mathbf{E}\mathbf{D}^{\xi}\mathbf{E}^{-1}$ where $\mathbf{D}$ is the diagonal matrix of eigenvalues of $\mathbf{P}$, and $\mathbf{E}$ is the corresponding matrix of eigenvectors. Note that $\mathbf{P}^{1+\xi} = \mathbf{P}\mathbf{P}^{\xi}$ and we can simulate misclassified data using the observed as true data and $\mathbf{P}^{\xi}$ as the misclassification rule. Küchenoff et al (2007) derives the asymptotic variance estimators for the MC-SIMEX estimator when the misclassification model is estimated from external data. MC-SIMEX can also apply to prevalence estimation.

Gustafson (2004) has a general discussion about the Bayesian method for epidemiological data with mismeasurement error and misclassification. Stamey et al. (2007) and Perez et al. (2007) use a Bayesian approach to address misclassified multinomial/binomial data. Stamey et al. (2007) compare Bayesian estimation of an intervention effect with the maximum likelihood estimators in Lin et al. (2005), and they find the Bayesian estimator's coverage is better. Perez et al. (2007) provide a Bayesian method for multinomial data with misclassification. Prescott and Garthwaite (2002) use a two-stage Bayesian method for the odds-ratio of a case-control study, and compare their method with the methods of Morrissey and Spiegelman (1999).

In the Bayesian approach, the unobserved true value is latent. Kuha (1997) uses data augmentation in generalized linear models with mismeasured covariates and misclassified covariates. Stephens and Dellaportas (1992) apply a Bayesian method to generalized linear models with mismeasured covariates. Müller and Roeder (1997) use a Dirichlet process prior on the joint distribution of covariates and the true unobserved variable, and they use a Gibbs sampling scheme to estimate the parameters of a logistic regression with a mismeasured covariate or a misclassified covariate and some validation data.

CHAPTER 2

ESTIMATORS FOR π

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed binary (0 or 1) random variables with $\pi = P(X_i = 1)$. Instead of observing X_i, we observe W_i. There are two common ways to describe the relationship between W_i and X_i. One is a misclassification model where $P(W_i = 1|X_i = 1) = \theta_{11}$ (sensitivity), and $P(W_i = 0|X_i = 0) = \theta_{00}$ (specificity) which is similar to the classical model for additive measurement error. The reclassification model approach is analogous to a Berkson model and specifies $P(X_i = 1|W_i = 1) = \gamma_{11}$ (positive predictive value) and $P(X_i = 0|W_i = 0) = \gamma_{00}$ (negative predictive value) .The bias of the naive estimator, $\widehat{\pi}_{naive} = \sum_{i=1}^{n} W_i/n$, can be shown to be $\pi(\theta_{00}+\theta_{11}-2)-\theta_{00}+1$ or $\frac{\pi(2-\gamma_{00}-\gamma_{11})+\gamma_{00}-1}{\gamma_{00}+\gamma_{11}-1}$ expressed in terms of the misclassification and reclassification models respectively. One consequence of these bias expressions is that a misclassification or reclassification parameter closer to one (the case of no error of a particular type) can actually result in more bias. See Section 2.2 for more discussion of bias. Recent reviews of this problem, including important extensions to two by two tables and odds ratios, can be found in Greenland (2008), van den Hout and van der Heijden (2002), Chen (1999), and Kuha et al. (1998). The Bayesian approach is discussed in Gustafson (2004, Chapter 5) and Prescott and Garthwaite (2002). We take a relative frequentist approach.

With either a known misclassification model or a known reclassification model, the respective bias expressions can be used to develop unbiased method of moments estimators of π. In the case of the misclassification model the estimator is:

$$\widehat{\pi}_{corrected,M} = \frac{\widehat{\pi}_{naive} + \theta_{00} - 1}{\theta_{00} + \theta_{11} - 1},$$

and in the case of the reclassification model it is:

$$\widehat{\pi}_{corrected,R} = (\gamma_{00} + \gamma_{11} - 1)\widehat{\pi}_{naive} - \gamma_{00} + 1.$$

The first correction method is the matrix method, and the second correction method is the inverse matrix method (Morissey and Spiegelman, 1999). Zelen and Haitovsky (1991) point out that the true and observed values have a positive correlation when $\theta_{00} + \theta_{11} - 1 > 0$.

If $\widehat{\pi}_{corrected,M}$ is modified by making it zero or one if $\widehat{\pi}_{corrected,M} < 0$ or $\widehat{\pi}_{corrected,M} > 1$ respectively, then the resulting estimator is also a maximum likelihood estimator (van den Hout and van der Heijden, 2002, Section 5). $\widehat{\pi}_{corrected,R}$ is also a maximum likelihood estimator (Lyles, 2002).

When the misclassification and reclassification parameters are known, the reclassification estimator generally has a smaller variance than the misclassification estimator since $|\theta_{00} + \theta_{11} - 1| < 1$ and $|\gamma_{00} + \gamma_{11} - 1| < 1$. On the other hand, the misclassification model can be estimated from a validation sample that is designed to contain a fixed number of $X = 1$ and $X = 0$ cases. That is the situation in which we are primarily interested, but when a reclassification model is available, $\widehat{\pi}_{corrected,R}$ should be used.

In the typical case when the misclassification model is unknown, it needs to be estimated. We consider the case of external validation data where W_i is observed $N_{.0}$ when $X_i = 0$ and $N_{.1}$ times when $X_i = 1$. With N_{jj} denoting the number of times $w_i = x_i = j, j = 0, 1$ in each sample, estimators of θ_{00}, θ_{11} are $\widehat{\theta}_{00} = \frac{N_{00}}{N_{.0}}$ and $\widehat{\theta}_{11} = \frac{N_{11}}{N_{.1}}$. Note that in these data we do not require the relative frequencies of $X = 0$ or $X_i = 1$ to have any connection to $Pr(X_i = 0)$ or $Pr(X_i = 1)$. As a result, these validation data cannot be used to estimate a reclassification model. Using the estimates of θ_{00} and θ_{11} above, an estimator of π is $\widehat{\pi}_{PlugIn} = \dfrac{\widehat{\pi}_{naive} + \widehat{\theta}_{00} - 1}{\widehat{\theta}_{00} + \widehat{\theta}_{11} - 1}$. If estimates of γ_{00} and γ_{11} are available, then the plugin version of the reclassification estimator is $\widehat{\pi}_{PlugIn,Re} = (\widehat{\gamma}_{00}+\widehat{\gamma}_{11}-1)\widehat{\pi}_{naive}-\widehat{\gamma}_{00}+1$. It can be shown that these are maximum likelihood estimators from the n main study and $N_{.0}$ and $N_{.1}$ validation study data points. Jensen's inequality can be used to show that $\widehat{\pi}_{PlugIn}$ is biased when $N_{.0}$ and $N_{.1}$ are finite. We develop two novel alternatives to these estimators next. One is a bias reduced version of $\widehat{\pi}_{PlugIn}$ (Section 2.1), and the other is a "partially corrected" estimator in Section 2.2.

Section 2.1 will be devoted to the bias reduced estimator, and section 2.2 is the partially corrected estimator, both for the binary case. In section 2.3, we will generalize these results to the case of categorical data with k categories.

2.1 Bias Reduced Estimator

In this section, we will discuss the bias of $\widehat{\pi}_{PlugIn}$. We begin with the following theorem that approximates the bias of $\widehat{\pi}_{PlugIn}$ as a function of the validation sample size.

Theorem 2.1.1 *Assume $\theta_{00}+\theta_{11}-1 \neq 0$, and $\theta^*_{00}+\theta^*_{11}-1 \neq 0$ for all $\theta^*_{00}, \theta^*_{11}$ in the rectangular box formed by $\theta_{00}, \widehat{\theta}_{00}$ and $\theta_{11}, \widehat{\theta}_{11}$, then*

$$E\left(\widehat{\pi}_{PlugIn}\right) = \pi + Var(\widehat{\theta}_{00})\frac{(\pi_{naive}-\theta_{11})}{(\theta_{00}+\theta_{11}-1)^3} + Var(\widehat{\theta}_{11})\frac{(\pi_{naive}+\theta_{00}-1)}{(\theta_{00}+\theta_{11}-1)^3} + O(min(N_{.1}, N_{.0})^{-2}).$$

Proof Let $f(\pi_{naive}, \theta_{00}, \theta_{11}) = \frac{\pi_{naive}+\theta_{00}-1}{\theta_{00}+\theta_{11}-1}$. Then from Taylor's expansion, we will have:

$$\begin{aligned} f(\widehat{\pi}_{naive}, \widehat{\theta}_{00}, \widehat{\theta}_{11}) &= f(\pi_{naive}, \theta_{00}, \theta_{11}) + \nabla f(\pi_{naive}, \theta_{00}, \theta_{11})^{\mathsf{T}} \begin{pmatrix} \widehat{\pi}_{naive}-\pi_{naive} \\ \widehat{\theta}_{00}-\theta_{00} \\ \widehat{\theta}_{11}-\theta_{11} \end{pmatrix} \\ &\quad + \begin{pmatrix} \widehat{\pi}_{naive}-\pi_{naive} \\ \widehat{\theta}_{00}-\theta_{00} \\ \widehat{\theta}_{11}-\theta_{11} \end{pmatrix}^{\mathsf{T}} \frac{\nabla^2 f(\pi_{naive}, \theta_{00}, \theta_{11})}{2} \begin{pmatrix} \widehat{\pi}_{naive}-\pi_{naive} \\ \widehat{\theta}_{00}-\theta_{00} \\ \widehat{\theta}_{11}-\theta_{11} \end{pmatrix} \\ &\quad + R_3(\widehat{\pi}_{naive}, \widehat{\theta}_{00}, \widehat{\theta}_{11}), \end{aligned}$$

where

$$R_3(\widehat{\pi}_{naive},\widehat{\theta}_{00},\widehat{\theta}_{11}) = \sum_{i_1+i_2+i_3=3} \frac{\partial^3 f(\widetilde{\pi}_{naive},\widetilde{\theta}_{00},\widetilde{\theta}_{11})}{3!\partial\pi_{naive}^{i_1}\partial\theta_{00}^{i_2}\partial\theta_{11}^{i_3}}(\widehat{\pi}_{naive}-\pi_{naive})^{i_1}(\widehat{\theta}_{00}-\theta_{00})^{i_2}(\widehat{\theta}_{11}-\theta_{11})^{i_3},$$

with $\widetilde{\theta}_{00}$ between θ_{00} and $\widehat{\theta}_{00}$, $\widetilde{\theta}_{11}$ between θ_{11} and $\widehat{\theta}_{11}$. Since $\frac{\partial^2 f}{\partial\pi_{naive}^2} = 0$, $\widehat{\theta}_{11},\widehat{\theta}_{00}$ are uncorrelated, $\left|\frac{\partial^3 f(\widetilde{\pi}_{naive},\widetilde{\theta}_{00},\widetilde{\theta}_{11})}{\partial\pi_{naive}^{i_1}\partial\theta_{00}^{i_2}\partial\theta_{11}^{i_3}}\right| < M$ for some M since the third order partial derivatives of f are continuous on a closed region, and $E(\widehat{\theta}_{ii}-\theta_{ii})^3 = \dfrac{2\theta_{ii}^3 - 3\theta_{ii}^2 + \theta_{ii}}{N_{.i}^2}, i = 0, 1,$ we will have

$$\left|E\left\{f(\widehat{\pi}_{naive},\widehat{\theta}_{00},\widehat{\theta}_{11})\right\} - \left\{f(\pi_{naive},\theta_{00},\theta_{11}) + \frac{\partial^2 f(\pi_{naive},\theta_{00},\theta_{11})}{2\partial\theta_{00}^2}\,\text{Var}(\widehat{\theta}_{00}) + \frac{\partial^2 f(\theta_{00},\theta_{11},\pi_{naive})}{2\partial\theta_{11}^2}\,\text{Var}(\widehat{\theta}_{11})\right\}\right| \leq 12M\text{min}(N_{.1},N_{.0})^{-2},$$

which yields

$$E\left(\widehat{\pi}_{PlugIn}\right) = \pi + \text{Var}(\widehat{\theta}_{00})\frac{(\pi_{naive}-\theta_{11})}{(\theta_{00}+\theta_{11}-1)^3} + \text{Var}(\widehat{\theta}_{11})\frac{(\pi_{naive}+\theta_{00}-1)}{(\theta_{00}+\theta_{11}-1)^3} + O(\text{min}(N_{.1},N_{.0})^{-2}).$$

As a consequence of the preceding result, we can create a first order bias corrected estimator:

$$\widehat{\pi}_{corrected,PI} = \widehat{\pi}_{PlugIn} - \frac{\widehat{\theta}_{00}(1-\widehat{\theta}_{00})(\widehat{\pi}_{naive}-\widehat{\theta}_{11})}{N_{.0}(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^3} - \frac{\widehat{\theta}_{11}(1-\widehat{\theta}_{11})(\widehat{\pi}_{naive}+\widehat{\theta}_{00}-1)}{N_{.1}(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^3}.$$

Also note that

$$\begin{aligned}\widehat{\pi}_{corrected,PI} &= \widehat{\pi}_{PlugIn} - \frac{\widehat{\text{Var}}(\widehat{\theta}_{00})(1-\widehat{\pi}_{PlugIn})}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^2} - \frac{\widehat{\text{Var}}(\widehat{\theta}_{11})\widehat{\pi}_{PlugIn}}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^2}\\ &= \widehat{\pi}_{PlugIn} + \frac{\widehat{\text{Var}}(\widehat{\theta}_{00})\widehat{\pi}_{PlugIn}}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)(\widehat{\pi}_{naive}+\theta_{00}-1)} - \frac{\widehat{\text{Var}}(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)\widehat{\pi}_{PlugIn}}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^2},\end{aligned}$$

and the estimator $\widehat{\pi}_{corrected,PI}$ can be rewritten as

$$\widehat{\pi}_{corrected,PI} = \widehat{\pi}_{PlugIn}\left\{1 + \frac{\widehat{\text{Cov}}(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1,\widehat{\pi}_{naive}+\widehat{\theta}_{00}-1)}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)(\widehat{\pi}_{naive}+\widehat{\theta}_{00}-1)} - \frac{\widehat{\text{Var}}(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^2}\right\}.$$

This estimator has the same structure as Tin's (1965) "modified ratio estimator". Our situation is more complicated than Tin's, since our ratio is a function of $\widehat{\pi}_{naive}$, $\widehat{\theta}_{00}$, and $\widehat{\theta}_{11}$, each of which may have different sample sizes, and our estimator also is derived by a different method. We find the similarities between the estimators surprising. Tin's estimator has been studied theoretically and via simulation by a number of subsequent authors, e.g. Hutchison (1971), Rao & Rao (1971), Dalabehera & Sahoo (1995), and they found Tin's estimator generally to be less biased and more efficient compared with other proposed ratio estimators. We use simulation to investigate the bias of $\widehat{\pi}_{PlugIn}$ and the performance of $\widehat{\pi}_{corrected,PI}$ in Chapter 5.

2.2 Partially Corrected Estimator

Theorem 2.2.1 *For any $0 \leq \pi \leq 1$, if $(\theta_{00}, \theta_{11})$ satisfy the relationship $\pi(\theta_{00} + \theta_{11} - 2) - \theta_{00} + 1 = 0$ for values inside the unit cube, then the bias of the naive estimator, $\widehat{\pi}_{naive}$, is zero. Similarly, if $(\gamma_{00}, \gamma_{11})$ satisfy $\frac{\pi(2-\gamma_{00}-\gamma_{11})+\gamma_{00}-1}{\gamma_{00}+\gamma_{11}-1} = 0$ for values in the unit cube (and $\gamma_{00} + \gamma_{11} \neq 1$), then the bias of the naive estimator, $\widehat{\pi}_{naive}$, is zero. Figure 1 illustrates these results.*

As noted before, since $|\theta_{00}+\theta_{11}-1| < 1$, the corrected estimator has a larger variance than the naive estimator, even if the misclassification model were known. The implication of that fact and the result above is that for certain combinations of π, θ_{00}, and θ_{11} the validation data should be ignored since the naive estimator is unbiased and has a lower sample variance. Although we would need to know π in order to use that fact directly (and if we knew π, we would be done!), we can create a "partially corrected" estimator that is an affine combination of the naive estimator and the plug in estimator: $\widehat{\pi}_{pc} = a\widehat{\pi}_{naive} + (1-a)\widehat{\pi}_{PlugIn}$. The tuning parameter $0 \leq a \leq 1$ needs to be estimated, and we do that by finding one to minimize an estimate of $MSE(\widehat{\pi}_{pc})$, subject to the constraint that $0 \leq a \leq 1$. Schafer (1986) used a similar idea in linear regression with covariate measurement error. Gustafson (2004), section 5.1 demonstrates that the mean squared error of naive estimator of a log odds ratio can be smaller than the corrected one when the sample size is small. Finally, while it is tempting to use $\widehat{\pi}_{corrected,PI}$ instead of

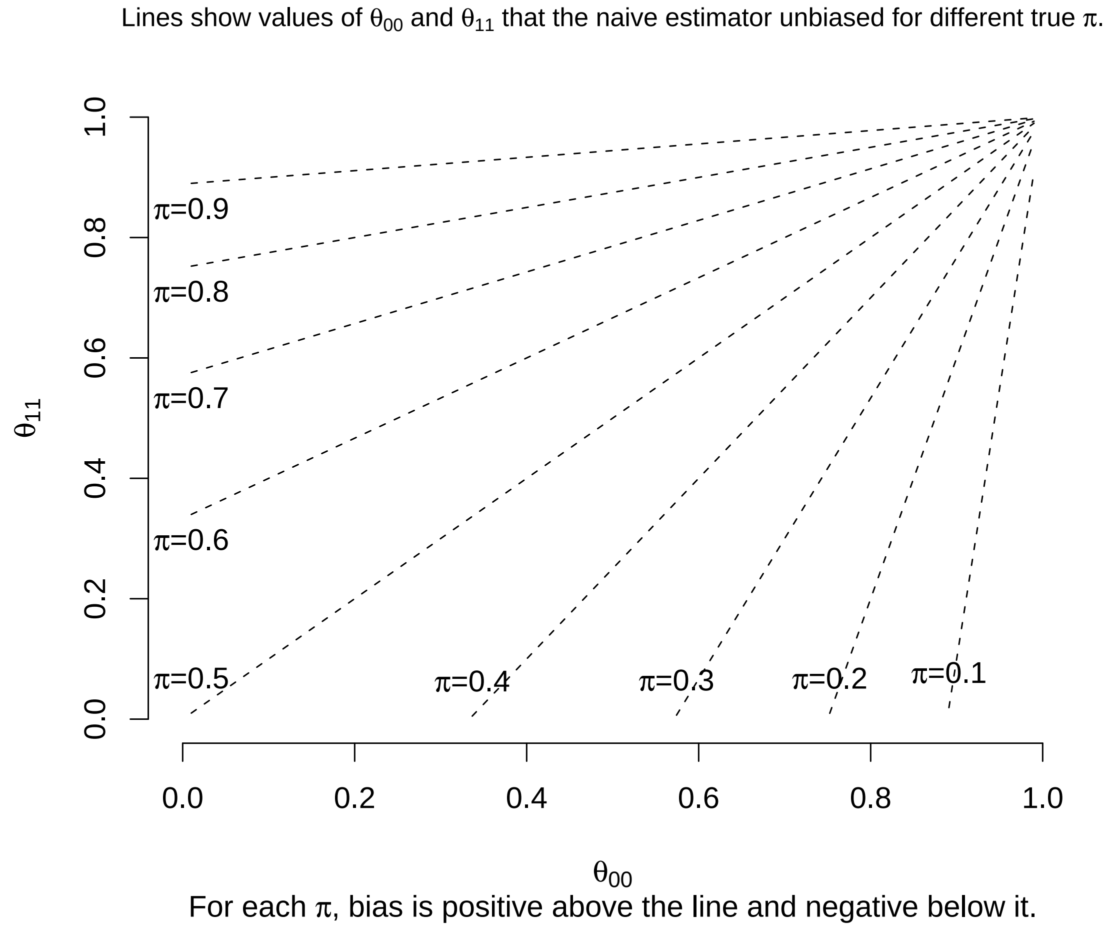


Figure 1. When the naive Estimator is Unbiased: this figure shows combinations of π, θ_{00}, and θ_{11} that result in zero bias for $\widehat{\pi}_{naive}$.

$\widehat{\pi}_{PlugIn}$, we found that a stable estimate of the variance of $\widehat{\pi}_{corrected,PI}$ (an involved expression derived via the the multivariate delta method) to be elusive.

The following two theorems give expressions for a, one for the case where the misclassification model is known, and the other for the case where the misclassification model is estimated from external data. We discuss how to estimate a after the theorems.

Theorem 2.2.2 *When θ_{00}, θ_{11} are known,* *(i)* $MSE(\widehat{\pi}_{pc})$ *has a minimum at*

$$a_{min} = \frac{MSE(\widehat{\pi}_{PlugIn}) - Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})}{MSE(\widehat{\pi}_{naive}) + MSE(\widehat{\pi}_{PlugIn}) - 2Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})}.$$

(ii) $0 < a_{min} < 1$ *if and only if* $MSE(\widehat{\pi}_{PlugIn}) > Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$ *and* $MSE(\widehat{\pi}_{naive}) > Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$.

(iii) If $MSE(\widehat{\pi}_{PlugIn}) < Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$, *then we set* $a_{min} = 0$, *and if* $MSE(\widehat{\pi}_{naive}) < Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$, *then we set* $a_{min} = 1$.

Proof

$$\begin{aligned}
\text{MSE}(\widehat{\pi}_{pc}) &= E[a\widehat{\pi}_{naive} + (1-a)\widehat{\pi}_{PlugIn} - [a\pi + (1-a)\pi]]^2 \\
&= a^2 \,\text{MSE}(\widehat{\pi}_{naive}) + (1-a)^2 \,\text{MSE}(\widehat{\pi}_{PlugIn}) + 2a(1-a)E[(\widehat{\pi}_{naive} - \pi)(\widehat{\pi}_{PlugIn} - \pi)] \\
&= a^2 \,\text{MSE}(\widehat{\pi}_{naive}) + (1-a)^2 \,\text{MSE}(\widehat{\pi}_{PlugIn}) + 2a(1-a)\,\text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) \qquad (2.1)
\end{aligned}$$

Let $f(a)$ refer to equation 2.1. Since $f(a)$ is a quadratic function of a with positive leading coefficient $\text{MSE}(\widehat{\pi}_{naive}) + \text{MSE}(\widehat{\pi}_{PlutIn}) - 2\,\text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$, $f(a)$ has a minimum. The derivative of f is

$$\begin{aligned}
f^{'}(a) &= \{\,\text{MSE}(\widehat{\pi}_{naive}) + \text{MSE}(\widehat{\pi}_{PlugIn}) - 2\,\text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})\}\,a \\
&\quad - \{\,\text{MSE}(\widehat{\pi}_{PlugIn}) - \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})\}\,, \text{ and}
\end{aligned}$$

MSE($\widehat{\pi}_{pc}$) has a minimum when $f^{'}(a) = 0$, that is when

$$a_{min} = \frac{\text{MSE}(\widehat{\pi}_{PlugIn}) - \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})}{\text{MSE}(\widehat{\pi}_{naive}) + \text{MSE}(\widehat{\pi}_{PlugIn}) - 2\,\text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})}.$$

So (i) follows.

By comparing a_{min} with 0 and 1, we obtain (ii).

If $f^{'}(0) = -\{\,\text{MSE}(\widehat{\pi}_{PlugIn}) - \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})\} > 0$ if $\text{MSE}(\widehat{\pi}_{PlugIn}) < \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$, then we have f increasing on interval $[0, 1]$. So we set $a_{min} = 0$ when $\text{MSE}(\widehat{\pi}_{PlugIn}) < \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$. We use the same argument for $\text{MSE}(\widehat{\pi}_{naive}) < \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$,then (iii) follows.

When the misclassification model is known, we have

$$\text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) \leq \frac{1}{2} \{ \text{ MSE}(\widehat{\pi}_{naive}) + \text{ MSE}(\widehat{\pi}_{PlugIn}) \} .$$

If $\text{MSE}(\widehat{\pi}_{PlugIn}) \leq \text{ Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$, we must have $\text{MSE}(\widehat{\pi}_{PlugIn}) \leq \text{ MSE}(\widehat{\pi}_{naive})$ and it is natural to think that $\widehat{\pi}_{PlugIn}$ has the smallest mean square error among all partial corrected estimators (i.e. $a = 0$). The same holds for the situation that $\text{MSE}(\widehat{\pi}_{naive}) \leq \text{ Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$. But if $\text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn})$ is less than both $\text{MSE}(\widehat{\pi}_{naive})$ and $\text{MSE}(\widehat{\pi}_{PlugIn})$, that means we can find a estimator with smaller mean square error.

Theorem 2.2.3 *When $\widehat{\theta}_{00}, \widehat{\theta}_{11}$ are estimated from validation data,*

(i) $MSE(\widehat{\pi}_{pc})$ has a minimum at

$$a_{min} = \frac{MSE(\widehat{\pi}_{PlugIn}) - [Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}]}{MSE(\widehat{\pi}_{naive}) + MSE(\widehat{\pi}_{PlugIn}) - 2\,[Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}]}.$$

(ii) $0 < a_{min} < 1$ if and only if $MSE(\widehat{\pi}_{PlugIn}) > Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}$ and $MSE(\widehat{\pi}_{naive}) > Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}$.

(iii) If $MSE(\widehat{\pi}_{PlugIn}) < Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}$, then we set $a_{min} = 0$, and if $MSE(\widehat{\pi}_{naive}) < Cov(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}$, then we set $a_{min} = 1$.

Proof The proof proceeds as in the case when the misclassification model is known, but now $\widehat{\pi}_{PlugIn}$ is not an unbiased estimator for π and we have

$$\begin{aligned} & E\{(\widehat{\pi}_{naive} - \pi)(\widehat{\pi}_{PlugIn} - \pi)\} \\ = \;& E[\{(\widehat{\pi}_{naive} - \lambda) + (\lambda - \pi)\}\{(\widehat{\pi}_{PlugIn} - E\widehat{\pi}_{PlugIn}) + (E\widehat{\pi}_{PlugIn} - \pi)\}] \\ = \;& \text{Cov}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\} \end{aligned}$$

We notice the difference of these two theorems: $(\lambda - \pi)\{E(\widehat{\pi}_{PlugIn}) - \pi\}$ will be zero if the misclassification model is known. From the above theorems, we know the a is a function of the unknown π. To estimate a, we can use the following algorithm:

1. To make notation simple, we will assume $\widehat{\theta}_{ii} = \theta_{ii}, i = 0, 1$ when the misclassification model is known. Also we define δ as an indicator, where $\delta = 0$ if the misclassification model is known and $\delta = 1$ if the misclassification is estimated from external data. Following these, we have

$$\widehat{\text{Var}}(\widehat{\theta}_{00}) = \delta \frac{\widehat{\theta}_{00}(1-\widehat{\theta}_{00})}{N_{.0}}, \ \widehat{\text{Var}}(\widehat{\theta}_{11}) = \delta \frac{\widehat{\theta}_{11}(1-\widehat{\theta}_{11})}{N_{.1}}.$$

2. We use Theorem 2.1.1 to estimate the bias for $\widehat{\pi}_{PlugIn}$, and define

$$\widehat{\text{Bias}} = \frac{(\ \widehat{\text{Var}}(\widehat{\theta}_{00})(\widehat{\pi}_{naive} - \widehat{\theta}_{11})}{(\widehat{\theta}_{00} + \widehat{\theta}_{11} - 1)^3} + \frac{\widehat{\text{Var}}(\widehat{\theta}_{11})(\widehat{\pi}_{naive} + \widehat{\theta}_{00} - 1)}{(\widehat{\theta}_{00} + \widehat{\theta}_{11} - 1)^3}.$$

3. We estimate the mean square error of $\widehat{\pi}_{naive}$ which is the sum of variance and square of the bias of $\widehat{\pi}_{naive}$. We treat $\widehat{\pi}_{PlugIn}$ as an estimator for π, we have

$$\widehat{\text{MSE}}(\widehat{\pi}_{naive}) = \frac{\widehat{\pi}_{naive}(1-\widehat{\pi}_{naive})}{n} + (\widehat{\pi}_{naive} - \widehat{\pi}_{PlugIn})^2.$$

4. We use the delta method to estimate the variance of $\widehat{\pi}_{PlugIn}$. The square of bias of $\widehat{\pi}_{PlugIn}$ will be small compared with the variance estimate of $\widehat{\pi}_{PlugIn}$ and we will ignore it. That is :

$$\widehat{\text{MSE}}(\widehat{\pi}_{PlugIn}) = \frac{\frac{\widehat{\lambda}(1-\widehat{\lambda})}{n} + \ \widehat{\text{Var}}(\widehat{\theta}_{00}) - 2\widehat{\pi}_{PlugIn}\,\widehat{\text{Var}}(\widehat{\theta}_{00}) + \widehat{\pi}^2_{PlugIn}\left\{\ \widehat{\text{Var}}(\widehat{\theta}_{00}) + \ \widehat{\text{Var}}(\widehat{\theta}_{11})\right\}}{(\widehat{\theta}_{00} + \widehat{\theta}_{11} - 1)^2}$$

5. Using the independence of validation and main study data, and the delta method, we will have the covariance estimate of $\widehat{\pi}_{naive}$ and $\widehat{\pi}_{PlugIn}$:

$$\widehat{\text{Cov}}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) = \frac{\widehat{\pi}_{naive}(1-\widehat{\pi}_{naive})}{n(\widehat{\theta}_{00} + \widehat{\theta}_{11} - 1)}$$

6. Combining the above together, we have our estimator for a_{min}:

$$\widehat{a}_{min} = \frac{\widehat{\text{MSE}}(\widehat{\pi}_{PlugIn}) - \ \widehat{\text{Cov}}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) - (\widehat{\pi}_{naive} - \widehat{\pi}_{PlugIn})\widehat{\text{Bias}}}{\widehat{\text{MSE}}(\widehat{\pi}_{naive}) + \ \widehat{\text{MSE}}(\widehat{\pi}_{PlugIn}) - 2\,\widehat{\text{Cov}}(\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}) + (\widehat{\pi}_{naive} - \widehat{\pi}_{PlugIn})\widehat{\text{Bias}}}$$

We use simulation to investigate the performance of $\widehat{\pi}_{pc}$ in Chapter 5.

2.3 K Category Misclassified Data with External Validation Data

In this section, we will consider categorical data with categories $0, \ldots, K-1$. All the matrices are using zero-based indices.

As before, let $X_1, \ldots, X_n$ be independent and identically multinomial distributed random variables, with $\pi_j = P(X_i = j), j = 0 \ldots K-1$. W_i is the observed value of X_i with $\lambda_j = P(W_i = j)$. Define $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\intercal, \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^\intercal$. Let $\mathbf{P}$ be a $K \times K$ matrix with m, lth element $\theta_{ml} = P(W_i = m|X_i = l)$. We assume $\mathbf{P}^{-1}$ exist. Let $\mathbf{Q}$ be a $K \times K$ matrix with m, lth element $\gamma_{ml} = P(X_i = m|W_i = l)$. $\mathbf{P}$ is the misclassification model and $\mathbf{Q}$ is the reclassification model. With $\widehat{\boldsymbol{\pi}}_{naive} = (\widehat{\pi}^0_{naive}, \ldots, \widehat{\pi}^{K-1}_{naive})^\intercal$ where $\widehat{\pi}^k_{naive} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{W_i=k\}}$ and a known misclassification or reclassification model, we also can develop unbiased method of moments estimators of $\boldsymbol{\pi}$, and the misclassification model requires some adjustment.

Unlike the binary case which only estimates one random variable, we are estimating multivariate correlated random variables when $k \geq 3$. If one of the elements of $\widehat{\boldsymbol{\pi}}$ is outside the parameter space, we no longer can just set that element to 0 or 1. We need to have the sum of all elements equal to 1. The method of moments estimator can run into that problem. In that case, we can use the maximum likelihood estimator (MLE). In van den Hout and van der Heijden (2002), it is proven that when the moment estimator is in the interior of the parameter space, the MLE is equal to the method of moments estimator $\mathbf{P}^{-1}\widehat{\boldsymbol{\pi}}_{naive}$. They also develop an EM algorithm for this situation.

Now we have a correction method with known misclassification model:

$$\widehat{\boldsymbol{\pi}}_{corrected,M} = \mathbf{P}^{-1}\widehat{\boldsymbol{\pi}}_{naive}.$$

and with known reclassification model,

$$\widehat{\boldsymbol{\pi}}_{corrected,R} = \mathbf{Q}\widehat{\boldsymbol{\pi}}_{naive}.$$

$\widehat{\boldsymbol{\pi}}_{corrected,M}, \widehat{\boldsymbol{\pi}}_{corrected,R}$ are both MLEs. When the reclassification model is estimated from independent external validation data, the correction estimator is still unbiased.

When the misclassification model is estimated from external data, $\widehat{\boldsymbol{\pi}}_{PlugIn} = \widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive}$ is biased, due to the fact that $E\widehat{\mathbf{P}}^{-1} \neq (E\widehat{\mathbf{P}})^{-1}$. We will prove an unbiased estimator of $\mathbf{P}^{-1}$ does not exist in section 2.3.1. We will still focus on the misclassification model. Section 2.3.2 contains a reduced bias estimator and section 2.3.3 discusses a partially corrected estimator.

2.3.1 The Proof of No Unbiased Estimator for $\mathbf{P}^{-1}$

In this section, we will prove that there does not exist an unbiased estimator for $\mathbf{P}^{-1}$. This generalizes a similar result for $\frac{1}{p}$ for binary data with unknown probability p.

$\mathbf{P}$ is a misclassification matrix of dimension K if $\mathbf{P}$ is a $K \times K$ matrix with the value of each entry between 0 and 1, and the sum of each column equal to 1. Let $\mathcal{P}_K = \{\mathbf{P}|\mathbf{P}$ is a $K \times K$ misclassification matrix $\}$.

Theorem 2.3.1 *Let* $\mathbf{A} = (\boldsymbol{a}_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_{K-1})$ *be a* $K \times K$ *random matrix with a distribution such that* $\boldsymbol{a}_i$ *is multinomial* $(N_{.i}, \boldsymbol{\theta}_i)$ *where* $\boldsymbol{a}_i = \begin{pmatrix} a_{0i} \\ \vdots \\ a_{(K-1)i} \end{pmatrix}, \boldsymbol{\theta}_i = \begin{pmatrix} \theta_{0i} \\ \vdots \\ \theta_{(K-1)i} \end{pmatrix}$, *and* $\mathbf{P} = (\boldsymbol{\theta}_0, \ldots \boldsymbol{\theta}_{K-1})$ *is an unknown misclassification matrix. Then there is no unbiased estimator of* $\mathbf{P}^{-1}$

Proof First we will prove that $\{\det(\mathbf{P}^{-1})|\mathbf{P} \in \mathcal{P}_K\}$ is unbounded. If we assume there does exist an unbiased estimator for $\mathbf{P}^{-1}$, then we can have $\{\det(\mathbf{P}^{-1})|\mathbf{P} \in \mathcal{P}_K\}$ is bounded, which leads to a contradiction.

For $m \in \mathbb{N}$, let

$$\mathbf{P}_m = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & \frac{m-1}{m(K-1)} \\ 0 & 1 & 0 & 0 & \ldots & \frac{m-1}{m(K-1)} \\ 0 & 0 & 1 & 0 & \ldots & \frac{m-1}{m(K-1)} \\ 0 & 0 & 0 & \ddots & \ldots & \frac{m-1}{m(K-1)} \\ 0 & 0 & \ldots & \ldots & \ldots & \frac{1}{m} \end{pmatrix}$$

$\mathbf{P}_m \in \mathcal{P}_K$ and $\det(\mathbf{P}_m) = \frac{1}{m}$, and therefore $\{\det(\mathbf{P}^{-1})|\mathbf{P} \in \mathcal{P}_K\}$ is unbounded.

Assume $\mathbf{T}$ is an unbiased estimator for $\mathbf{P}^{-1}$ with element $T_{ij} = \sum_{l=0}^{n_{ij}} G_{ij}^{l}(\mathbf{A}), i, j = 0 \ldots K-1$ where $G_{ij}^{l}(\mathbf{A})$ is a polynomial of elements in $\mathbf{A}$ with degree l. Let the sum of the absolute value of coefficients of $G_{ij}^{l}(\mathbf{A})$ be C_{ij}^{l}. Then,

$E[T_{ij}] = (\mathbf{P}^{-1})_{ij} = E[\sum_{l=0}^{n_{ij}} G_{ij}^{l}(\mathbf{A})] \leq \sum_{l=0}^{n_{ij}} C_{ij}^{l}$ from the property of multinomial distribution, $P(\bigcap_{\substack{i=0\\j=1}}^{K-1} \{0 \leq a_{i,j} \leq 1\}) = 1$.

So $\mathbf{P}_{ij}^{-1}$ is bounded for all $\mathbf{P} \in \mathcal{P}$ and all $i, j = 0 \ldots K-1$ As a result, $\det(\mathbf{P}^{-1})$ is bounded above for all $\mathbf{P} \in \mathcal{P}$ which contradicts our earlier statement, therefore there is no unbiased estimator for $\mathbf{P}^{-1}$.

2.3.2 Bias Reduced Estimator for $k \geq 3$

In this section, we will discuss the bias for $\widehat{\boldsymbol{\pi}}_{PlugIn}$ for misclassified data with K categories.

Assume we have K categories with misclassification model

$$\mathbf{P} = \begin{pmatrix} \theta_{00} & \theta_{01} & \cdots & \theta_{0(K-1)} \\ \theta_{10} & \theta_{11} & \cdots & \theta_{1(K-1)} \\ & & \ddots & \\ 1-\sum_{i=0}^{K-2}\theta_{i0} & 1-\sum_{i=0}^{K-2}\theta_{i1} & \cdots & 1-\sum_{i=0}^{K-2}\theta_{i(K-1)} \end{pmatrix}.$$

Assume $\widehat{\mathbf{P}}$ is an estimator for $\mathbf{P}$ from external data for $\mathbf{P}$ and $\det(\widehat{\mathbf{P}}) \neq 0$. Then $\widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive}$ is a simple estimator for $\boldsymbol{\pi}$. It is a biased estimator since $E\widehat{\mathbf{P}}^{-1} \neq \mathbf{P}^{-1}$.

Let $f(\boldsymbol{\theta}, \boldsymbol{\pi}_{naive}) = \mathbf{P}^{-1}\boldsymbol{\pi}_{naive}$ and let $\mathbf{M}_{ij}$ be a $K \times K$ matrix with 0 everywhere except in position i, j which a 1 , and $K-1, j$ which contains a -1, $i, j = 0, \ldots K-1$. Then using the delta method and any scalar variable x, $\dfrac{\partial \mathbf{P}^{-1}}{\partial x} = -\mathbf{P}^{-1}\dfrac{\partial \mathbf{P}}{\partial x}\mathbf{P}^{-1}$ (Harville, 197, Section 15.8) we will have

$$E\widehat{\boldsymbol{\pi}}_{PlugIn} \quad = \quad \boldsymbol{\pi} + \sum_{j=0}^{K-1}\sum_{i=0}^{K-2} \operatorname{Var}(\widehat{\theta}_{ij})\mathbf{P}^{-1}\mathbf{M}_{ij}\mathbf{P}^{-1}\mathbf{M}_{ij}\mathbf{P}^{-1}\boldsymbol{\pi}_{naive}$$

$$+\sum_{j=0}^{K-1}\sum_{i<i'}^{K-2} \mathrm{Cov}(\widehat{\theta}_{ij},\widehat{\theta}_{i'j})\Big\{\mathbf{P}^{-1}\mathbf{M}_{i'j}\mathbf{P}^{-1}\mathbf{M}_{ij}\mathbf{P}^{-1}\boldsymbol{\pi}_{naive}$$
$$+\mathbf{P}^{-1}\mathbf{M}_{ij}\mathbf{P}^{-1}\mathbf{M}_{i'j}\mathbf{P}^{-1}\boldsymbol{\pi}_{naive}\Big\} + O(\min_i(N_{.k}^{-2}))$$

where $N_{.k}$ is the sample of validation size in category k. This result gives a first order bias corrected estimator:

$$\begin{aligned}\widehat{\boldsymbol{\pi}}_{corrected,PI} &= \widehat{\boldsymbol{\pi}}_{PlugIn} - \sum_{j=0}^{K-1}\sum_{i=0}^{K-2} \mathrm{Var}(\widehat{\theta}_{ij})\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive} \\ &\quad -\sum_{j=0}^{K-1}\sum_{i<i'}^{K-2} \mathrm{Cov}(\widehat{\theta}_{ij},\widehat{\theta}_{i'j})\Big\{\widehat{\mathbf{P}}^{-1}\mathbf{M}_{i'j}\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive} \\ &\quad +\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\mathbf{M}_{i'j}\widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive}\Big\}\end{aligned}$$

Even though $\widehat{\boldsymbol{\pi}}_{corrected,PI}$ can reduce the bias, we can't guarantee that it is in the parameter space. Unlike the binary case in which we know how to adjust if the estimator is not in the parameter space, we do not know how to adjust in this situation if $K \geq 3$. If $\widehat{\boldsymbol{\pi}}_{corrected,PI}$ is outside the parameter space, we should compare it with the pseudo MLE estimator (i.e. treat $\widehat{\mathbf{P}}$ as known and use the EM algorithm to get MLE).

2.3.3 Partially Corrected Estimator

When the dimension K is greater than 2, we have more than one parameter to estimate in $\boldsymbol{\pi}$. We will define the mean square error as the sum of mean square error of each parameter.

Definition Assume $\widehat{\boldsymbol{\pi}}$ is an estimator for $\boldsymbol{\pi}$, then

$$\mathrm{MSE}(\widehat{\boldsymbol{\pi}}) = E(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi})^{\mathsf{T}}(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi}) = \text{ trace}(\mathrm{Var}(\widehat{\boldsymbol{\pi}})) + (E\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi})^{\mathsf{T}}(E\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi}).$$

As before, we define the partially corrected estimator $\widehat{\boldsymbol{\pi}}_{pc} = a\widehat{\boldsymbol{\pi}}_{naive} + (1-a)\widehat{\boldsymbol{\pi}}_{plugIn}$.

Lemma 2.3.2 *When the misclassification model* $\mathbf{P}$ *is known,* *(i)* $MSE(\widehat{\boldsymbol{\pi}}_{pc})$ *has a minimum at*

$$a_{min} = \frac{MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) - trace(\,Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))}{MSE(\widehat{\boldsymbol{\pi}}_{naive}) + MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) - 2trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))}.$$

(ii) $0 < a_{min} < 1$ *if and only if* $MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) > trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))$ *and* $MSE(\widehat{\boldsymbol{\pi}}_{naive}) > trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))$.

(iii) If $MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) < trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))$, *then we set* $a_{min} = 0$, *and if* $MSE(\widehat{\boldsymbol{\pi}}_{naive}) < trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))$, *then we set* $a_{min} = 1$.

Proof

$$\begin{aligned}
& \text{MSE}(\widehat{\boldsymbol{\pi}}_{pc}) \\
= \;& E\left[a\widehat{\boldsymbol{\pi}}_{naive} + (1-a)\widehat{\boldsymbol{\pi}}_{PlugIn} - \{a\boldsymbol{\pi} + (1-a)\boldsymbol{\pi})\}\right]^{\mathsf{T}}\left[a\widehat{\boldsymbol{\pi}}_{naive} + (1-a)\widehat{\boldsymbol{\pi}}_{PlugIn} - \{a\boldsymbol{\pi} + (1-a)\boldsymbol{\pi}\}\right] \\
= \;& a^2\,\text{MSE}(\widehat{\boldsymbol{\pi}}_{naive}) + (1-a)^2\,\text{MSE}(\widehat{\boldsymbol{\pi}}_{PlugIn}) + 2a(1-a)E(\widehat{\boldsymbol{\pi}}_{naive} - \boldsymbol{\pi})^{\mathsf{T}}(\widehat{\boldsymbol{\pi}}_{PlugIn} - \boldsymbol{\pi}) \\
= \;& a^2\,\text{MSE}(\widehat{\boldsymbol{\pi}}_{naive}) + (1-a)^2\,\text{MSE}(\widehat{\boldsymbol{\pi}}_{PlugIn}) + 2a(1-a)\,\text{trace}(\,\text{Cov}(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))
\end{aligned}$$

is a quadratic function of a as before. The only difference is instead of $\text{Cov}(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn})$, we have $\text{trace}(\,\text{Cov}(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))$. So we can use the proof of Theorem 2.2.2 to prove this lemma.

Lemma 2.3.3 *When misclassification model* $\widehat{\mathbf{P}}$ *is estimated from validation data,*

(i) $MSE(\widehat{\boldsymbol{\pi}}_{pc})$ *has a minimum at*

$$a_{min} = \frac{MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \left[C + (\boldsymbol{\lambda} - \boldsymbol{\pi})^{\mathsf{T}}\{E(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \boldsymbol{\pi}\}\right]}{MSE(\widehat{\boldsymbol{\pi}}_{naive}) + MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) - 2\left[C + (\boldsymbol{\lambda} - \boldsymbol{\pi})^{\mathsf{T}}\{E(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \boldsymbol{\pi}\}\right]},$$

where $C = trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}))$.

(ii) $0 < a_{min} < 1$ *if and only if* $MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) > trace(\,Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn})) + (\boldsymbol{\lambda} - \boldsymbol{\pi})^{\mathsf{T}}\{E(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \boldsymbol{\pi}\}$ *and* $MSE(\widehat{\boldsymbol{\pi}}_{naive}) > trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn})) + (\boldsymbol{\lambda} - \pi)^{\mathsf{T}}\{E(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \boldsymbol{\pi}\}$.

(iii) If $MSE(\widehat{\boldsymbol{\pi}}_{PlugIn}) < trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn})) + (\boldsymbol{\lambda} - \pi)^{\mathsf{T}}\{E(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \boldsymbol{\pi}\}$, *then we set* $a_{min} = 0$, *and if* $MSE(\widehat{\boldsymbol{\pi}}_{naive}) < trace(Cov(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn})) + (\boldsymbol{\lambda} - \boldsymbol{\pi})^{\mathsf{T}}\{E(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \boldsymbol{\pi}\}$, *then we set* $a_{min} = 1$.

As in the binary case, to estimate a, we need to evaluate some estimators involving the unknown parameters:

1. To make notation simple, we will assume $\widehat{\mathbf{P}} = \mathbf{P}$ when the misclassification model is known. Also we define δ as an indicator, $\delta = 0$ if the misclassification model is known and $\delta = 1$ if the misclassification is estimated from external data. Following these, we have $\widehat{\text{Var}}(\widehat{\boldsymbol{\pi}}_{naive}) = \delta \langle \widehat{v}_{ij} \rangle_{i,j=0 \ldots K-1}$, where
$$\widehat{v}_{ij} = \begin{cases} -\widehat{\pi}^{i}_{naive}\widehat{\pi}^{j}_{naive}/n & \text{if} \quad i \neq j \\ \widehat{\pi}^{i}_{naive}(1-\widehat{\pi}^{j}_{naive})/n & \text{if} \quad i = j \end{cases}$$

2. We use section 2.3.2 to estimate the bias for $\widehat{\boldsymbol{\pi}}_{PlugIn}$, and define
$$\begin{aligned} \widehat{\text{Bias}} \quad = \quad & \sum_{j=0}^{K-1}\sum_{i=0}^{K-2} \widehat{\text{Var}}(\widehat{\theta}_{ij})\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive} \\ & + \sum_{j=0}^{K}\sum_{i<i'}^{K-2} \widehat{\text{Cov}}(\widehat{\theta}_{ij}, \widehat{\theta}_{i'j}) \Big\{ \widehat{\mathbf{P}}^{-1}\mathbf{M}_{i'j}\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\boldsymbol{\pi}_{naive} \\ & + \widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}\mathbf{M}_{i'j}\widehat{\mathbf{P}}^{-1}\widehat{\boldsymbol{\pi}}_{naive} \Big\}. \end{aligned}$$

3. We estimate the mean square error of $\widehat{\boldsymbol{\pi}}_{naive}$. Treating $\widehat{\boldsymbol{\pi}}_{PlugIn}$ as an estimator for $\boldsymbol{\pi}$, we have
$$\widehat{\text{MSE}}(\widehat{\boldsymbol{\pi}}_{naive}) = \sum_{i=0}^{K-1} \widehat{\lambda}_i(1-\widehat{\lambda}_i)/n + (\widehat{\boldsymbol{\pi}}_{naive} - \widehat{\boldsymbol{\pi}}_{PlugIn})^{\mathsf{T}}(\widehat{\boldsymbol{\pi}}_{naive} - \widehat{\boldsymbol{\pi}}_{PlugIn}).$$

4. We use the delta method to estimate the variance of $\widehat{\pi}_{PlugIn}$. The inner product part of the bias of $\widehat{\pi}_{PlugIn}$ will be too small compared with the trace of variance estimate of $\widehat{\pi}_{PlugIn}$ and we will ignore it. That is :
$$\begin{aligned} \widehat{\text{MSE}}(\widehat{\boldsymbol{\pi}}_{PlugIn}) \quad = \quad & \text{trace}\Bigg(\widehat{\mathbf{P}}^{-1}\, \text{Var}(\widehat{\boldsymbol{\pi}}_{naive})(\widehat{\mathbf{P}}^{\mathsf{T}})^{-1} \\ & + \sum_{j=0}^{K-1}\sum_{i=0}^{K-2} \text{Var}(\widehat{\theta}_{ij})\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1}(\widehat{\mathbf{P}}^{-1}\mathbf{M}_{ij}\widehat{\mathbf{P}}^{-1})^{\mathsf{T}} \Bigg) \end{aligned}$$

5. Using the independence of validation and main study data, and the delta method, we will have the covariance estimate of $\widehat{\pi}_{naive}$ and $\widehat{\pi}_{PlugIn}$:
$$\widehat{\text{Cov}}(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn}) = \text{Var}(\widehat{\boldsymbol{\pi}}_{naive})(\widehat{\mathbf{P}}^{\mathsf{T}})^{-1}, \text{ and } \widehat{\mathbf{C}} = \text{trace}(\widehat{\text{Cov}}(\widehat{\boldsymbol{\pi}}_{naive}, \widehat{\boldsymbol{\pi}}_{PlugIn})).$$

6. Combining the above together, we have our estimator for a_{min}:

$$\widehat{a}_{min} = \frac{\widehat{\text{MSE}}(\widehat{\boldsymbol{\pi}}_{PlugIn}) - \left\{\widehat{\mathbf{C}} + (\widehat{\boldsymbol{\pi}}_{naive} - \widehat{\boldsymbol{\pi}}_{PlugIn})^{\mathsf{T}}\widehat{\text{Bias}}\right\}}{\widehat{\text{MSE}}(\widehat{\boldsymbol{\pi}}_{naive}) + \widehat{\text{MSE}}(\widehat{\boldsymbol{\pi}}_{PlugIn}) - 2\left\{\widehat{\mathbf{C}} + (\widehat{\boldsymbol{\pi}}_{naive} - \widehat{\boldsymbol{\pi}}_{PlugIn})^{\mathsf{T}}\widehat{\text{Bias}}\right\}}.$$

CHAPTER 3

CONFIDENCE INTERVALS FOR π

In this chapter, we will focus on methods to find confidence intervals for the probability of interest of misclassified binary data with an estimated misclassification model. As we noted earlier, Schwartz (1985) describes how misclassification will affect the coverage probability of the traditional Wald 95% confidence interval. Without correction, the Wald confidence interval is not reliable for misclassified data.

From Chapter 2, with external validation data, we consider $\widehat{\pi}_{PlugIn} = \dfrac{\widehat{\pi}_{naive} + \widehat{\theta}_{00} - 1}{\widehat{\theta}_{00} + \widehat{\theta}_{11} - 1}$ as an estimator for π, the proportion of interest. We consider two novel ways to get a confidence interval for π. One way is to get the confidence intervals for $\pi_{naive}, \theta_{00},$ and θ_{11}, and then proceed to get the Bonferroni joint confidence interval for π ($\pi_{naive}, \theta_{00},$ and θ_{11} are parameters from independent binary distributions and there are a number of ways to get confidence intervals for each of those parameters). This idea is also adapted from Buonaccorsi (2010). It is known that Wald confidence intervals for a binomial proportion perform poorly in terms of coverage probability when π is near 0 or 1. Vollset (1993), Newcombe (1998), Brown et al. (2001) and many other authors compare different confidence intervals for the population of proportion, using methods including Wald, Wilson, Agresti, Jefferys, Clopper-Pearson and continuity correction. Brown et al. (2001) show that due to the nature of discreteness and skewness in the binomial distribution, the actual coverage of the Wald interval can be significantly smaller than the nominal level for moderate and even large sample sizes (such as 1876) and not just for π near 0 or 1.

A second way to get a confidence interval in our situation is Fieller's method (derived by Fieller (1954); see Buonaccorsi (2001) for a detailed discussion). As von Luxburg and Franz (2008) point out, the confidence interval constructed by Fieller's theorem is in fact a projected confidence region of bivariate normal data. Guiard (1989) and Milliken (1982) contains related results.

In the following, we introduce two interval estimators for π when the misclassification model is estimated from external validation data. Both intervals account for the potentially substantial variability that is introduced by the validation data. Section 3.1 will be devoted to the confidence interval for π using projection and Bonferroni correction. We will use Fieller's method in Section 3.2. Different methods to get confidence intervals for $\widehat{\pi}_{naive}, \theta_{00}$ and θ_{11} are described in Appendix A. We assess the performance of these methods and compare them to a SIMEX approach (Kuchenhoff et al, 2007), a multivariate delta method approach, and an interval that does not include variability from validation data in Chapter 5.

3.1 Optimization Based Projected Interval

Since $\pi = \dfrac{\pi_{naive} + \theta_{00} - 1}{\theta_{00} + \theta_{11} - 1}$, one way to find a $100(1-\alpha)\%$ confidence interval for π is to find confidence intervals for π_{naive}, θ_{00}, and θ_{11}, each with $100(1-\alpha)^{1/3}\%$ confidence level. Denote these intervals as $[L_{\pi_{naive}}, U_{\pi_{naive}}]$, $[L_{00}, U_{00}]$, and $[L_{11}, U_{11}]$ respectively (See Appendix A for different methods to find those intervals). Let

$$R = \{\pi | \pi = \frac{\pi_{naive} + \theta_{00} - 1}{\theta_{00} + \theta_{11} - 1}, \pi_{naive} \in [L_{\pi_{naive}}, U_{\pi_{naive}}], \theta_{00} \in [L_{00}, U_{00}], \theta_{11} \in [L_{11}, U_{11}]\}.$$

Then $P(R) \geq \alpha$, since $\widehat{\pi}_{naive}, \widehat{\theta}_{00}$ and $\widehat{\theta}_{11}$ are independent, and the mapping $f : [0,1]^3 \rightarrow [0,1], (\pi_{naive}, \theta_{00}, \theta_{11}) \mapsto \frac{\pi_{naive}+\theta_{00}-1}{\theta_{00}+\theta_{11}-1}$ is not one-to one. Therefore, R is a $100(1-\alpha)\%$ confidence set for π. In the following, we would like to determine conditions under which R is an interval, and determine its upper and lower bounds. This is a constrained optimization problem. An interval is optimal if it contains R and is as short as possible while maintaining level $1-\alpha$. We find the interval by solving two optimization problems. First, we find the left endpoint by minimizing π subject to the constraints: $\pi_{naive} \in [L_{\pi_{naive}}, U_{\pi_{naive}}], \theta_{00} \in [L_{00}, U_{00}]$ and $\theta_{11} \in [L_{11}, U_{11}]$. The right endpoint is found by maximizing π subject to the same constraints. The proof below uses compactness to show that every π between the two endpoints is in the interval. It is tempting to use $\left[\frac{L_{\pi_{naive}}+L_{00}-1}{U_{00}+U_{11}-1}, \frac{U_{\pi_{naive}}+U_{00}-1}{L_{00}+L_{11}-1}\right]$, but this does not necessarily solve the optimization problems above since the same θ_{00} must be used in the numerators and denominators in both end points. As a result, we use constrained optimization to find the interval.

The following example illustrates the problem with the tempting interval. Suppose we have $[L_{\pi_{naive}}, U_{\pi_{naive}}] = [0.37, 0.51]$, $[L_{00}, U_{00}] = [0.871, 0.975]$, and $[L_{11}, U_{11}] = [0.662, 0.838]$, then the tempting interval is $[0.30, 0.91]$ as a confidence interval for π. But if we use the optimization method, we will get $R = [0.34, 0.761]$ as a confidence interval for π.

Theorem 3.1.1 *Define* $f(\pi_{naive}, \theta_{00}, \theta_{11}) = \frac{\pi_{naive}+\theta_{00}-1}{\theta_{00}+\theta_{11}-1}$ *for* $\pi_{naive} \in [L_{\pi_{naive}}, U_{\pi_{naive}}], \theta_{00} \in [L_{00}, U_{00}], \theta_{11} \in [L_{11}, U_{11}]$ *and assume* $(L_{00}+L_{11}-1)(U_{00}+U_{11}-1) > 0$. *Then the maximum and minimum values of* f, M, *and* m *respectively, occur at endpoints of these intervals. The optimal interval of* $\mathbb{R}$ *is* $[m, M]$.

Proof Since $(L_{00} + L_{11} - 1)(U_{00} + U_{11} - 1) > 0$, f is continuous on $[L_{\pi_{naive}}, U_{\pi_{naive}}] \times [L_{00}, U_{00}] \times [L_{11}, U_{11}]$, a compact set, so $f([L_{\pi_{naive}}, U_{\pi_{naive}}] \times [L_{00}, U_{00}] \times [L_{11}, U_{11}])$ is compact too. That means f has maximum/minimum values on this region. Assume f attains its minimum at $(\pi^*_{naive}, \theta^*_{00}, \theta^*_{11})$.

f is defined on the region $[L_{\pi_{naive}}, U_{\pi_{naive}}] \times [L_{00}, U_{00}] \times [L_{11}, U_{11}]$, and the region can transfer to the constraints:

$$\begin{pmatrix} g_1(\pi_{naive}, \theta_{00}, \theta_{11}) \\ g_2(\pi_{naive}, \theta_{00}, \theta_{11}) \\ g_3(\pi_{naive}, \theta_{00}, \theta_{11}) \end{pmatrix} = \begin{pmatrix} (\pi_{naive} - L_{\pi_{naive}})(\pi_{naive} - U_{\pi_{naive}}) \\ (\theta_{00} - L_{00})(\theta_{00} - U_{00}) \\ (\theta_{11} - L_{11})(\theta_{11} - U_{11}) \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

According to Kuhn-Tucker conditions (Luenberger, 1973), there exist $a_i \geq 0, i =$

$1 \ldots 3$ such that :

$(\nabla f + a_1 \nabla g_1 + a_2 \nabla g_2 + a_3 \nabla g_3)|_{\pi^*_{naive}, \theta^*_{00}, \theta^*_{11}} = \mathbf{0}$ and $a_i g_i(\pi^*_{naive}, \theta^*_{00}, \theta^*_{11}) = 0$ for $i = 1 \ldots 3$.

From $a_i g_i(\pi^*_{naive}, \theta^*_{00}, \theta^*_{11}) = 0$, we have :

$$\text{either } a_1 = 0 \text{ or } \pi^*_{naive} = L_{\pi_{naive}} \text{ or } \pi_{naive} = U_{\pi_{naive}},$$

$$\text{either } a_2 = 0 \text{ or } \theta^*_{00} = L_{00} \text{ or } \theta_{00} = U_{00},$$

$$\text{either } a_3 = 0 \text{ or } \theta^*_{11} = L_{11} \text{ or } \theta_{11} = U_{11}.$$

$$\begin{aligned}
(\nabla f + a_1 \nabla g_1 + a_2 \nabla g_2 a_3 \nabla g_3)|_{\pi^*_{naive}, \theta^*_{00}, \theta^*_{11}} &= \begin{pmatrix} \frac{1}{\theta^*_{00}+\theta^*_{11}-1} + a_1[2\pi^*_{naive} - (L_{\pi_{naive}} + U_{\pi_{naive}})] \\ \frac{\theta^*_{11}-\pi^*_{naive}}{(\theta^*_{00}+\theta^*_{11}-1)^2} + a_2[2\theta^*_{00} - (L_{00} + U_{00})] \\ \frac{-(\pi^*_{naive}+\theta^*_{00}-1)}{(\theta^*_{00}+\theta^*_{11}-1)^2} + a_3[2\theta^*_{11} - (L_{11} + U_{11})] \end{pmatrix} \\
&= \mathbf{0},
\end{aligned}$$

Therefore, we have $a_1 \neq 0$ and $\pi^*_{naive} = L_{\pi_{naive}}$ or $\pi_{naive} = U_{\pi_{naive}}$.

If $a_2 = 0,$ then $\theta^*_{11} = \pi^*_{naive}$, and $f(\pi^*_{naive}, \theta_{00}, \theta^*_{11}) = 1$ for all $\theta_{00} \in [L_{00}, U_{00}]$ and we have $\theta^*_{00} = L_{00}$ or $\theta^*_{00} = U_{00}$.

If $a_3 = 0,$ then $\pi^*_{naive} + \theta^*_{00} - 1 = 0$, and $f(\pi^*_{naive}, \theta^*_{00}, \theta_{11}) = 0$ for all $\theta_{11} \in [L_{11}, U_{11}]$ and we have $\theta^*_{11} = L_{11}$ or $\theta^*_{11} = U_{11}$.

So the minimum values of f occur at endpoints of these intervals.

Using the same argument for $h = -f$, if h has a minimum value, it is at an endpoint of these intervals. So the maximum/minimum values of f occur at endpoints of these intervals.

From the above theorem, we know all the a_i are non-zero and positive with respect to the endpoints of these intervals, that is, the relative minimum values of f . We can find necessary conditions for an endpoint to have a minimum value by solving for a_i in the gradient equations and get

$$a_1 = -\frac{1}{\{(\theta^*_{00} + \theta^*_{11} - 1)2\pi^*_{naive} - (L_{\pi_{naive}} + U_{\pi_{naive}})\}} > 0,$$

$$\begin{aligned} a_2 &= -\frac{\theta^*_{11} - \pi^*_{naive}}{(\theta^*_{00} + \theta^*_{11} - 1)^2 \{2\theta^*_{00} - (L_{00} + U_{00})\}} > 0, \\ a_3 &= \frac{\pi^*_{naive} + \theta^*_{00} - 1}{(\theta^*_{00} + \theta^*_{11} - 1)^2 \{2\theta^*_{11} - (L_{11} + U_{11})\}} > 0. \end{aligned}$$

So $2\pi^*_{naive} - (L_{\pi_{naive}} + U_{\pi_{naive}})$ and $\theta^*_{11} + \theta^*_{00} - 1$ have different signs, $\theta^*_{11} - \pi^*_{naive}$ and $2\theta^*_{00} - (L_{00} + U_{00})$ have different signs, and $\pi^*_{naive} + \theta^*_{00} - 1$ and $2\theta^*_{11} - (L_{11} + U_{11})$ have the same sign. We can find some similar relationship for endpoints with relative maximum values. By observing the relationships of the signs, we will summarize the necessary conditions for an endpoint to have a relative minimum /maximum value in the following lemma.

Lemma 3.1.2 *This table summarizes the necessary conditions for an endpoint of confidence intervals of $\pi_{naive}, \theta_{00}, \theta_{11}$ to have a relative minimum/maximum value of f. The upper one is for relative minimum, and the lower one is for relative maximum:*

Endpoints			**Sign Pattern**		
π_{naive}	θ_{00}	θ_{11}	$\theta_{00} + \theta_{11} - 1$	$\theta_{11} - \pi_{naive}$	$\pi_{naive} + \theta_{00} - 1$
$U_{\pi_{naive}}$	U_{00}	U_{11}	-**/+**	-/+	+**/-**
$L_{\pi_{naive}}$	U_{00}	U_{11}	+/-	-/+	+/-
$U_{\pi_{naive}}$	L_{00}	U_{11}	-*/+*	+*/-*	+*/-*
$L_{\pi_{naive}}$	L_{00}	U_{11}	+/-	+/-	+/-
$U_{\pi_{naive}}$	U_{00}	L_{11}	-/+	-/+	-/+
$L_{\pi_{naive}}$	U_{00}	L_{11}	+*/-*	-*/+*	-*/+*
$U_{\pi_{naive}}$	L_{00}	L_{11}	-/+	+/-	-/+
$L_{\pi_{naive}}$	L_{00}	L_{11}	+**/-**	+/-	-**/+**

** indicates sign pattern is impossible*

*** indicates function value is negative, therefore is not a probability*

The above are mathematical results, but the resulting interval is not necessarily contained in $[0, 1]$. The algorithm to find the upper and lower bound for π is present in Appendix B.

3.2 Fieller's Method Based Interval

Fieller's method provides a way to develop a confidence interval for the ratio of two parameters, such as

$$\pi = \frac{\pi_{naive} + \theta_{00} - 1}{\theta_{11} + \theta_{00} - 1} := \frac{N}{D}.$$

A recent paper by von Luxburg and Franz (2009) reviews the literature on Fieller's method comprehensively. That paper also provides the following geometric interpretation of the method: an elliptical confidence region can be developed for N and D such that Fieller's interval is equivalent to the set of all N/D that are in the ellipse. Similar results can be found in Guiard (1989) and Milliken (1982). We slightly modify that procedure and use the interval that is the intersection of Fieller's interval and [0,1], the domain for our ratio.

In our situation, Fieller's procedure can result in four types of confidence sets for the ratio: a simple bounded interval that is contained in [0,1], an ”unbounded interval” that becomes [0,1] when intersected with the domain of the ratio, a disjoint interval, or an empty interval. Figure 2 illustrates the first three of these cases. The first type (simple bounded) of set can occur when the confidence ellipse for N and D is in quadrant 1 or quadrant 3 and does not intersect the y-axis. The second type of set (unbounded) occurs when the origin is in the ellipse. The third type of set (disjoint) occurs when the ellipse intersects the y-axis, but does not contain the origin. The fourth type of set (empty) occurs when all of the ratios formed by the set of Ns and Ds inside the ellipse are outside of the [0,1] domain for the ratio. As the main study and validation sample sizes become large, the intervals will be of the simple bounded type. In our case, from the central limit theorem, we have

$$\begin{pmatrix} \widehat{\pi}_{naive} + \widehat{\theta}_{00} - 1 \\ \widehat{\theta}_{11} + \widehat{\theta}_{00} - 1 \end{pmatrix} \xrightarrow{D} MVN \left\{ \begin{pmatrix} \pi_{naive} + \theta_{00} - 1 \\ \theta_{00} + \theta_{11} - 1 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right\}$$

with $\sigma_{11} = \frac{\pi_{naive}(1-\pi_{naive})}{n} + \frac{\theta_{00}(1-\theta_{00})}{N_{11}}$, $\sigma_{22} = \frac{\theta_{11}(1-\theta_{11})}{N_{11}} + \frac{\theta_{00}(1-\theta_{00})}{N_{11}}$, $\sigma_{12} = \frac{\theta_{00}(1-\theta_{00})}{N_{11}}$, and $z_{\alpha/2}$ the $1 - \alpha/2$ quantile of the standard normal distribution, the ellipse for a $1 - \alpha$

interval is

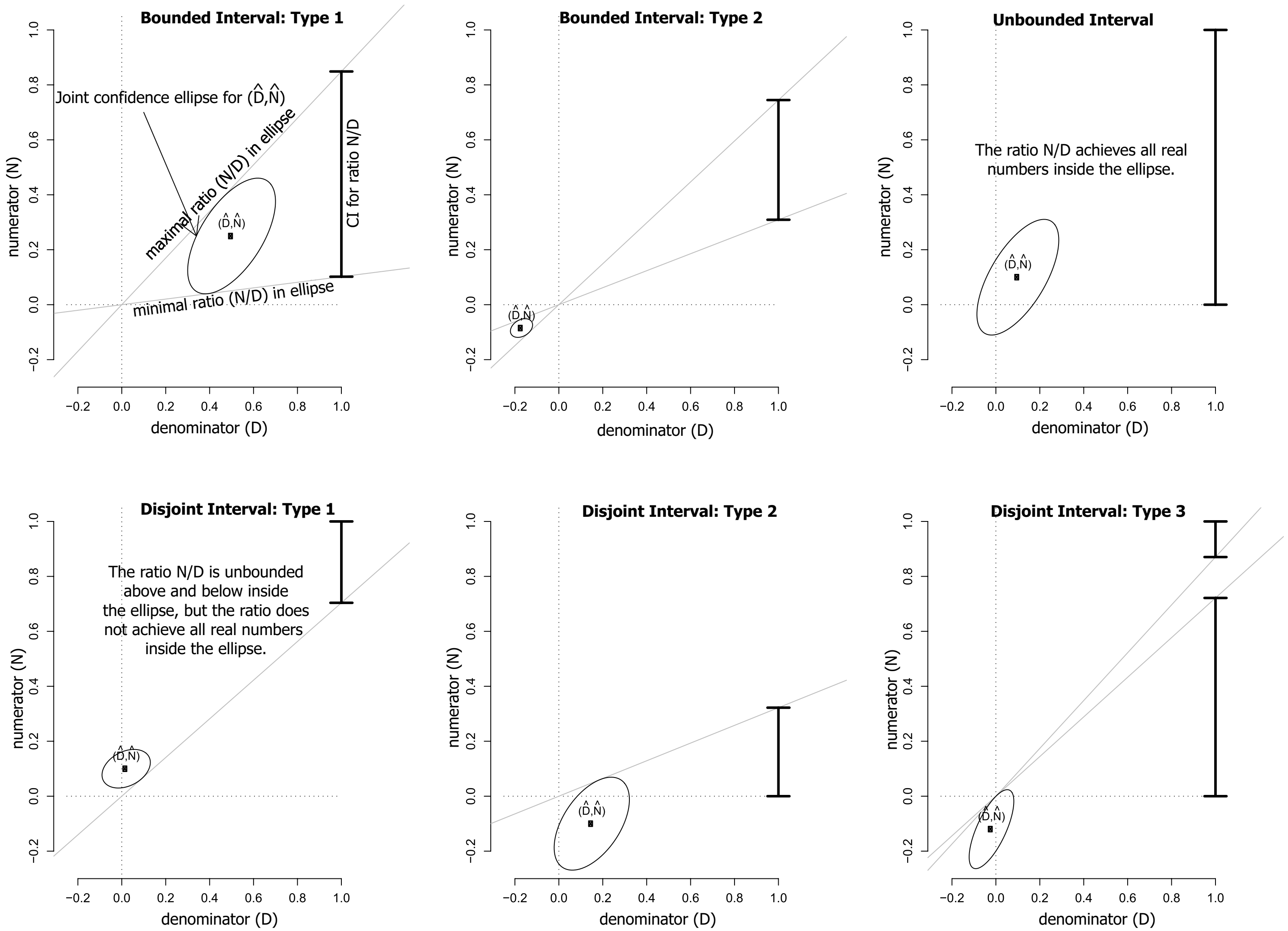
$$\left\{N, D : \begin{pmatrix} N - \widehat{N} \\ D - \widehat{D} \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} N - \widehat{N} \\ D - \widehat{D} \end{pmatrix} \leq z_{\alpha/2}^2 \right\}.$$

Note that the $z_{\alpha/2}$ comes from the approximate normality of $\widehat{N} - \pi\widehat{D}$. An alternate region could be defined using the approximate χ_2^2 distribution of the quadratic form.

Algebraically, Fieller's interval is from

$$P\left(|\frac{\widehat{N} - \pi\widehat{D}}{\widehat{\sigma}_{11} + \pi^2\widehat{\sigma}_{22} - 2\pi\widehat{\sigma}_{12}}| \leq z_{\alpha/2}) = P(f_0 - 2f_1\pi + f_2\pi^2 \leq 0\right) \approx 1 - \alpha$$

since $\widehat{N} - \pi\widehat{D} \xrightarrow{D} N(0, \sigma_{11} + \pi^2\sigma_{22} - 2\pi\sigma_{12})$, where $f_0 = \widehat{N}^2 - z_{\alpha/2}^2\widehat{\sigma}_{11}$, $f_1 = \widehat{D}\widehat{N} - z_{\alpha/2}^2\widehat{\sigma}_{12}$, and $f_2 = \hat{D}^2 - z_{\alpha/2}^2\widehat{\sigma}_{22}$. Further, let $C = f_1^2 - f_2 f_0$, $r1 = (f_1 + \sqrt{C})/f_2$, and $r2 = (f_1 - \sqrt{C})/f_2$. If $C \geq 0$ and $f_2 \geq 0$, then the confidence interval is $[r2, r1] \cap [0, 1]$. If $C \geq 0$ and $f_2 < 0$, then the confidence interval is $[0, r1] \cup [r2, 1]$. If $C < 0$, then the confidence interval is $[0, 1]$. We evaluate the performance of this method in a simulation experiment in Chapter 5.



The intersection of the disjoint interval and [0,1] gives an interval of the form [a,1].

The intersection of the disjoint interval and [0,1] gives an interval of the form [0,a].

The intersection of the disjoint interval and [0,1] gives an interval of the form [0,a]+[b,1].

Figure 2. Geometric Interpretation of Fieller's Method: this figure shows three of the four possible types of confidence sets that Fieller's method can produce. These types and the fourth type are explained in the text.

CHAPTER 4

REGRESSION MODELS WITH MISCLASSIFIED COVARIATES

In this chapter, we will focus on linear regression models with mismeasured discrete covariates, that is, misclassified covariates. It is known that coefficient estimators of a simple regression model with mismeasured covariates will always be biased unless the slope is zero. Section 3.2 of Carroll et al. (2006) and Buonaccorsi et al. (2005), both give a bias expression for a linear regression model with misclassified binary covariates and a possibly perfectly measured univariate covariate.

Akazawa et al. (1998) prove that the corrected score function for a generalized linear model with misclassified covariates exists if the misclassification model $\mathbf{P}$ is known (see Section 2.3 for definition of $\mathbf{P}$). In this case, we can use a corrected score function to obtain asymptotically unbiased estimators for the true coefficients (Nakamura, 1990). We should note that a corrected score function does not always exist for regression models with mismeasured continuous covariates (Nakamura, 1990, Section 4.6).

The corrected score function for a regression model with misclassified covariates will involve $\mathbf{P}^{-1}$. We have proven that $\widehat{\mathbf{P}}^{-1}$ is not an unbiased estimator for $\mathbf{P}^{-1}$ when $\widehat{\mathbf{P}}$ is an unbiased estimator for $\mathbf{P}$ (except in trivial cases), and, in fact, an unbiased estimator of $\mathbf{P}^{-1}$ does not exist (see Section 2.3.1). If the misclassification model $\widehat{\mathbf{P}}$ is estimated from validation data, a "corrected" score function that plugs in $\widehat{\mathbf{P}}^{-1}$ for $\mathbf{P}^{-1}$ without modification, is not a corrected score function.

In this chapter, we will use the approach of Akazawa et al. (1998) and study the impact of corrected estimators using the score function approach when the misclassification model is estimated from external data. We will provide improved methods for estimating the regression coefficients and making inference if a reclassification model is known or estimated. We also develop confidence intervals for the slope of simple linear regression with misclassified binary covariates.

This chapter is organized as follows: first we will establish the notation, then in section 4.1 we will study the bias of the naive least squares estimator for coefficients of linear regression models with misclassified covrariates. We will use Fieller's method to find confidence intervals for the slope of a regression model with misclassified binary data in section 4.2. In section 4.3 we will discuss a corrected score function approach, and in section 4.4 we will explore using the reclassification model to correct the coefficients of regression models with misclassified covariates. Sections 4.1 through 4.4 deal with linear regression models with only misclassified covariates. In section 4.5 we will discuss linear regression with misclassified data and perfectly measured data.

In this chapter, we assume that the categorical data has K categories, from $0, \ldots K-1$. We will use a $K \times 1$ vector with 1 in the position of the category and 0 elsewhere to represent a single categorical random variable. Throughout the chapter, all vectors will be underlined, and all matrices will be bold.

We will use $\underline{X}$ for a true value, and $\underline{W}$ for an observation that is subject to misclassification.

For $\underline{W} = (w_0, \ldots, w_{K-1})^{\mathsf{T}}$, we will refer to $\underline{W} = m$ for $\underline{W}$ in the mth category, that is $w_m = 1$ and $w_j = 0$ for $j \neq m$.

Also for $\underline{X} = (x_0, \ldots, x_{K-1})^{\mathsf{T}}$, we will refer to $\underline{X} = m$ for $\underline{X}$ in the mth category, that is $x_m = 1$ and $x_j = 0$ for $j \neq m$.

We will use the notation $\underline{e}_k$ to denote a $K \times 1$ vector with a 1 in kth position and 0 elsewhere, $k = 0, \ldots K-1$. We also let $\mathbf{M}_{ml}$ be a $K \times K$ matrix with 0 everywhere except in position m, l which is 1 , and in position K, l which contains a -1, $m, l = 0 \ldots K-1$.

As we defined in Section 2.3, $\theta_{ml} = P(\underline{W} = m|\underline{X} = l), \gamma_{ml} = P(\underline{X} = m|\underline{W} = l)$,

$$\mathbf{P} = \begin{pmatrix} \theta_{00} & \theta_{01} & \cdots & \theta_{0(K-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{(K-1)0} & \theta_{(K-1)1} & \cdots & \theta_{(K-1)(K-1)} \end{pmatrix} = (\underline{\theta}_0, \underline{\theta}_1, \ldots, \underline{\theta}_{K-1}), \text{ and}$$

$$\mathbf{Q} = \begin{pmatrix} \gamma_{00} & \gamma_{01} & \cdots & \gamma_{0(K-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{(K-1)0} & \gamma_{(K-1)1} & \cdots & \gamma_{(K-1)(K-1)} \end{pmatrix} = (\underline{\gamma}_0, \underline{\gamma}_1, \ldots, \underline{\gamma}_{K-1}).$$

$\mathbf{P}$ and $\mathbf{Q}$ are the misclassification model and reclassification model respectively. Also we define $\pi_j = P(\underline{X} = j), \lambda_j = P(\underline{W} = j), \underline{\pi} = (\pi_0, \ldots, \pi_{K-1})^{\mathsf{T}}$, and $\underline{\lambda} = (\lambda_0, \ldots, \lambda_{K-1})^{\mathsf{T}}$. We know $\underline{\lambda} = \mathbf{P}\underline{\pi}$ and $\underline{\pi} = \mathbf{Q}\underline{\lambda}$.

From the definition of $\mathbf{P}$ and $\mathbf{Q}$, $\theta_{ml} = P(\underline{W} = m|\underline{X} = l), \gamma_{ml} = P(\underline{X} = m|\underline{W} = l)$, we do not have $\mathbf{P} = \mathbf{Q}^{-1}$. Let $\mathbf{D}_{\underline{\lambda}}$ be a diagonal matrix with $\underline{\lambda}$ along the diagonal and $\mathbf{D}_{\underline{\pi}}$ be a diagonal matrix with $\underline{\pi}$ along the diagonal. Using

$$\theta_{ml} = P(\underline{W} = m|\underline{X} = l) = P(\underline{W} = m)P(\underline{X} = l|\underline{W} = m)/P(\underline{X} = l) = \lambda_m \gamma_{ml}/\pi_l,$$

the relationship between $\mathbf{P}$ and $\mathbf{Q}$ is

$$\mathbf{P} = \mathbf{D}_{\underline{\lambda}} \mathbf{Q}^{\mathsf{T}} \mathbf{D}_{\underline{\pi}}^{-1}, \text{ or } \mathbf{Q} = \mathbf{D}_{\underline{\pi}} \mathbf{P}^{\mathsf{T}} \mathbf{D}_{\underline{\lambda}}^{-1}.$$

In this chapter, we will assume $\mathbf{P}$ is invertible and $\pi_k \neq 0, k = 0 \ldots K-1$.

4.1 Naive Estimators

In this section, we derive the behavior of least squares estimators of linear regression coefficients in the presence of misclassified covariates.

Consider the linear regressions $E(Y|\underline{X}) = \underline{X}^{\mathsf{T}}\underline{\beta}$ and $E(Y|\underline{W}) = \underline{W}^{\mathsf{T}}\underline{\beta}_W$, where $\underline{\beta} = (\beta_0, \ldots, \beta_{K-1})^{\mathsf{T}}$, and $\underline{\beta}_W$ is the coefficient vector under the observed data. Linear regression with covariates that have K categories is like a one factor experimental design model with K levels. As a result, the regression model that we consider can be written as a cell means model. Of course, we could write the regression model as a factor effect model (see Christopher and Kupper (1995) for that approach). Next, we will derive the relationship between $\underline{\beta}$ and $\underline{\beta}_W$ when reclassification model $\mathbf{Q}$ is available.

Theorem 4.1.1 *We consider the observed data $y_i, \underline{W}_i, i = 1 \dots n$ where the dimension of $\underline{W}_i$ is $K \times 1$, and $\underline{W}_i$ is the observed for $\underline{X}_i$. Let*

$$\mathbf{W} = \begin{pmatrix} \underline{W}_1^\top \\ \underline{W}_2^\top \\ \vdots \\ \underline{W}_n^\top \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \underline{X}_1^\top \\ \underline{X}_2^\top \\ \vdots \\ \underline{X}_n^\top \end{pmatrix}, \underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

We have the linear regression $E(Y|\underline{X}) = \underline{X}^\top \underline{\beta}$ and $E(Y|\underline{W}) = \underline{W}^\top \underline{\beta}_W$, then the least squares naive estimator $\widehat{\underline{\beta}}_W = (\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\underline{Y} \xrightarrow{P} \mathbf{Q}^\top\underline{\beta}$.

Proof From the model assumption, $E\left(\widehat{\underline{\beta}}_W|\mathbf{W}, \mathbf{X}\right) = (\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\underline{\beta}$.

Also, $E\left\{(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}|\mathbf{W}\right\} = (\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{W}\mathbf{Q}^\top = \mathbf{Q}^\top$, and we have

$$E\widehat{\underline{\beta}}_W = \mathbf{Q}^\top\underline{\beta}.$$

Therefore, $\widehat{\underline{\beta}}_W$ is an unbiased estimator for $\mathbf{Q}^\top\underline{\beta}$. Since $\widehat{\underline{\beta}}_W$ is a least squares estimator, it is, a consistent estimator for $\underline{\beta}_W$ (see Shaw, 2003, Theorem 3.11 for additional technical conditions), and $\widehat{\underline{\beta}}_W \xrightarrow{P} \mathbf{Q}^\top\underline{\beta}$.

Christopher and Kupper (1995) also have a result that is similar to Theorem 4.1.1, but ours is proven differently.

Remark We note that $\mathbf{W}^\top\mathbf{W}$ is a diagonal matrix with diagonal the number of observed data in each category, that is $E(\mathbf{W}^\top\mathbf{W}) = n\mathbf{D}_{\underline{\lambda}}$. We also observe that

$$\begin{aligned} E(\mathbf{W}^\top\mathbf{X})_{ij} &= \sum_{l=1}^{n} E(w_{il}x_{jl}) = \sum_{l=1}^{n} P(\underline{W}_l = i, \underline{X}_l = j) = nP(\underline{W} = i, \underline{X} = j) \\ &= n\gamma_{ji}\lambda_i = n\left(\mathbf{D}_{\underline{\lambda}}\mathbf{Q}^\top\right)_{ij}, \text{ and we have } E(\mathbf{W}^\top\underline{Y}) = n\mathbf{D}_{\underline{\lambda}}\mathbf{Q}^\top\underline{\beta}. \end{aligned}$$

We could have used $\underline{\beta}_w = \text{Var}(\underline{W})^{-1}\,\text{Cov}(\underline{W}, Y)$to derive the relationship between $\underline{\beta}_W$ and $\underline{\beta}$. But $\text{Var}(\underline{W})$ is singular, since $\underline{W}$ is multinomial distributed with parameter $\underline{\lambda}$. We would need to use a factor effect model, that is, a regression model with intercept.

From Theorem 4.1.1, if $\mathbf{Q}$ is known and invertible, we can have a consistent corrected estimator $\widehat{\underline{\beta}}_c = (\mathbf{Q}^{\mathsf{T}})^{-1}\widehat{\underline{\beta}}_W$. If $\mathbf{Q}$ is estimated from external validation data and $\widehat{\mathbf{Q}}$ is consistent for $\mathbf{Q}$, $(\widehat{\mathbf{Q}}^{\mathsf{T}})^{-1}\widehat{\underline{\beta}}_W$ is not a consistent estimator for $\underline{\beta}$ unless the validation sample size also goes to infinity.

If $\mathbf{P}$ is known, we can use the relationship between $\mathbf{P}$ and $\mathbf{Q}$ and get $\widehat{\mathbf{Q}} = \mathbf{D}_{\widehat{\underline{\pi}}}\mathbf{P}^{\mathsf{T}}\mathbf{D}_{\widehat{\underline{\lambda}}}^{-1}$. Then we can have $\widehat{\underline{\beta}}_c = (\widehat{\mathbf{Q}}^{\mathsf{T}})^{-1}\widehat{\underline{\beta}}_W$ as an estimator for $\underline{\beta}$. If $\mathbf{P}$ is estimated from external data, and $\widehat{\mathbf{P}}$ is an unbiased estimator for $\mathbf{P}$, we still can get $\widehat{\mathbf{Q}} = \mathbf{D}_{\widehat{\underline{\pi}}}\widehat{\mathbf{P}}^{\mathsf{T}}\mathbf{D}_{\widehat{\underline{\lambda}}}^{-1}$, then use the same formula above to get a correction estimator for $\underline{\beta}$. To use this method, we need an estimate of $\underline{\pi}$ (see Chapter 2).

4.2 Confidence Interval for the Slope

In this section, we will do two things. We will derive the relationship between the naive estimators and true coefficients in the simple linear regression model with misclassified binary covariates and a misclassification model. After that, we will use Fieller's method to get a confidence interval for the slope. Note that this is another application of the general method we described in the Section 4.1.

Now we consider $K = 2$ and $\pi = P(X = 1)$.

Corollary 4.2.1 *Consider the model $y = \beta_0 + \beta_1 x + \epsilon$, where $x = 0$ or 1, and $\epsilon \sim N(0, \sigma^2)$. Observe w_i instead of x_i, $w_i = 0$ or 1. Given the observed data and letting $\widehat{\beta}_{wi}$ be the naive least square estimator, then $\lim_{n\to\infty}\widehat{\beta}_{w1} = \dfrac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)}\beta_1$.*

Proof

$$\lim_{n\to\infty}\widehat{\beta}_{w1} \quad = \quad \frac{\text{Cov}(W, Y)}{\text{Var}(W)}$$

$$
\begin{aligned}
&= \frac{E(WY) - EWEY}{\lambda(1-\lambda)} \\
&= \frac{\lambda\beta_0 + \theta_{11}\beta_1\pi - \lambda(\beta_0 + \beta_1\pi)}{\lambda(1-\lambda)} \\
&= \frac{\beta_1(\theta_{00} + \theta_{11} - 1)\pi(1-\pi)}{\lambda(1-\lambda)},
\end{aligned}
$$

where

$$
\begin{aligned}
E(WY) &= E[E\left(WY|X\right)] \\
&= E\left[E\left(W|X\right)\left(\beta_0 + \beta_1 X\right)\right] \\
&= E[\,\{(1-\theta_{00})\left(1-X\right) + \theta_{11}X\}\left(\beta_0 + \beta_1 X\right)] \\
&= \beta_0(1-\theta_{00})(1-\pi) + \theta_{11}\beta_0\pi + \theta_{11}\beta_1\pi \\
&= \lambda\beta_0 + \theta_{11}\beta_1\pi, \\
&\quad \text{from } \lambda = \theta_{11}\pi + (1-\theta_{00})(1-\pi) \text{ and } \theta_{11} - \lambda = (1-\pi)(\theta_{00} + \theta_{11} - 1).
\end{aligned}
$$

We should make a note that when $\lambda = 1$ or $\lambda = 0$, the slope of a naive estimator can not be calculated since then there is a unique value for W. We also note that

$$\beta_1 = E[Y|X=1] - E[Y|X=0].$$

From Corollary 4.2.1, we can make the following remarks:

Remark As a result, the coefficient causing bias in the slope is $\dfrac{(\theta_{00} + \theta_{11} - 1)\pi(1-\pi)}{\lambda(1-\lambda)}$. In simple linear regression with a continuous covariate that suffers from nondifferential additive error, the coefficient that causes bias is an attenuation factor that biases $\widehat{\beta}_1$ toward zero. For the misclassification case, this is also true. We prove that in the following corollary. After the corollary, we also investigate when the inequality in the result is strict.

Corollary 4.2.2

$$\left|\frac{(\theta_{00} + \theta_{11} - 1)\pi(1-\pi)}{\lambda(1-\lambda)}\right| \leq 1.$$

Proof Assume $\left|\frac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)}\right| > 1$, and without loss of generality, we will assume $\theta_{00}+\theta_{11}-1>0$. Using the fact that $\pi = \dfrac{\lambda+\theta_{00}-1}{\theta_{00}+\theta_{11}-1}$, we will have

$$\begin{aligned}
(\theta_{00}+\theta_{11}-1)\pi(1-\pi) &> \lambda(1-\lambda) \\
(\lambda+\theta_{00}-1)(\theta_{11}-\lambda) &> (\theta_{00}+\theta_{11}-1)\lambda(1-\lambda) \\
-\lambda^2+(\theta_{11}-\theta_{00}+1)\lambda-\theta_{11}(1-\theta_{00}) &> (1-\theta_{00}-\theta_{11})\lambda^2+(\theta_{00}+\theta_{11}-1)\lambda \\
(2-\theta_{00}-\theta_{11})\lambda^2-2(1-\theta_{00})\lambda+\theta_{11}(1-\theta_{00}) &< 0.
\end{aligned}$$

Let $f(\lambda)=(2-\theta_{00}-\theta_{11})\lambda^2-2(1-\theta_{00})\lambda+\theta_{11}(1-\theta_{00})$. Then $f(\lambda)$ is a quadratic equation with non-negative leading coefficients. If $2-\theta_{00}-\theta_{11}=0$, we have $\theta_{00}=\theta_{11}=1$, and we will have the result. If $2-\theta_{00}-\theta_{11}>0$, $f(\lambda)$ will achieve a negative value if

$$\begin{aligned}
(1-\theta_{00})^2-(2-\theta_{00}-\theta_{11})\theta_{11}(1-\theta_{00}) &\geq 0 \\
(1-\theta_{00})-(2-\theta_{00}-\theta_{11})\theta_{11} &\geq 0 \text{ or } \theta_{00}=1 \\
\theta_{11}^2-\theta_{11}(1-\theta_{00})+(1-\theta_{00}-\theta_{11}) &\geq 0 \text{ or } \theta_{00}=1 \\
(1-\theta_{00}-\theta_{11})(1-\theta_{11}) &\geq 0 \text{ or } \theta_{00}=1.
\end{aligned}$$

Since $\theta_{00}+\theta_{11}-1>0$, we should have $\theta_{11}=1$ or $\theta_{00}=1$. If $\theta_{11}=1$, we have $\lambda = \theta_{00}\pi+(1-\theta_{00})$, and $1-\lambda=\theta_{00}(1-\pi)$. Then

$$\frac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)} = \frac{\theta_{00}\pi(1-\pi)}{\{\theta_{00}\pi+(1-\theta_{00})\}\theta_{00}(1-\pi)} = \frac{\pi}{\theta_{00}\pi+(1-\theta_{00})} > 1$$

and we will have $\pi>1$ which is impossible. We will do the same argument for $\theta_{00}=1$. So we have

$$\left|\frac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)}\right| \leq 1.$$

Remark $\widehat{\beta}_{w1}$ is an unbiased estimator for β_1 if either $\beta_1=0$ or $\theta_{00}+\theta_{11}-1=1$ (trivial case and no misclassification) or $\dfrac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)}=1$. We will prove the latter case is impossible in the following corollary.

Corollary 4.2.3

$$\frac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)} \neq 1 \textit{ if } \theta_{00}+\theta_{11}-1\neq 1.$$

Proof Assume $\dfrac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)} = 1$, and $\theta_{00}+\theta_{11}-1 \neq 1$. From the assumption, we will have:

$$\theta_{00}+\theta_{11}-1 > 0, \pi \neq 0, \pi \neq 1 \text{ and}$$
$$(\theta_{00}+\theta_{11}-1)\pi(1-\pi) = \lambda(1-\lambda).$$

Since $\pi = \dfrac{\lambda+\theta_{00}-1}{\theta_{00}+\theta_{11}-1}$, we have:

$$\begin{aligned}(\lambda+\theta_{00}-1)(\theta_{11}-\lambda) &= (\theta_{00}+\theta_{11}-1)\lambda(1-\lambda)\\ (2-\theta_{00}-\theta_{11})\lambda^2+2(\theta_{00}-1)\lambda+(1-\theta_{00})\theta_{11} &= 0. \qquad (4.1)\end{aligned}$$

In order to have a solution for λ in Equation 4.1, we should have

$$(1-\theta_{00})^2-\theta_{11}(1-\theta_{00})(2-\theta_{00}-\theta_{11}) = (1-\theta_{00})(1-\theta_{11})(1-\theta_{00}-\theta_{11}) \geq 0.$$

Since $\theta_{00}+\theta_{11}-1>0$, we will have $\theta_{00}=1$ or $\theta_{11}=1$. If $\theta_{00}=1$, then $\lambda=0$ and $\pi=0$. If $\theta_{11}=1$, then $\lambda=1$ and $\pi=1$. This proves that we cannot have the attenuation factor $\dfrac{(\theta_{00}+\theta_{11}-1)\pi(1-\pi)}{\lambda(1-\lambda)} = 1$ for the nontrivial case.

As a result, $\widehat{\beta}_{w1}$ is a always biased estimator for β_1, unless $\beta_1=0$ or there is no misclassification in the data.

Remark If the conditional variance of Y given X is constant, the conditional variance of Y given W is not necessarily constant. It is constant if $P(X=1|W=1)=P(X=1|W=0)$ or $P(X=1|W=1)=P(X=0|W=0)$ (assuming the misclassification is non-differential, i.e., the conditional distribution of $W|X,Y$ is the same as $W|X$). This result can be seen from:

$$\begin{aligned}&\text{Var}(Y|W)\\ =\ & E[\text{ Var}(Y|W,X)|W] + \text{ Var}[E(Y|W,X)|W]\\ =\ & E[\text{ Var}(Y|X)|W] + \text{ Var}[E(Y|X)|W]\\ =\ & \sigma^2+\beta_1^2\left\{P(X=1|W=1)P(X=0|W=1)W+P(X=1|W=0)P(X=0|W=0)(1-W)\right\}\\ =\ & \sigma^2+\beta_1^2\left\{\frac{\theta_{11}\pi\theta_{10}(1-\pi)}{\lambda^2}W+\frac{\theta_{01}\pi\theta_{00}(1-\pi)}{(1-\lambda)^2}(1-W)\right\}.\end{aligned}$$

We can use $\mathrm{Var}(Y) = E[\,\mathrm{Var}(Y|W)] + \mathrm{Var}[E(Y|W)]$ and the above expression of $\mathrm{Var}[Y|W]$ to estimate σ^2.

Also from Corollary 4.2.1, we can define a corrected estimator for the slope

$$\widehat{\beta}_{c1} = \frac{\widehat{\lambda}(1-\widehat{\lambda})}{\widehat{\pi}(1-\widehat{\pi})(\widehat{\theta}_{11}+\widehat{\theta}_{00}-1)}\widehat{\beta}_{w1}.$$

See sections 2.1, 2.2 for methods of estimating π.

We will use Fieller's method to get a confidence interval for the corrected slope

$$\beta_1 = \frac{\lambda(1-\lambda)\beta_{w1}}{\pi(1-\pi)(\theta_{00}+\theta_{11}-1)} = \frac{\lambda(1-\lambda)(\theta_{00}+\theta_{11}-1)\beta_{w1}}{(\lambda+\theta_{00}-1)(\theta_{11}-\lambda)}.$$

See Section 3.2 for a more general discussion of Fieller's method. Let the numerator be $\widehat{N} = \widehat{\lambda}(1-\widehat{\lambda})(\theta_{00}+\theta_{11}-1)\widehat{\beta}_{w1}$ when the misclassification is known, (and $\widehat{N} = \widehat{\lambda}(1-\widehat{\lambda})(\widehat{\theta}_{11}+\widehat{\theta}_{00}-1)\widehat{\beta}_{w1}$ when misclassification is estimated from external data). Let the denominator be $\widehat{D} = (\widehat{\lambda}+\theta_{00}-1)(\theta_{11}-\widehat{\lambda})$ when the misclassification is known, (and $\widehat{D} = (\widehat{\lambda}+\widehat{\theta}_{00}-1)(\widehat{\theta}_{11}-\widehat{\lambda})$ when the misclassification is estimated from external data). Let $\sigma_{11} = \mathrm{Var}(\widehat{N})$, $\sigma_{22} = \mathrm{Var}(\widehat{D})$ and $\sigma_{12} = \mathrm{Cov}(\widehat{N},\widehat{D})$. As before, we compute $f_0 = \widehat{N}^2 - z^2_{\alpha/2}\widehat{\sigma}_{11}, f_1 = \widehat{D}\widehat{N} - z^2_{\alpha/2}\widehat{\sigma}_{12}, f_2 = \widehat{D}^2 - z^2_{\alpha/2}\widehat{\sigma}_{22}, C = f_1^2 - f_2 f_0$ and $r1 = \dfrac{f_1 + C^{.5}}{f_2}, r2 = \dfrac{f_1 - C^{.5}}{f_2}$. If $D \geq 0$ and $f_2 \geq 0$, then $[r1, r2]$ is a $100(1-\alpha)\%$ confidence interval for β_1. If $C \geq 0$ and $f_2 < 0$,then $(-\infty, r2] \cup [r1, \infty)$ is a $100(1-\alpha)\%$ confidence interval for β_1. If $C < 0$, then the confidence interval is $(-\infty, \infty)$.

When the misclassification is estimated from external data. We can rewrite

$$\begin{aligned}\widehat{N} &= \frac{(\widehat{\theta}_{11}+\widehat{\theta}_{00}-1)\sum_i Y_i(W_i-\widehat{\lambda})}{n} = \frac{Z_1(\widehat{\theta}_{11},\widehat{\theta}_{00})Z_2(\mathbf{Y},\mathbf{W})}{n},\\ \widehat{D} &= Z_3(\widehat{\lambda}) + Z_4(\widehat{\lambda},\widehat{\theta}_{11},\widehat{\theta}_{00}) + Z_5(\widehat{\theta}_{11},\widehat{\theta}_{00}),\end{aligned}$$

where

$$\begin{aligned}Z_1(\widehat{\theta}_{11},\widehat{\theta}_{00}) &= \widehat{\theta}_{11}+\widehat{\theta}_{00}-1,\\ Z_2(\mathbf{Y},\mathbf{W}) &= \sum_i Y_i(W_i-\widehat{\lambda})\end{aligned}$$

$$\begin{aligned} Z_3(\widehat{\theta}_{11}, \widehat{\theta}_{00}) &= \widehat{\theta}_{11}(\widehat{\theta}_{00} - 1), \\ Z_4(\widehat{\lambda}, \widehat{\theta}_{11}, \widehat{\theta}_{00}) &= (\widehat{\theta}_{11} - \widehat{\theta}_{00})\widehat{\lambda} \text{ and} \\ Z_5(\widehat{\lambda}) &= \widehat{\lambda}(1 - \widehat{\lambda}). \end{aligned}$$

The following lemmas develop the computations we need for $\sigma_{11} = \text{Var}(\widehat{N}), \sigma_{22} = \text{Var}(\widehat{D})$ and $\sigma_{12} = \text{Cov}(\widehat{N}, \widehat{D})$.

Lemma 4.2.4 *$Var(h_1 h_2) = Var(h_1)Var(h_2) + Var(h_1)[E(h_2)]^2 + [E(h_1)]^2$ Var(h_2) where h_1, and h_2 are independent random variables.*

Proof

$$\begin{aligned} \text{Var}(h_1 h_2) &= E[\text{ Var}(h_1 h_2 | h_1)] + \text{ Var}[E\,(h_1 h_2 | h_1)\,] = E(h_1^2) \text{ Var}(h_2) + \text{ Var}[h_1 E\,(h_2)\,] \\ &= [\text{ Var}(h_1) + (Eh_1)^2] \text{ Var}(h_2) + \text{ Var}(h_1)(Eh_2)^2. \end{aligned}$$

Lemma 4.2.5 *If (h_1, h_2) and (g_1, g_2) are independent, then*

$Cov(h_1 g_1, h_2 g_2) = Cov(h_1, h_2)Cov(g_1, g_2) + E(h_1)E(h_2)Cov(g_1, g_2) + Cov(h_1, h_2)E(g_1)E(g_2).$

Proof

$$\begin{aligned} \text{Cov}(h_1 g_1, h_2 g_2) &= E[\text{ Cov}\,(h_1 g_1, h_2 g_2 | h_1, h_2)\,] + \text{ Cov}[E\,(h_1 g_1 | h_1, h_2)\,, E\,(h_2 g_2 | h_1, h_2)\,] \\ &= E(h_1 h_2) \text{ Cov}(g_1, g_2) + \text{ Cov}(h_1, h_2)E(g_1)E(g_2) \\ &= \text{ Cov}(h_1, h_2) \text{ Cov}(g_1, g_2) + E(h_1)E(h_2) \text{ Cov}(g_1, g_2) + \text{ Cov}(h_1, h_2)E(g_1)E(g_2). \end{aligned}$$

Lemma 4.2.6

$$\begin{aligned} E(Z_2) &= (n-1)\lambda(1-\lambda)\beta_{w1}, \\ Var(Z_2) &= \beta_{w1}^2 Var[n\widehat{\lambda}(1-\widehat{\lambda})] + \left\{\sigma^2 + \frac{\beta_1^2 \theta_{01}\pi\theta_{00}(1-\pi)}{(1-\lambda)^2}(n-1)\lambda(1-\lambda)\right\} \\ &\quad + n\beta_1^2 \left\{\frac{\theta_{11}\pi\theta_{10}(1-\pi)}{\lambda^2} - \frac{\theta_{01}\pi\theta_{00}(1-\pi)}{(1-\lambda)^2}\right\} E(\widehat{\lambda} - 2\widehat{\lambda}^2 + \widehat{\lambda}^3) \\ Cov(Z_2, \widehat{\lambda}) &= n\beta_{w1} Cov[\widehat{\lambda}(1-\widehat{\lambda}), \widehat{\lambda}] \\ Cov[Z_2, \widehat{\lambda}(1-\widehat{\lambda})] &= n\beta_{w1} Var[\widehat{\lambda}(1-\widehat{\lambda})] \end{aligned}$$

Proof

$$
\begin{aligned}
E(Z_2) &= E\left[E\left\{\sum_i Y_i(W_i-\widehat{\lambda})|\mathbf{W}\right\}\right] \\
&= E\left[\sum_i (W_i-\widehat{\lambda})(\beta_{w0}+\beta_{w1}W_i)\right] \\
&= n\beta_{w1}E[\widehat{\lambda}(1-\widehat{\lambda})] \\
&= = (n-1)\beta_{w1}\lambda(1-\lambda) \\
\text{Var}(Z_2) &= E\left[\ \text{Var}\left\{\sum_i Y_i(W_i-\widehat{\lambda})|\mathbf{W}\right\}\right] +\ \text{Var}\left[E\left\{\sum_i Y_i(W_i-\widehat{\lambda})|\mathbf{W}]\right.\right] \\
&= E\left[\sum_i (W_i-\widehat{\lambda})^2\left\{\sigma^2+\beta_1^2[\frac{\theta_{11}\pi\theta_{10}(1-\pi)}{\lambda^2}W_i+\frac{\theta_{01}\pi\theta_{00}(1-\pi)}{(1-\lambda)^2}(1-W_i)]\right\}\right] \\
&\quad + \text{Var}\left[\sum_i(\beta_{w0}+\beta_{w1}W_i)(W_i-\widehat{\lambda})\right] \\
&= \beta_{w1}^2\ \text{Var}\left[n\widehat{\lambda}(1-\widehat{\lambda})\right]+\left\{\sigma_2+\frac{\beta_1^2\theta_{01}\pi\theta_{00}(1-\pi)}{(1-\lambda)^2}\right\}(n-1)\lambda(1-\lambda) \\
&\quad + n\beta_1^2\left\{\frac{\theta_{11}\pi\theta_{10}(1-\pi)}{\lambda^2}-\frac{\theta_{01}\pi\theta_{00}(1-\pi)}{(1-\lambda)^2}\right\}E(\widehat{\lambda}-2\widehat{\lambda}^2+\widehat{\lambda}^3) \\
\text{Cov}(Z_2,\widehat{\lambda}) &= E\left[\ \text{Cov}\left\{\sum_i Y_i(W_i-\widehat{\lambda}),\widehat{\lambda}\middle|\mathbf{W}\right\}\right] +\ \text{Cov}\left[E\left\{\sum_i Y_i(W_i-\widehat{\lambda})|\mathbf{W}\right\},E\left\{\widehat{\lambda}|\mathbf{W}\right\}\right] \\
&= \ \text{Cov}\left[\sum_i(\beta_{w0}+\beta_{w1}W_i)(W_i-\widehat{\lambda}),\widehat{\lambda}\right] \\
&= n\beta_{w1}\ \text{Cov}[\widehat{\lambda}(1-\widehat{\lambda}),\widehat{\lambda}]
\end{aligned}
$$

Lemma 4.2.7

$$
\begin{aligned}
E\widehat{\lambda}^3 &= \frac{(n-1)(n-2)\lambda^3+3(n-1)\lambda^2+\lambda}{n^2} \\
E\widehat{\lambda}^4 &= \frac{(n-1)(n-2)(n-3)\lambda^4+6(n-1)(n-2)\lambda^3+7(n-1)\lambda^2+\lambda}{n^3} \\
\mathit{Cov}[\widehat{\lambda}(1-\widehat{\lambda}),\widehat{\lambda}] &= \frac{\lambda(1-\lambda)(1+\lambda)}{n}+\lambda^3-E\widehat{\lambda}^3 \\
\mathit{Var}[\widehat{\lambda}(1-\widehat{\lambda})] &= \frac{\lambda(1-\lambda)}{n}+2\left\{\lambda^3+\frac{\lambda^2(1-\lambda)}{n}-E\widehat{\lambda}^3\right\}+E\widehat{\lambda}^4-\left\{\lambda^2+\frac{\lambda(1-\lambda)}{n}\right\}^2
\end{aligned}
$$

From those lemmas, when the misclassification model is estimated, we will have

$$
\sigma_{11} = \frac{\left\{\frac{\theta_{11}(1-\theta_{11})}{N_{.1}}+\frac{\theta_{00}(1-\theta_{00})}{N_{.0}}\right\}\left\{\ \text{Var}(Z_2)+(EZ_2)^2\right\}+(\theta_{00}+\theta_{11}-1)^2\ \text{Var}(Z_2)}{n^2}
$$

$$
\begin{aligned}
\sigma_{12} &= \frac{\operatorname{Cov}(Z_1Z_2, Z_3) + \operatorname{Cov}(Z_1Z_2, Z_4) + \operatorname{Cov}(Z_1Z_2, Z_5)}{n} \\
&= \frac{1}{n}\Big\{E(Z_2)\operatorname{Cov}(Z_1, Z_3) + E(Z_1)\operatorname{Cov}(Z_2, Z_5) + \operatorname{Cov}(Z_1, \widehat{\theta}_{11} - \widehat{\theta}_{00})\operatorname{Cov}(Z_2, \widehat{\lambda}) \\
&\quad + E(Z_1)(\theta_{11} - \theta_{00})\operatorname{Cov}(Z_2, \widehat{\lambda}) + \operatorname{Cov}(Z_1, \widehat{\theta}_{11} - \widehat{\theta}_{00})E(Z_2)\lambda\Big\} \\
&= \frac{1}{n}\Big[E(Z_2)\Big\{\operatorname{Var}(\widehat{\theta}_{11})(\theta_{00} - 1) + \theta_{11}\operatorname{Var}(\widehat{\theta}_{00})\Big\} + (\theta_{00} + \theta_{11} - 1)\operatorname{Cov}(Z_2, Z_5) \\
&\quad + \Big\{\operatorname{Var}(\widehat{\theta}_{11}) - \operatorname{Var}(\widehat{\theta}_{00})\Big\}\Big\{\operatorname{Cov}(Z_2, \widehat{\lambda}) + \lambda E(Z_2)\Big\} \\
&\quad + (\theta_{00} + \theta_{11} - 1)(\theta_{11} - \theta_{00})\operatorname{Cov}(Z_2, \widehat{\lambda})\Big] \\
\sigma_{22} &= \operatorname{Var}(Z_3) + \operatorname{Var}(Z_4) + \operatorname{Var}(Z_5) + 2\{\operatorname{Cov}(Z_3, Z_4) + \operatorname{Cov}(Z_3, Z_5) + \operatorname{Cov}(Z_4, Z_5)\} \\
&= \operatorname{Var}(\widehat{\theta}_{11})\Big\{\operatorname{Var}(\widehat{\theta}_{00}) + (1 - \theta_{00})^2\Big\} + \theta_{11}^2\operatorname{Var}(\widehat{\theta}_{00}) \\
&\quad + \Big\{\operatorname{Var}(\widehat{\theta}_{11}) + \operatorname{Var}(\theta_{00})\Big\}\Big\{\operatorname{Var}(\widehat{\lambda}) + \lambda^2\Big\} + (\theta_{11} - \theta_{00})^2\operatorname{Var}(\widehat{\lambda}) + \operatorname{Var}[\widehat{\lambda}(1 - \widehat{\lambda})] \\
&\quad + 2\Big\{(\theta_{00} - 1)\operatorname{Var}(\widehat{\theta}_{11}) - \theta_{11}\operatorname{Var}(\widehat{\theta}_{00})\Big\}\lambda + 2(\theta_{11} - \theta_{00})\operatorname{Cov}(\widehat{\lambda}, 1 - \widehat{\lambda}) \\
&= \frac{\theta_{11}(1 - \theta_{11})\theta_{00}(1 - \theta_{00})}{N_{.1}N_{.0}} + \frac{\theta_{00}(1 - \theta_{00})}{N_{.0}}\left\{\frac{\lambda(1 - \lambda)}{n} + (\lambda - \theta_{11})^2\right\} \\
&\quad + 2(\theta_{11} - \theta_{00})\operatorname{Cov}[\widehat{\lambda}, \widehat{\lambda}(1 - \widehat{\lambda})] + \frac{\theta_{11}(1 - \theta_{11})}{N_{.1}}\left\{\frac{\lambda(1 - \lambda)}{n} + (1 - \lambda - \theta_{00})^2\right\} \\
&\quad + (\theta_{11} - \theta_{00})^2\operatorname{Var}(\widehat{\lambda}) + \operatorname{Var}[\widehat{\lambda}(1 - \widehat{\lambda})]
\end{aligned}
$$

When the misclassification matrix is known, using the above lemmas, we will have :

$$
\begin{aligned}
\sigma_{11} &= \frac{(\theta_{00} + \theta_{11} - 1)^2\operatorname{Var}(Z_2)}{n^2}, \\
\sigma_{12} &= \frac{(\theta_{00} + \theta_{11} - 1)\operatorname{Cov}[Z_2, \widehat{\lambda}(1 - \widehat{\lambda})] + (\theta_{00} + \theta_{11} - 1)(\theta_{11} - \theta_{00})\operatorname{Cov}(Z_2, \widehat{\lambda})}{n}, \text{ and} \\
\sigma_{22} &= \operatorname{Var}[\widehat{\lambda}(1 - \widehat{\lambda})] + (\theta_{11} - \theta_{00})^2\operatorname{Var}(\widehat{\lambda}) + 2(\theta_{11} - \theta_{00})\operatorname{Cov}[\widehat{\lambda}(1 - \widehat{\lambda}), \widehat{\lambda}].
\end{aligned}
$$

4.3 Score Function Approach

In this section, we will first use the result from Akazawa et al. (1998) to get the corrected score for linear regression model with misclassified covariates when the misclassification model $\mathbf{P}$ is known, and show how it can be used to estimate $\underset{\sim}{\beta}$. Then we extend the approach to the case when the misclassification model is estimated from external data.

Assume the linear regression model $Y = \underline{X}^{\mathsf{T}}\underline{\beta} + \epsilon = \sum\limits_{k=0}^{K-1} x_k\beta_k + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ with unknown σ^2. Instead of observing $\underline{X}$, we observe $\underline{W} = (w_0, ..., w_{K-1})^{\mathsf{T}}$.

Assume we have data $(y_1, \underline{W}_1), (y_2, \underline{W}_2), \ldots, (y_n, \underline{W}_n)$. Let

$$\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{W}^{\mathsf{T}} = \begin{pmatrix} \underline{W}_1^{\mathsf{T}} \\ \underline{W}_2^{\mathsf{T}} \\ \vdots \\ \underline{W}_n^{\mathsf{T}} \end{pmatrix}.$$

Then $\ell(\underline{\beta}, \underline{Y}, \mathbf{W}) = -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum\limits_{i=1}^{n}(y_i - \underline{W}_i^{\mathsf{T}}\underline{\beta})^2$ is the log-likelihood function, $S(\underline{\beta}, \underline{Y}, \mathbf{W}) = \frac{\partial \ell}{\partial \underline{\beta}}$ is the score function, and $\mathbf{I}(\underline{\beta}, \mathbf{W}) = E\frac{\partial S}{\partial \underline{\beta}}$ is the Fisher information. We should note that the solution of $S(\underline{\beta}, \underline{Y}, \mathbf{W}) = \underline{0}$ is an estimator of $\underline{\beta}_W$, and it is often biased for $\underline{\beta}$.

We note that $g(\underline{\beta}, \underline{Y}, \mathbf{W})$ is called a corrected log likelihood function if

$$E[g(\underline{\beta}, \underline{Y}, \mathbf{W})|\underline{Y}, \mathbf{X}] = \ell(\underline{\beta}, \underline{Y}, \mathbf{X}),$$

for $\underline{\beta}$ in an open convex subset of the parameter space and where $\mathbf{X} = (\underline{X}_1, \underline{X}_2 \ldots, \underline{X}_n)$ are the true (unobserved) values of $\mathbf{W}$. In this case, $\dfrac{\partial g(\underline{\beta}, \underline{Y}, \mathbf{W})}{\partial \underline{\beta}}$ is called a corrected score function (Nakamura, 1990).

Let

$$\ell_{\mathbf{P}}(\underline{\beta}, \underline{Y}, \mathbf{W}) = -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}\left[y_i^2 - 2y_i\underline{\beta}^{\mathsf{T}}\mathbf{P}^{-1}\underline{W}_i + \underline{\beta}^{\mathsf{T}}\sum_{k=0}^{K-1}\left\{(\mathbf{P}^{-1}\underline{W}_i)^{\mathsf{T}}\underline{e}_k\right\}\underline{e}_k\underline{e}_k^{\mathsf{T}}\underline{\beta}\right],$$

where $\mathbf{P}$ is the known misclassification model. We note that $\sum\limits_{i=1}^{n}\sum\limits_{k=0}^{K-1}\left\{(\mathbf{P}^{-1}\underline{W}_i)^{\mathsf{T}}\underline{e}_k\right\}\underline{e}_k\underline{e}_k^{\mathsf{T}}$ is the diagonal matrix with estimated true category frequencies on the diagonal, that is

$$\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left\{(\mathbf{P}^{-1}\underline{W}_i)^{\mathsf{T}}\underline{e}_k\right\}\underline{e}_k\underline{e}_k^{\mathsf{T}} = n\mathbf{D}_{\widehat{\underline{\pi}}} \text{ with } \widehat{\underline{\pi}} = \mathbf{P}^{-1}\widehat{\underline{\lambda}}.$$

Using that fact that $E[\underline{W}_i|\underline{X}_i = m] = \begin{pmatrix} \theta_{0m} \\ \vdots \\ \theta_{(K-1)m} \end{pmatrix} = \underline{\theta}_m$, and $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I} = \mathbf{P}^{-1}(\underline{\theta}_0, \underline{\theta}_1, \ldots, \underline{\theta}_{K-1})$, we have $\mathbf{P}^{-1}\underline{\theta}_m = \underline{e}_m$ and $E[\mathbf{P}^{-1}\underline{W}_i|\underline{X}_i = m] = \mathbf{P}^{-1}\underline{\theta}_m = \underline{e}_m$. So,

$$
\begin{aligned}
& E[\ell_{\mathbf{P}}(\underline{\beta}, \underline{Y}, \mathbf{W})|\underline{Y}, \mathbf{X}] \\
= \; & -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n} E\Big[y_i^2 - 2y_i\underline{\beta}^\mathsf{T}\mathbf{P}^{-1}\underline{W}_i + \underline{\beta}^\mathsf{T}\sum_{k=0}^{K-1}\{(\mathbf{P}^{-1}\underline{W}_i)^\mathsf{T}\underline{e}_k\}\underline{e}_k\underline{e}_k^\mathsf{T}\underline{\beta}|y_i, \underline{X}_i\Big] \\
= \; & -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}\sum_{m=0}^{K-1} E\Big[y_i^2 - 2y_i\underline{\beta}^\mathsf{T}\mathbf{P}^{-1}\underline{W}_i \\
& + \underline{\beta}^\mathsf{T}\sum_{k=0}^{K-1}\Big\{(\mathbf{P}^{-1}\underline{W}_i)^\mathsf{T}\underline{e}_k\Big\}\underline{e}_k\underline{e}_k^\mathsf{T}\underline{\beta}|y_i, \underline{X}_i = m\Big]\mathbf{1}_{\underline{\mathbf{X}}_\mathbf{i}=\mathbf{m}} \\
= \; & -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}\sum_{m=0}^{K-1}\Big[y_i^2 - 2y_i\underline{\beta}\underline{e}_m + \underline{\beta}^\mathsf{T}\sum_{k=0}^{K-1}\{(\underline{e}_m)^\mathsf{T}\underline{e}_k\}\underline{e}_k\underline{e}_k^\mathsf{T}\underline{\beta}\Big]\mathbf{1}_{\underline{\mathbf{X}}_\mathbf{i}=\mathbf{m}} \\
= \; & -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}\sum_{m=0}^{K-1}(y_i^2 - 2y_i\beta_m + \beta_m^2)\mathbf{1}_{\underline{\mathbf{X}}_\mathbf{i}=\mathbf{m}} \qquad (4.2) \\
= \; & -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}(y_i - \underline{X}_i^\mathsf{T}\underline{\beta})^2 \\
= \; & \ell(\underline{\beta}, \sigma^2\underline{Y}, \mathbf{X})
\end{aligned}
$$

Therefore $\ell_{\mathbf{P}}(\underline{\beta}, \underline{Y}, \mathbf{W}, \mathbf{P})$ is a corrected log likelihood function and

$$
\begin{aligned}
S_{\mathbf{P}}(\underline{\beta}, \underline{Y}, \mathbf{W}, \mathbf{P}) &= \frac{\partial \ell_{\mathbf{P}}}{\partial \underline{\beta}} \\
&= \sigma^{-2}\Big[\sum_{i=1}^{n} y_i\mathbf{P}^{-1}\underline{W}_i - \sum_{i=1}^{n}\sum_{k=0}^{K-1}\Big\{(\mathbf{P}^{-1}\underline{W}_i)^\mathsf{T}\underline{e}_k\Big\}\underline{e}_k\underline{e}_k^\mathsf{T}\underline{\beta}\Big] \\
&= \sigma^{-2}(\mathbf{P}^{-1}\mathbf{W}^\mathsf{T}\underline{Y} - n\mathbf{D}_{\underline{\widehat{\pi}}}\underline{\beta})
\end{aligned}
$$

is a corrected score function for $\underline{\beta}$. The solution of $S_{\mathbf{P}}(\underline{\beta}, \underline{Y}, \mathbf{W}, \mathbf{P}) = \underline{0}$ is

$$
\begin{aligned}
\underline{\widehat{\beta}}_{\mathbf{P}} &= \Big[\sum_{i=1}^{n}\sum_{k=0}^{K-1}\Big\{(\mathbf{P}^{-1}\underline{W}_i)^\mathsf{T}\underline{e}_k\Big\}\underline{e}_k\underline{e}_k^\mathsf{T}\Big]^{-1}(\sum_{i=1}^{n} y_i\mathbf{P}^{-1}\underline{W}_i) \qquad (4.3) \\
&= \mathbf{D}_{\underline{\widehat{\pi}}}^{-1}\mathbf{P}^{-1}\mathbf{W}^\mathsf{T}\underline{Y}/n.
\end{aligned}
$$

This gives an asymptotically unbiased estimator for $\underline{\beta}$. Actually $\underline{\widehat{\beta}}_{\mathbf{P}}$ is an unbiased estimator for $\underline{\beta}$ (further explanation follows in the remark at the end of this section). Nakamura (1990) proves that under certain regularity conditions, the solution of a corrected score function is asymptotically normal with mean $\underline{\beta}$, the true parameter, and covariance matrix $\mathbf{I}(\underline{\beta},\mathbf{W})^{-1}\mathbf{C}(\underline{\beta},\underline{W})\mathbf{I}(\underline{\beta},\mathbf{W})^{-1}$ where $\mathbf{C}(\underline{\beta},\underline{W})$ is the unconditional covariance matrix of $S_{\mathbf{P}}$.

Next, we will find the estimator for σ^2, then use Nakamura (1990) to find the covariance matrix of $\underline{\widehat{\beta}}_{\mathbf{P}}$. To make notation simpler, we will let $\underline{\widehat{X}}_i = \mathbf{P}^{-1}\underline{W}_i$, then

$$\sum_{k=0}^{K-1}\left\{(\mathbf{P}^{-1}\underline{W}_i)^{\intercal}\underline{e}_k\right\}\underline{e}_k\underline{e}_k^{\intercal} = \mathbf{D}_{\underline{\widehat{X}}_i}.\text{Letting}$$

$$\widehat{\mathbf{X}} = \begin{pmatrix}\underline{\widehat{X}}_1^{\intercal}\\ \dots \\ \underline{\widehat{X}}_n^{\intercal}\end{pmatrix} = \mathbf{W}(\mathbf{P}^{-1})^{\intercal},$$

by differentiating corrected loglikelihood function $l_{\mathbf{P}}$, an estimate for σ^2 is

$$\widehat{\sigma^2_{\mathbf{P}}} = \sum_{i=1}^{n}(y_i^2 - 2y_i\underline{\widehat{\beta}}_{\mathbf{P}}^{\intercal}\underline{\widehat{\mathbf{X}}}_i + \underline{\widehat{\beta}}_{\mathbf{P}}^{\intercal}\mathbf{D}_{\underline{\widehat{\mathbf{X}}}_i}\underline{\widehat{\beta}}_{\mathbf{P}})/n.$$

The corrected observed information $\mathbf{I_P}$ is

$$\mathbf{I_P}(\underline{\beta},\underline{\mathbf{Y}},\mathbf{W}) = \widehat{\sigma}_{\mathbf{P}}^{-2}\sum_{i=1}^{n}\sum_{k=0}^{K-1}\left\{(\mathbf{P}^{-1}\underline{W}_i)^{\intercal}\underline{e}_k\right\}\underline{e}_k\underline{e}_k^{\intercal} = n\mathbf{D}_{\underline{\widehat{\pi}}}/\widehat{\sigma}^2_{\mathbf{P}}.$$

According to Nakamura (1990), there are two ways to estimate the covariance matrix of $\underline{\widehat{\beta}}_{\mathbf{P}}$. We will first use the simple one. Let

$$\mathbf{V}(\underline{\beta},\underline{Y},\mathbf{W}) = \sum_{i=1}^{n}\left[y_i\underline{\widehat{X}}_i - \mathbf{D}_{\underline{\widehat{X}}_i}\underline{\beta}\right]\left[y_i\underline{\widehat{X}}_i - \mathbf{D}_{\underline{\widehat{X}}_i}\underline{\beta}\right]^{\intercal}/\widehat{\sigma}^4_{\mathbf{P}},$$

then the asymptotic covariance matrix of $\underline{\widehat{\beta}}_{\mathbf{P}}$ is

$$\begin{aligned}\widehat{\Sigma}_{1(\underline{\widehat{\beta}}_{\mathbf{P}})} &= \mathbf{I_P}(\underline{\widehat{\beta}}_{\mathbf{P}},\underline{\mathbf{Y}},\mathbf{W})^{-1}\mathbf{V}(\underline{\widehat{\beta}}_{\mathbf{P}},\underline{Y},\mathbf{W})\mathbf{I_P}(\underline{\widehat{\beta}}_{\mathbf{P}},\underline{\mathbf{Y}},\mathbf{W})^{-1}\\ &= \mathbf{D}_{\underline{\widehat{\pi}}}^{-1}\sum_{i=1}^{n}\left[y_i\underline{\widehat{X}}_i - \mathbf{D}_{\underline{\widehat{X}}_i}\underline{\beta}\right]\left[y_i\underline{\widehat{X}}_i - \mathbf{D}_{\underline{\widehat{X}}_i}\underline{\beta}\right]^{\intercal}\mathbf{D}_{\underline{\widehat{\pi}}}^{-1}/n^2,\end{aligned}$$

which is equation (4) of Nakamura (1990).

Nakamura (1990) points out that if there is $\omega(\underline{\beta},y,\underline{W})$ such that

$$E[\omega(\underline{\beta},y,\underline{W})|y,\underline{X}] = S(\underline{\beta},y,\underline{X})S(\underline{\beta},y,\underline{X})^{\intercal}$$

where $S(\underline{\beta}, y, \underline{X}) = \frac{\partial \ell(\underline{\beta}, y, \underline{X})}{\partial \underline{\beta}}$ and $\ell(\underline{\beta}, y, \underline{X})$ is the log-likelihood function of $\underline{\beta}$ given data $y, \underline{X}$, then we can use

$$\begin{aligned}\Sigma_{2(\widehat{\underline{\beta}}_{\mathbf{P}})} &= \mathbf{I_P}(\widehat{\underline{\beta}}_{\mathbf{P}}, \underline{\mathbf{Y}}, \mathbf{W})^{-1} \left\{ \mathbf{V}(\widehat{\underline{\beta}}_{\mathbf{P}}, \underline{Y}, \mathbf{W}) - \sum_{i=1}^{n} \omega(\underline{\beta}_{\mathbf{P}}, y_i, \underline{W}_i) \right\} \mathbf{I_P}(\widehat{\underline{\beta}}_{\mathbf{P}}, \underline{\mathbf{Y}}, \mathbf{W})^{-1} \\ &\quad + \mathbf{I_P}(\widehat{\underline{\beta}}_{\mathbf{P}}, \underline{\mathbf{Y}}, \mathbf{W})^{-1}\end{aligned}$$

as an asymptotic covariance matrix for $\widehat{\underline{\beta}}_{\mathbf{P}}$. Nakamura (1990) also uses simulation to demonstrate that the covariance matrix involving $\omega(\underline{\beta}, y, \underline{W})$ is more efficient than using the other one.

We should note that such $\omega(\underline{\beta}, y, \underline{W})$ is not always available. Also in our case

$$\begin{aligned} S(\underline{\beta}, y, \underline{X}) &= (y - \underline{X}^{\mathsf{T}}\underline{\beta})\underline{X}/\sigma^2, \\ S(\underline{\beta}, y, \underline{X})S(\underline{\beta}, y, \underline{X})^{\mathsf{T}} &= (y - \underline{X}^{\mathsf{T}}\underline{\beta})^2 \underline{X}\,\underline{X}^{\mathsf{T}}/\sigma^4. \end{aligned}$$

In the following lemma, we will prove $\omega(\underline{\beta}, y, \underline{W})$ exists for our case.

Lemma 4.3.1 *Let*

$$\omega(\underline{\beta}, y, \underline{W}) = \sum_{k=0}^{K-1} \left[\left\{ y^2 - 2y\underline{\beta}^{\mathsf{T}}\underline{e}_k + \underline{\beta}^{\mathsf{T}}\underline{e}_k\underline{e}_k^{\mathsf{T}}\underline{\beta} \right\} (\mathbf{P}^{-1}\underline{W})^{\mathsf{T}}\underline{e}_k \right] \underline{e}_k\underline{e}_k^{\mathsf{T}}/\sigma^4.$$

Then

$$E[\omega(\underline{\beta}, y, \underline{W})|y, \underline{X}] = (y - \underline{X}^{\mathsf{T}}\underline{\beta})^2 \underline{X}\,\underline{X}^{\mathsf{T}}/\sigma^4.$$

Proof Following the lines in equations 4.2, we should have

$$E\left[\sum_{k=0}^{K-1} \left\{ \mathbf{P}^{-1}\underline{W})^{\mathsf{T}}\underline{e}_k \right\} \underline{e}_k\underline{e}_k^{\mathsf{T}} | y, \underline{X} \right] = \underline{X}\,\underline{X}^{\mathsf{T}}$$

and $\underline{\beta}^{\mathsf{T}}\underline{e}_k = \beta_k$, so we have $E[\omega(\underline{\beta}, y, \underline{W})|y, \underline{X}] = (y^2 - 2y\underline{\beta}\underline{X} + \underline{\beta}\underline{X}\underline{\beta}\underline{X})\underline{X}\,\underline{X}^{\mathsf{T}}/\sigma^4$.

We should note that

$$\omega(\underline{\beta}, y, \underline{W}) \neq \left[y\mathbf{P}^{-1}\underline{W} - \sum_{k=0}^{K-1} \left\{ (\mathbf{P}^{-1}\underline{W})^{\mathsf{T}}\underline{e}_k \right\} \underline{e}_k\underline{e}_k^{\mathsf{T}} \right] \left[y\mathbf{P}^{-1}\underline{W} - \sum_{k=0}^{K-1} \left\{ (\mathbf{P}^{-1}\underline{W})^{\mathsf{T}}\underline{e}_k \right\} \underline{e}_k\underline{e}_k^{\mathsf{T}} \right]^{\mathsf{T}} /\sigma^4,$$

and $E[S_{\mathbf{P}}(\underline{\beta}, y, \underline{W}, \mathbf{P})S_{\mathbf{P}}(\underline{\beta}, y, \underline{W}, \mathbf{P})^{\mathsf{T}}|y, \underline{X}] \neq S(\underline{\beta}, y, \underline{X})S(\underline{\beta}, y, \underline{X})^{\mathsf{T}}$.

We will use simulation to to compare the efficiency of $\widehat{\Sigma}_{1(\widehat{\underline{\beta}}_{\mathbf{P}})}$ and $\widehat{\Sigma}_{1(\widehat{\underline{\beta}}_{\mathbf{P}})}$ in Chapter 5.

Remark We could regress y_i on $\widehat{\underline{X}}_i$ and estimate regression coefficients. This is a regression calibration method. The least squares estimator for regression calibration is $\widehat{\underline{\beta}}_{\mathbf{P}R} = (\widehat{\mathbf{X}}^\mathsf{T}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^\mathsf{T}\underline{Y}$. By comparing the structure of $\widehat{\underline{\beta}}_{\mathbf{P}}$ and $\widehat{\underline{\beta}}_{\mathbf{P}R}$, $n\mathbf{D}_{\widehat{\underline{\pi}}}$ is a diagonal matrix but $\widehat{\mathbf{X}}^\mathsf{T}\widehat{\mathbf{X}}$ is not necessarily diagonal. We should note again that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is a diagonal matrix due to the nature of categorical data. Also $E\widehat{\underline{\beta}}_{\mathbf{P}} = \mathbf{D}_{\underline{\pi}}^{-1}\mathbf{P}^{-1}\mathbf{D}_{\underline{\lambda}}\mathbf{Q}^\mathsf{T}\underline{\beta} = (\mathbf{Q}^{-1})^\mathsf{T}\mathbf{Q}^\mathsf{T}\underline{\beta} = \underline{\beta}$, using $E(\mathbf{W}^\mathsf{T}\underline{Y}) = n\mathbf{D}_{\underline{\lambda}}\mathbf{Q}^\mathsf{T}\underline{\beta}$ from the remark after Theorem 4.1.1, so $\widehat{\underline{\beta}}_{\mathbf{P}}$ is an unbiased estimator for $\widehat{\underline{\beta}}$, even for small sample size.

$E\widehat{\underline{\beta}}_{\mathbf{P}R} = \left\{n(\mathbf{P}^{-1})\mathbf{D}_{\underline{\lambda}}(\mathbf{P}^{-1})^\mathsf{T}\right\}^{-1}\mathbf{P}^{-1}n\mathbf{D}_{\underline{\lambda}}\mathbf{Q}^\mathsf{T}\underline{\beta} = \mathbf{P}^\mathsf{T}\mathbf{Q}\underline{\beta}$ is not an unbiased estimator for $\underline{\beta}$ unless $\mathbf{P} = \mathbf{Q}^{-1}$.

Remark In the above we assume normality, but it is not really necessary. Using the least squares approach, we get the estimating equation

$$\mathbf{W}^\mathsf{T}(\underline{Y} - \mathbf{W}\underline{\beta}) = \underline{0}.$$

$E[\mathbf{W}^\mathsf{T}(\underline{Y} - \mathbf{W}\underline{\beta})|\mathbf{X}, \underline{Y}] = \mathbf{P}\mathbf{X}^\mathsf{T}\underline{Y} - nD_{\widehat{\underline{\lambda}}}\underline{\beta}$ assuming the misclassification in $\underline{W}$ is non-differential with y. Then $E[\mathbf{W}^\mathsf{T}(\underline{Y} - \mathbf{W}\underline{\beta})] = n\mathbf{P}\mathbf{D}_{\underline{\pi}}\underline{\beta} - nD_{\widehat{\underline{\lambda}}}\underline{\beta}$. We can use the modified estimation equation approach (Buonaccorsi,1996), and solve for $\underline{\beta}$ in

$$\mathbf{W}^\mathsf{T}(\underline{Y} - \mathbf{W}\underline{\beta}) - (n\mathbf{P}\mathbf{D}_{\underline{\pi}}\underline{\beta} - nD_{\underline{\lambda}}\underline{\beta}) = \underline{0}.$$

This gives $\widehat{\underline{\beta}} = \mathbf{D}_{\widehat{\underline{\pi}}}^{-1}\mathbf{P}^{-1}\mathbf{W}^\mathsf{T}\underline{Y}/n$. So we do not really need the normality assumption if we just want to get an estimator for the coefficient.

4.3.1 Inference for $\underline{\beta}$ When $\mathbf{P}$ is Estimated

Sometimes the misclassification model $\mathbf{P}$ is not known, and it is often estimated from some validation data. In this section, we will assume $\mathbf{P}$ is estimated from external validation data, and we will derive sampling covariance of $\widehat{\underline{\beta}}_{\widehat{\mathbf{P}}}$ that includes variability from the validation data.

When $\mathbf{P}$ is estimated from external data, $(\widehat{\mathbf{P}})^{-1}$ is not an unbiased estimator for $\mathbf{P}^{-1}$, and $\ell_{\widehat{\mathbf{P}}}(\underline{\beta}, \underline{Y}, \mathbf{W})$ from previous section is not a corrected log likelihood function. Similarly, $S_{\widehat{\mathbf{P}}}(\underline{\beta}) = \frac{\partial \ell_{\widehat{\mathbf{P}}}}{\partial \underline{\beta}}$ is not a corrected score function, and any corrected function involving $\widehat{\mathbf{P}}^{-1}$ is not a corrected score function. In section 2.3.1 we prove that an unbiased estimator for $\mathbf{P}^{-1}$ does not exist.

Let $(\underline{Y}_n, \mathbf{W}_n, \widehat{\mathbf{P}}_{N_n})$ be a sequence of data where $(\underline{Y}_n, \mathbf{W}_n)$ is the observed data with sample size n, and $\widehat{\mathbf{P}}_{N_n}$ is an unbiased estimator for $\mathbf{P}$ with N_n the minimum of validation size over all categories. Then we have $\widehat{\mathbf{P}}_{N_n}^{-1}$ as a sequence of consistent estimators for $\mathbf{P}^{-1}$ as $N_n \to \infty$. We know $\underline{\widehat{\beta}}_{\mathbf{P}} \xrightarrow{p} \underline{\beta}$ when the main study size n goes to infinity. Then we have $\underline{\widehat{\beta}}_{\widehat{\mathbf{P}}_{N_n}} \xrightarrow{p} \underline{\beta}$ as $N_n, n \to \infty$ according to the generalized Slutsky theorem (Demidenko, 2004).

We should use the sandwich method or pseudo likelihood approach to get the variance for $\underline{\widehat{\beta}}_{\widehat{\mathbf{P}}}$. We know the misclassification model

$$\mathbf{P} = \begin{pmatrix} \theta_{00} & \theta_{01} & \dots & \theta_{0(K-1)} \\ \theta_{10} & \theta_{11} & \dots & \theta_{1(K-1)} \\ \dots & \dots & \dots & \dots \\ 1-\sum_{i=0}^{K-2}\theta_{i0} & 1-\sum_{i=0}^{K-2}\theta_{i1} & \dots & 1-\sum_{i=0}^{K-2}\theta_{i(K-1)} \end{pmatrix}$$

is a function of

$$(\theta_{00}, \dots, \theta_{(K-2)0}, \theta_{01}, \dots, \theta_{(K-2)1}, \dots, \theta_{0(K-1)}, \dots, \theta_{(K-2)(K-1)})^{\mathsf{T}} = \underline{\theta}.$$

The corrected observed information matrix for $(\underline{\beta}^{\mathsf{T}}, \underline{\mathbf{p}})$ is $\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$ where
$H_{11} = -\dfrac{\partial^2 \ell^*}{\partial \underline{\beta} \partial \underline{\beta}^{\mathsf{T}}}, H_{12} = H_{21}^{\mathsf{T}} = \dfrac{\partial^2 \ell^*}{\partial \underline{\beta} \partial \underline{\theta}^{\mathsf{T}}}, H_{22} = \dfrac{\partial^2 \ell^*}{\partial \underline{\theta} \partial \underline{\theta}^{\mathsf{T}}}.$
Also the covariance estimate for $\underline{\theta}$ is

$$\Sigma_{\underline{\theta}} = \begin{pmatrix} \Sigma_0 & \mathbf{0} & \dots & \mathbf{0} \\ \underline{0} & \Sigma_2 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_{K-1} \end{pmatrix}, \text{where}$$

$$\Sigma_k = \frac{1}{N_{.k}} \begin{pmatrix} \widehat{\theta}_{0k}(1-\widehat{\theta}_{0k}) & -\widehat{\theta}_{0k}\widehat{\theta}_{1k} & \dots & -\widehat{\theta}_{0k}\widehat{\theta}_{(K-2)k} \\ -\widehat{\theta}_{1k}\widehat{\theta}_{0k} & \widehat{\theta}_{1k}(1-\widehat{\theta}_{1k}) & \dots & -\widehat{\theta}_{1k}\widehat{\theta}_{(K-2)k} \\ \dots & \dots & \dots & \dots \\ -\widehat{\theta}_{(K-2)k}\widehat{\theta}_{0k} & -\widehat{\theta}_{(K-2)k}\widehat{\theta}_{1k} & \dots & \widehat{\theta}_{(K-2)k}(1-\widehat{\theta}_{(K-2)k}) \end{pmatrix}$$

Assume Σ_K is the asymptotic covariance matrix of $\underline{\beta}_{\mathbf{P}}$ if $\mathbf{P}$ is known using $\widehat{\Sigma}_{1(\widehat{\underline{\beta}}_{\mathbf{P}})}$ or $\widehat{\Sigma}_{2(\widehat{\underline{\beta}}_{\mathbf{P}})}$. Then, from Parke (1986), we can estimate the covariance of $\widehat{\beta}_{\widehat{\mathbf{P}}}$ by:

$$\Sigma_{(\widehat{\underline{\beta}}_{\widehat{\mathbf{P}}})} = \Sigma_K + H_{11}^{-1} H_{12} \Sigma_{\underline{\theta}} H_{12}^{\mathsf{T}} (H_{11}^{-1})^{\mathsf{T}}, \text{ where } H_{ij} \text{ are}$$

$$\begin{aligned} H_{11} &= \widehat{\sigma}_{\widehat{\mathbf{P}}}^{-2} n \mathbf{D}_{\widehat{\underline{\pi}}}, \text{and} \\ H_{12} &= (\underline{h}_{00}, \underline{h}_{10}, \dots, \underline{h}_{(K-2)0}, \underline{h}_{01}, \dots, \underline{h}_{(K-2)1}, \dots, \underline{h}_{0(K-1)}, \dots, \underline{h}_{(K-2)(K-1)K}) \text{ where} \\ \underline{h}_{ml} &= \widehat{\sigma}_{\widehat{\mathbf{P}}}^{-2} \left[\widehat{\mathbf{P}}^{-1} \mathbf{M}_{ml} \widehat{\mathbf{P}}^{-1} \mathbf{W}^{\mathsf{T}} \underline{Y} - n \mathbf{D}_{\widehat{\underline{\pi}}^*} \widehat{\underline{\beta}}_{\widehat{\mathbf{P}}} \right] \text{ with} \\ \widehat{\underline{\pi}}^* &= \widehat{\mathbf{P}}^{-1} \mathbf{M}_{ml} \widehat{\underline{\pi}}. \end{aligned}$$

4.4 Reclassification Model

In this section, we will focus on a reclassification model, where the relationship between $\underline{X}$ and $\underline{W}$ is specified by $\gamma_{ij} = P(\underline{X} = i | \underline{W} = j)$. From Section 4.1, we know $E\widehat{\underline{\beta}}_W = \mathbf{Q}^{\mathsf{T}} \underline{\beta}$ where $\underline{\beta}_W$ is the least squares naive estimator and $\mathbf{Q}$ is the reclassification model. If $\mathbf{Q}$ is known, $(\mathbf{Q}^{\mathsf{T}})^{-1} \widehat{\underline{\beta}}_W$ is a method of moments correction estimator . If $\mathbf{Q}$ is estimated from external data and $\widehat{\mathbf{Q}}$ is an unbiased estimator for $\mathbf{Q}$, $(\widehat{\mathbf{Q}}^{\mathsf{T}})^{-1} \widehat{\underline{\beta}}_W$ is not an unbiased estimator for $\underline{\beta}$.

In Section 4.4.1, we will make a connection between regression with misclassified covariates and mixture models. In section 4.4.2, we will use a regression calibration approach to get an unbiased corrected estimator for the regression coefficients.

4.4.1 Likelihood Function Approach: Mixture

We will use the likelihood approach to find a corrected estimator that is appropriate when a reclassification model is available. We reindex the observed data $y_i, \underline{W}_i$ such that $y_{k(1)}, \ldots, y_{k(n_k)}$ with $\underline{W}_{k(j)} = k$ for $k = 0 \ldots K-1, j = 1 \ldots n_k$.

We assume $\mathbf{Q}$ is known. If we assume non-differential measurement error, then the conditional density function for y given $\underline{W} = j$ is

$$\begin{aligned} f(y|\underline{W} = j) &= \sum_{k=0}^{K-1} f(y|\underline{W} = j, \underline{X} = k) P(\underline{X} = k|\underline{W} = j) \\ &= \sum_{k=0}^{K-1} f(y|\underline{X} = k) P(\underline{X} = k|\underline{W} = j) \\ &= \sum_{k=0}^{K-1} (\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} \exp(-\frac{(y-\beta_k)^2}{2\sigma^2})\gamma_{kj}. \end{aligned}$$

The log likelihood function of $\underline{\beta}$ given $\underline{Y}, \mathbf{W}$ is

$$\ell(\underline{\beta}|\underline{Y}, \mathbf{W}, \mathbf{Q}) = \sum_{m=0}^{K-1}\sum_{j=1}^{n_m} \log\left[\sum_{k=0}^{K-1} (\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} \exp\left\{-\frac{(y_{m(j)}-\beta_k)^2}{2\sigma^2}\right\}\gamma_{km}\right].$$

Note that the data $\mathbf{W}$ appears in the index of $y_{m(j)}$.

This can be shown to be a mixture problem. Let $\underline{X}_{m(j)} = (\, x_{m(j)0} \quad \ldots \quad x_{m(j)(K-1)}\,)^{\mathsf{T}}$ be the unobserved value of $\underline{W}_{m(j)}$ for $m = 0 \ldots K-1, j = 1 \ldots n_m$. Then the log density of $y_{m(j)}$ given $\underline{X}_{m(j)}$ is

$$\sum_{k=0}^{K-1} x_{m(j)k} \left\{\frac{1}{2}\log(\frac{1}{2\pi\sigma^2}) - \frac{(y_{m(j)}-\beta_k)^2}{2\sigma^2}\right\}$$

and the log likelihood for the complete data is

$$\ell(\underline{\beta}|\underline{Y}, \mathbf{X}, \mathbf{W}, \mathbf{Q}) = \sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\sum_{k=0}^{K-1} x_{m(j)k}\left\{\log(\gamma_{km}) + \frac{1}{2}\log(\frac{1}{2\pi\sigma^2}) - \frac{(y_{m(j)}-\beta_k)^2}{2\sigma^2}\right\}.$$

First, we want to find the density of $\underline{X} = l$ given $\underline{W} = m, y$:

$$f(\underline{X} = l|\underline{W} = m, y) = \frac{f(\underline{X} = l, \underline{W} = m, y)}{\sum_{k=0}^{K-1} f(\underline{X} = k, \underline{W} = m, y)}$$

$$= \frac{\dfrac{f(\underline{X}=l,\underline{W}=m,y)f(\underline{X}=l,\underline{W}=m)}{f(\underline{X}=l,\underline{W}=m)}}{\displaystyle\sum_{k=0}^{K-1}\frac{f(\underline{X}=k,\underline{W}=m,y)f(\underline{X}=k,\underline{W}=m)}{f(\underline{X}=k,\underline{W}=m)}}$$

$$= \frac{f(y|\underline{X}=l)f(\underline{X}=l,\underline{W}=m)}{\displaystyle\sum_{k=0}^{K-1} f(y|\underline{X}=k)f(\underline{X}=k,\underline{W}=m)}$$

$$= \frac{f(y|\underline{X}=l)\gamma_{lm}}{\displaystyle\sum_{k=0}^{K-1} f(y|\underline{X}=k)\gamma_{km}}.$$

The EM algorithm can be applied to fit mixture models (McLachlan and Basford, 1988). Let

$$\tau_{m(j)k} = P(x_{m(j)k}=1|\underline{W}=m, Y_{m(j)}) = E(x_{m(j)k}|\underline{W}=m, y_{m(j)}) = \frac{f(y_{m(j)}|\underline{X}_{m(j)}=k)\gamma_{km}}{\displaystyle\sum_{l=0}^{K-1} f(y_{m(j)}|\underline{X}=l)\gamma_{lm}}.$$

We should note that $\tau_{m(j)}$ is not necessary equal to $\gamma_{km} = P(\underline{X}=k|\underline{W}=m)$. For the EM algorithm, we need to have initial value of $\tau^{(0)}_{m(j)k}$ and after pth iteration, we have $(\tau^{(p)}_{m(j)k}, \beta^{(p)}_k)$ for $m = 0 \dots K-1, j = 1 \dots n_m, k = 0 \dots K-1$. For the $(p+1)$th iteration, the E-step is

$$E[\ell(\underline{\beta}|\underline{Y}, \mathbf{X}, \mathbf{W}, \mathbf{Q})|\mathbf{Y}, \mathbf{W}] = \sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\sum_{k=0}^{K-1}\tau^{(p)}_{m(j)k}\left\{[\log(\gamma_{km}) + \frac{1}{2}\log(\frac{1}{2\pi\sigma^2}) - \frac{(y_{m(j)}-\beta_k)^2}{2\sigma^2}\right\}$$

and in the M-step, we will solve β_k in

$$\frac{\partial E[\ell(\underline{\beta}|\underline{Y}, \mathbf{X}, \mathbf{W}, \mathbf{Q})|\mathbf{Y}, \mathbf{W}]}{\partial\beta_k} = \sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\tau^{(p)}_{m(j)k}\frac{(y_{m(j)}-\beta_k)}{\sigma^2} = 0.$$

Solving it, we have

$$\beta^{(p+1)}_k = \frac{\displaystyle\sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\tau^{(p)}_{m(j)k}y_{mj}}{\displaystyle\sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\tau^{(p)}_{m(j)k}},$$

$$\sigma^{2(p+1)} = \frac{\displaystyle\sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\sum_{k=0}^{K-1}\tau^{(p)}_{m(j)k}(y_{m(j)}-\beta^{(p+1)}_k)^2}{\displaystyle\sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\sum_{k=0}^{K-1}\tau^{(p)}_{m(j)k}},$$

and

$$\tau_{m(j)k}^{(p+1)} = \frac{f(y_{m(j)}, \beta_k^{(p+1)})\gamma_{km}}{\sum_{l=0}^{K-1} f(y_{m(j)}, \beta_l^{(p+1)})\gamma_{lm}} \text{ with}$$

$$f(y_{m(j)}, \beta_l^{(p+1)}, \sigma^{2(p+1)}) = \frac{1}{2\pi\sigma^{2(p+1)}} exp\left\{ -\frac{(y_{m(j)} - \beta_l^{(p+1)})^2}{2\sigma^{2(p+1)}} \right\}.$$

Let $\underline{\beta}_{EM}^{(p)} = (\,\beta_0^{(p)} \quad \dots \quad \beta_{K-1}^{(p)}\,)^{\mathsf{T}}$. The EM algorithm ensures the log likelihood values increase monotonically: $\ell(\underline{\beta}_{EM}^{(p+1)}|\underline{Y}, \mathbf{W}, \mathbf{Q}) \geq \ell(\underline{\beta}_{EM}^{(p)}|\underline{Y}, \mathbf{W}, \mathbf{Q})$. The convergence of $\underline{\beta}_{EM}^{(p)}$is consistent for the MLE of $\underline{\beta}$ (McLachlan and Peel, 2000), and we will denote it by $\underline{\beta}_{EM}$.

The observed information matrix is:

$$\mathbf{I}(\underline{\beta}_{EM}) = -\frac{\partial^2 \ell(\underline{\beta}|\underline{Y}, \mathbf{W}, \mathbf{Q})}{\partial \underline{\beta} \partial \underline{\beta}^{\mathsf{T}}}|_{\underline{\beta}=\underline{\beta}_{EM}}.$$

We could use the observed information to get an asymptotic covariance estimator for $\widehat{\underline{\beta}}_{EM}$, but the computation is hard. Louis (1982) shows that the observed information can be expressed in terms of complete-data gradeint vector or second derivative matrix. Since we have independent data, the observed information matrix can be approximated in terms of the gradient of the complete-data log likelihood, where the unobservable variables are replaced by their fitted conditional expectations (McLachlan and Peel, 2000):

$$\widehat{\Sigma}_{\underline{\beta}_{EM}} = \widehat{\sigma}_{EM}^2 \left\{ \sum_{m=0}^{K-1} \sum_{j=1}^{n_m} \underline{h}_{m(j)} \underline{h}_{m(j)}^{\mathsf{T}} \right\}^{-1},$$

where $\underline{h}_{m(j)} = (\,\widehat{\tau}_{m(j)0}(y_{m(j)} - \widehat{\beta}_{EM_1}) \quad \dots \quad \widehat{\tau}_{m(j)(K-1)}(y_{m(j)} - \widehat{\beta}_{EM_{K-1}})\,)^{\mathsf{T}}$.

Remark In the above, we assume the reclassification model $\mathbf{Q}$ is known, and we can use the probabilities γ_{km} to get $\tau_{m(j)k}$. It would be interesting to know when the reclassification model $\mathbf{Q}$ is not available, whether or not we can recover γ_{km} from the EM algorithm. We will leave this a topic for future research.

Remark We could use the likelihood approach when a misclassification model $\mathbf{P}$ is available. From the non-differential assumption, we have the density function $f(y, \underline{X}, \underline{W}) = f(y|\underline{X})P(\underline{X}, \underline{W})$. Let $\eta_{mk} = P(\underline{W} = m, \underline{X} = k)$. So the complete-data log likelihood is:

$$\ell(\underline{\beta}|\underline{Y}, \mathbf{X}, \mathbf{W}) = \sum_{m=0}^{K-1}\sum_{j=1}^{n_m}\sum_{k=0}^{K-1} x_{m(j)k}\left\{\log(\eta_{mk}) + \log(f(y_{m(j)}|\underline{X}_{m(j)} = k)\right\}.$$

We could have $\tau_{m(j)k} = P(x_{m(j)k} = 1|\underline{W} = m, y_{m(j)}) = \dfrac{f(y_{m(j)}|\underline{X}_{m(j)} = k)\eta_{mk}}{\sum\limits_{l=1}^{K-1} f(y_{m(j)}|\underline{X} = l)\eta_{ml}}$, and we could estimate η_{ml} by $\widehat{\eta}_{ml} = \theta_{ml}\widehat{\pi}_l$.

4.4.2 Imputed Data and Regression Calibration

In this section, we develop a regression calibration approach to get a corrected estimator when a reclassification model is available. First, we assume the reclassification model $\mathbf{Q}$ is known.

We impute $\widehat{\underline{X}}_i = \mathbf{Q}\underline{W}_i$, and set $\widehat{\mathbf{X}} = \begin{pmatrix} \widehat{\underline{X}}_1^{\mathsf{T}} \\ \vdots \\ \widehat{\underline{X}}_n \end{pmatrix} = \mathbf{W}\mathbf{Q}^{\mathsf{T}}$, the least squares estimator $\widehat{\underline{\beta}}_{R_{\mathbf{Q}}}$ from regressing y_i on $\widehat{\underline{X}}_i$ is

$$\begin{aligned}
\widehat{\underline{\beta}}_{R_{\mathbf{Q}}} &= (\widehat{\mathbf{X}}^{\mathsf{T}}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^{\mathsf{T}}\underline{Y} \\
&= (\mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}})^{-1}\mathbf{Q}\mathbf{W}^{\mathsf{T}}\underline{Y} \\
&= (\mathbf{Q}^{\mathsf{T}})^{-1}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}\underline{Y} \\
&= (\mathbf{Q}^{\mathsf{T}})^{-1}\widehat{\underline{\beta}}_W
\end{aligned}$$

where $\widehat{\underline{\beta}}_W$ is the least squares naive estimator. Now, since

$$\begin{aligned}
E\left\{\widehat{\underline{\beta}}_{R_{\mathbf{Q}}}\right\} &= E\left\{(\mathbf{Q}^{\mathsf{T}})^{-1}(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}\underline{Y}\right\} \\
&= (\mathbf{Q}^{\mathsf{T}})^{-1}(n\mathbf{D}_{\lambda})^{-1}n\mathbf{D}_{\underline{\lambda}}\mathbf{Q}^{\mathsf{T}}\underline{\beta} \\
&= \underline{\beta},
\end{aligned}$$

$\widehat{\underline{\beta}}_{R_{\mathbf{Q}}}$ is an unbiased estimator for $\underline{\beta}$.

The asymptotic covariance matrix of $\underline{\widehat{\beta}}_{R_{\mathbf{Q}}}$ is

$$\widehat{\Sigma}_{\underline{\widehat{\beta}}_{R_{\mathbf{Q}}}} = (\mathbf{Q}^{\mathsf{T}})^{-1}\widehat{\Sigma}_{\underline{\widehat{\beta}}_{W}}(\mathbf{Q})^{-1}$$

where

$$\widehat{\Sigma}_{\underline{\widehat{\beta}}_{W}} = \widehat{\sigma}^2_W(\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}$$

and $\widehat{\sigma}^2_W = (\underline{Y} - \mathbf{W}\underline{\widehat{\beta}}_W)^{\mathsf{T}}(\underline{Y} - \mathbf{W}\underline{\widehat{\beta}}_W)/n$.

We should note that σ^2_W is not an estimator of σ^2. Since

$$\begin{aligned} \text{Var}(y) &= E\{\ \text{Var}(y|\underline{X})\} +\ \text{Var}\{E(y|\underline{X})\} \\ &= \sigma^2 +\ \text{Var}(\underline{\beta}^{\mathsf{T}}\underline{X}) \\ &= \sigma^2 + \underline{\beta}^{\mathsf{T}}\ \text{Var}(\underline{X})\underline{\beta}, \end{aligned}$$

we could get an estimator for σ^2 :

$$\widehat{\sigma}^2 = \widehat{\text{Var}}(y) - \underline{\widehat{\beta}}^{\mathsf{T}}_{R_{\mathbf{Q}}}\ \widehat{\text{Var}}(\underline{X})\underline{\widehat{\beta}}_{R_{\mathbf{Q}}},$$

where $\widehat{\text{Var}}(y) = \dfrac{\sum_{i=0}^{K-1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2}{\sum_{i=0}^{K-1} n_i}, y_{..} = \dfrac{\sum_{i=0}^{K-1}\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=0}^{K-1} n_i}$, and

$$\widehat{\text{Var}}(\underline{X}) = \begin{pmatrix} \widehat{\pi}_0(1-\widehat{\pi}_0) & -\widehat{\pi}_0\widehat{\pi}_1 & \dots & \widehat{\pi}_0\widehat{\pi}_{K-1} \\ -\widehat{\pi}_1\widehat{\pi}_0 & \widehat{\pi}_1(1-\widehat{\pi}_1) & \dots & -\widehat{\pi}_1\widehat{\pi}_{K-1} \\ \dots & \dots & \ddots & \dots \\ -\widehat{\pi}_{K-1}\widehat{\pi}_0 & -\widehat{\pi}_{K-1}\widehat{\pi}_0 & \dots & \widehat{\pi}_{K-1}(1-\widehat{\pi}_{K-1}) \end{pmatrix}.$$

4.4.3 Inferences When Q is Estimated

In this section, we want to discuss the situation when the reclassification model is estimated from external data. We will use the sandwich method to get an asymptotic covariance matrix to account the variability that comes from an estimated **Q**.

We will only develop the asymptotic covariance matrix for the regression calibration estimator. The asymptotic covariance of the EM estimator uses the same technique.

Suppose $\mathbf{Q}$ is estimated from external data and $\widehat{\mathbf{Q}}$ is an unbiased estimator for $\mathbf{Q}$. We know the reclassification model

$$\mathbf{Q} = \begin{pmatrix} \gamma_{00} & \gamma_{01} & \cdots & \gamma_{0(K-1)} \\ \gamma_{10} & \gamma_{11} & \cdots & \gamma_{1(K-1)} \\ \cdots & \cdots & \cdots & \cdots \\ 1-\sum_{k=0}^{K-2}\gamma_{k0} & 1-\sum_{k=0}^{K-2}\gamma_{k1} & \cdots & 1-\sum_{k=0}^{K-2}\gamma_{k(K-1)} \end{pmatrix}$$

is a function of

$$(\gamma_{00}, \ldots, \gamma_{(K-2)0}, \gamma_{01}, \ldots, \gamma_{(K-2)1}, \ldots, \gamma_{0(K-1)}, \ldots, \gamma_{(K-2)(K-1)})^{\mathsf{T}} = \underline{\gamma}.$$

As a result, the regression calibration estimator $\widehat{\underline{\beta}}_{R_{\widehat{\mathbf{Q}}}} = (\widehat{\mathbf{Q}}^{\mathsf{T}})^{-1}\widehat{\underline{\beta}}_W$ is still a consistent estimator for $\underline{\beta}$ if the validation sample size and main study size go to infinity, and the estimating equation is

$$S(\underline{\beta}|\underline{Y}, \mathbf{W}, \mathbf{Q}) = \mathbf{Q}\mathbf{W}^{\mathsf{T}}\underline{Y} - \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}}\underline{\beta}.$$

Let

$$\begin{aligned} H_{11} &= \frac{\partial S(\underline{\beta}|\underline{Y}, \mathbf{W}, \mathbf{Q})}{\partial \underline{\beta}} = -\mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}}, \\ H_{12} &= \left.\frac{\partial S(\underline{\beta}|\underline{Y}, \mathbf{W}, \mathbf{Q})}{\partial \underline{\gamma}}\right|_{\underline{\beta}=\widehat{\underline{\beta}}_{R_{\widehat{\mathbf{Q}}}}} \\ &= (\underline{h}_{00}, \underline{h}_{10}, \ldots, \underline{h}_{(K-2)0}, \underline{h}_{01}, \ldots, \underline{h}_{(K-2)1}, \ldots, \underline{h}_{0(K-1)}, \ldots, \underline{h}_{(K-2)(K-1)}), \end{aligned}$$

with $\underline{h}_{ml} = \mathbf{M}_{ml}\mathbf{W}^{\mathsf{T}}\underline{Y} - \mathbf{M}_{ml}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}}\underline{\beta} - \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{M}_{ml}^{\mathsf{T}}\underline{\beta}$.

Then the asymptotic covariance matrix of $\widehat{\underline{\beta}}_{R_{\widehat{\mathbf{Q}}}}$ is

$$\widehat{\Sigma}_{\widehat{\underline{\beta}}_{R_{\widehat{\mathbf{Q}}}}} = \widehat{\Sigma}_K + H_{11}^{-1}H_{12}\Sigma_{\underline{\gamma}}H_{12}^{\mathsf{T}}(H_{11}^{-1})^{\mathsf{T}}, \text{ where}$$

$\widehat{\Sigma}_K$ is the asymptotic covariance matrix of $\widehat{\underline{\beta}}_{R_{\widehat{\mathbf{Q}}}}$, treating $\widehat{\mathbf{Q}}$ as known, and

$$\Sigma_{\underline{\gamma}} = \Sigma_{\widehat{\underline{\gamma}}} = \begin{pmatrix} \Sigma_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_1 & \cdots \mathbf{0} & \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{K-1} \end{pmatrix} \text{ with}$$

$$\Sigma_k = \frac{1}{N_{.k}} \begin{pmatrix} \widehat{\gamma}_{0k}(1-\widehat{\gamma}_{0k}) & -\widehat{\gamma}_{0k}\widehat{\gamma}_{1k} & \dots & -\widehat{\gamma}_{0k}\widehat{\gamma}_{(K-2)k} \\ -\widehat{\gamma}_{1k}\widehat{\gamma}_{0k} & \widehat{\gamma}_{1k}(1-\widehat{\gamma}_{1k}) & \dots & -\widehat{\gamma}_{1k}\widehat{\gamma}_{(K-2)k} \\ \dots & \dots & \dots & \dots \\ -\widehat{\gamma}_{(K-2)k}\widehat{\gamma}_{0k} & -\widehat{\gamma}_{(K-2)k}\widehat{\gamma}_{1k} & \dots & \widehat{\gamma}_{(K-2)k}(1-\widehat{\gamma}_{(K-2)k}) \end{pmatrix} \text{ where}$$

$N_{.k}$ is the validation size for category k.

4.5 Linear Regression with Categorical Covariates and Perfectly Measured Covariates

In this section, we will study regression with misclassified covariates and perfectly measured covariates. In section 4.5.1, we will study the bias of the naive least squares estimator. In 4.5.2, we use the score function approach to create a consistent estimator, and in 4.5.3 we will demonstrate how to use the reclassification model to get a corrected estimator.

4.5.1 Bias of Naive Estimator

Assume $Y = \underline{X}^{\mathsf{T}}\underline{\beta} + \underline{Z}^{\mathsf{T}}\underline{\beta}_Z + \epsilon$ where $\underline{X}$ is a categorical variable with K categories, $\underline{Z}$ is a vector of variables with no measurement error, and ϵ is random error with mean 0, and independent from $\underline{X}$, and $\underline{Z}$.

Let $y_i, \underline{W}_i, \underline{Z}_i, i = 1\dots n$ where the dimension of $\underline{W}_i$ is $k \times 1$ (These are the observed values for X_i), and the dimension of $\underline{Z}_i$ is $p \times 1$. Let

$$\mathbf{W} = \begin{pmatrix} \underline{W}_1^{\mathsf{T}} \\ \underline{W}_2^{\mathsf{T}} \\ \vdots \\ \underline{W}_n^{\mathsf{T}} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \underline{X}_1^{\mathsf{T}} \\ \underline{X}_2^{\mathsf{T}} \\ \vdots \\ \underline{X}_n^{\mathsf{T}} \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} \underline{Z}_1^{\mathsf{T}} \\ \underline{Z}_2^{\mathsf{T}} \\ \vdots \\ \underline{Z}_n^{\mathsf{T}} \end{pmatrix}, \underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

It is easy to show that for linear regression models with mismeasured continuous covariates and perfectly measured covariates, the bias of the naive least squares estimator $\begin{pmatrix} \underline{\widehat{\beta}}_W \\ \underline{\widehat{\beta}}_{Z_W} \end{pmatrix}$ for $\begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}$ is a function of $\underline{\beta}$, and if $\underline{X}$ and $\underline{Z}$ are uncorrelated, the naive least squares estimator $\widehat{\underline{\beta}}_{Z_W}$ is also an unbiased estimator for $\underline{\beta}_Z$. Gustafson (2004) proves the above statement is true for binary misclassified covariates and a perfectly measured covariate with a nondifferential misclassification model. Christopher and Kupper (1995) prove that if the reclassification model is nondifferential (or independent from $\underline{Z}$), then $\widehat{\underline{\beta}}_{Z_W}$ is an unbiased estimator for $\underline{\beta}_Z$. Buonaccorsi et al (2005) demonstrate that if a misclassification model is independent of the perfectly measured $\underline{Z}$, and $\underline{X}$ is independent of $\underline{Z}$, then there will be no bias in the naive estimator of the coefficients associated with the perfectly measured covariates. The following lemma restates the result of Christopher and Kupper (1995) for misclassified data, with a different proof.

Lemma 4.5.1 *Assume $Y = \underline{X}^{\mathsf{T}}\underline{\beta} + \underline{Z}^{\mathsf{T}}\underline{\beta}_Z + \epsilon$ where $\underline{X}$ is a categorical variable with K categories, $\underline{Z}$ is a vector of variables with no measurement error, and ϵ is random error with mean 0, and independent from $\mathbf{X}$ and $\mathbf{Z}$. Suppose we observe $y_i, \underline{W}_i, \underline{Z}_i, i = 1 \ldots n$ where $\underline{W}_i$ is the observed value of $\underline{X}_i$. Then the bias of the naive least squares estimator $\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix}$ for $\begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}$ is a function of $\underline{\beta}$, not a function of $\underline{\beta}_Z$. If $E[\underline{X}|\underline{W}, \underline{Z}] = E[\underline{X}|\underline{W}]$, i.e. the reclassification model is nondifferential , then the naive least squares estimator $\widehat{\underline{\beta}}_Z$ is an unbiased estimator for $\underline{\beta}_Z$.*

Proof The least squares naive estimator is $\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} = \left\{ \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{W} & \mathbf{Z} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \underline{Y}$, and

$$
\begin{aligned}
& E\left\{ \begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} \middle| \mathbf{W}, \mathbf{X}, \mathbf{Z} \right\} \\
= \quad & E\left[\left\{ \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{W} & \mathbf{Z} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \underline{Y} \middle| \mathbf{W}, \mathbf{X}, \mathbf{Z} \right] \\
= \quad & \left\{ \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{W} & \mathbf{Z} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{X} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix} \\
= \quad & \left\{ \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{W} & \mathbf{Z} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix} \left[\begin{pmatrix} \mathbf{W} & \mathbf{Z} \end{pmatrix} + \begin{pmatrix} \mathbf{X} - \mathbf{W} & \mathbf{0} \end{pmatrix} \right] \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix} \\
= \quad & \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix} + \begin{pmatrix} \mathbf{W}^{\mathsf{T}}\mathbf{W} & \mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{W}^{\mathsf{T}}(\mathbf{X} - \mathbf{W})\underline{\beta} \\ \mathbf{Z}^{\mathsf{T}}(\mathbf{X} - \mathbf{W})\underline{\beta} \end{pmatrix}.
\end{aligned}
$$

So the bias of naive estimator is a function of $\underline{\beta}$ only, not $\underline{\beta}_Z$.

If $E[\underline{X}|\underline{W},\underline{Z}] = E[\underline{X}|\underline{W}] = \mathbf{Q}\underline{W}$ where $\mathbf{Q}$ is the reclassification model, then

$$E\left\{\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} |\mathbf{W},\mathbf{Z}\right\} \;=\; \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix} + \begin{pmatrix} \mathbf{W}^\mathsf{T}\mathbf{W} & \mathbf{W}^\mathsf{T}\mathbf{Z} \\ \mathbf{Z}^\mathsf{T}\mathbf{W} & \mathbf{Z}^\mathsf{T}\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{W}^\mathsf{T}\mathbf{W}(\mathbf{Q}^\mathsf{T} - \mathbf{I})\underline{\beta} \\ \mathbf{Z}^\mathsf{T}\mathbf{W}(\mathbf{Q}^\mathsf{T} - \mathbf{I})\underline{\beta} \end{pmatrix}.$$

Using the fact:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix},$$

we have $E\left\{\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} |\mathbf{W},\mathbf{Z}\right\} = \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix} + \begin{pmatrix} (\mathbf{Q}^\mathsf{T} - \mathbf{I})\underline{\beta} \\ \mathbf{0} \end{pmatrix}$. So,

$$E\left\{\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} |\mathbf{W},\mathbf{Z}\right\} = \begin{pmatrix} \mathbf{Q}\underline{\beta} \\ \underline{\beta}_Z \end{pmatrix},$$

and $\widehat{\underline{\beta}}_{Z_W}$ is an unbiased for $\underline{\beta}_Z$ if the reclassification is nondifferential.

4.5.2 Score Function Approach

In this section, we will assume a misclassification model is available, and we will use a score function approach like in section 4.3 to get a corrected estimator. Since $P(\underline{W}|\underline{X})$ might depend on the value of $\underline{Z}$, we need to have a misclassification model for each value of $\underline{Z}$ in the range of the perfectly measured variables.

We define

$$\theta_{ij\underline{Z}} = P(\underline{W} = j|\underline{X} = j, \underline{Z}).$$

Let

$$\mathbf{P}_{\underline{Z}} = \begin{pmatrix} \theta_{00\underline{Z}} & \theta_{01\underline{Z}} & \cdots & \theta_{0(K-1)\underline{Z}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{(K-1)0\underline{Z}} & \theta_{(K-1)1\underline{Z}} & \cdots & \theta_{(K-1)(K-1)\underline{Z}} \end{pmatrix} = (\underline{\theta}_{0\underline{Z}}, \underline{\theta}_{1\underline{Z}}, \ldots, \underline{\theta}_{(K-1)\underline{Z}})$$

be a misclassification model for each $\underline{Z}$ in the range of the perfectly measured covariates. We will have $\mathbf{P}_{\underline{Z}}^{-1}\underline{\theta}_{m\underline{Z}} = \underline{e}_m$.

$Y_i = \underline{X}_i^\mathsf{T}\underline{\beta} + \underline{Z}_i^\mathsf{T}\underline{\beta}_Z + \epsilon$ where $\epsilon \sim N(0,\sigma^2), i = 1 \ldots n$, and σ^2 is unknown. From the observed data $y_i, \underline{W}_i, \underline{Z}_i$, the log-likelihood function is

$$\ell(\underline{\beta}, \underline{\beta}_Z, \underline{Y}, \mathbf{W}, \mathbf{Z}) = -\frac{n}{2}\log(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}(y_i - \underline{W}_i^\mathsf{T}\underline{\beta} - \underline{Z}_i^\mathsf{T}\underline{\beta}_Z)^2. \text{ Let}$$

$$\begin{aligned}\ell^*(\underline{\beta},\underline{\beta}_Z,\underline{Y},\mathbf{W},\mathbf{Z}) &= -\frac{n}{2}\log(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}\Big[y_i^2-2y_i(\underline{\beta}^\top\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+\underline{Z}_i^\top\underline{\beta}_Z)\\ &\quad +2\underline{Z}_i^\top\beta_Z\underline{\beta}^\top\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+\underline{\beta}^\top\sum_{k=0}^{K-1}\Big\{(\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i)^\top\underline{e}_k\Big\}\,\underline{e}_k\underline{e}_k^\top\underline{\beta}+(\underline{Z}_i^\top\underline{\beta}_Z)^2\Big].\end{aligned}$$

As in section 4.3,

$$\begin{aligned}&E[\ell^*(\underline{\beta},\sigma^2,\underline{Y},\mathbf{W},\mathbf{Z})|\underline{Y},\mathbf{X},\mathbf{Z}]\\
=\;& -\frac{n}{2}\log(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}E\Big[y_i^2-2y_i(\underline{\beta}^\top\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+\underline{Z}_i^\top\underline{\beta}_Z)+\\
&2\underline{Z}_i^\top\beta_Z\underline{\beta}^\top\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+\underline{\beta}^\top\sum_{k=0}^{K-1}\Big\{(\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i)^\top\underline{e}_k\Big\}\,\underline{e}_k\underline{e}_k^\top\underline{\beta}+(\underline{Z}_i^\top\underline{\beta}_Z)^2\Big|y_i,\underline{X}_i,\underline{Z}_i\Big]\\
=\;& -\frac{n}{2}\log(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}\sum_{m=0}^{K-1}E\Big[y_i^2-2y_i(\underline{\beta}^\top\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+\underline{Z}_i^\top\underline{\beta}_Z)+\\
&2\underline{Z}_i^\top\beta_Z\underline{\beta}^\top\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+\underline{\beta}^\top\sum_{k=0}^{K-1}\Big\{(\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i)^\top\underline{e}_k\Big\}\,\underline{e}_k\underline{e}_k^\top\underline{\beta}+(\underline{Z}_i^\top\underline{\beta}_Z)^2\Big|y_i,\underline{X}_i=m,\underline{Z}_i\Big]\mathbf{1}_{\underline{\mathbf{X}}_\mathbf{i}=\mathbf{m}}\\
=\;& -\frac{n}{2}\log(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}\sum_{m=0}^{K-1}\Big[y_i^2-2y_i(\underline{\beta}^\top\underline{e}_m+\underline{Z}_i^\top\underline{\beta}_Z)+\\
&2\underline{Z}_i^\top\beta_Z\underline{\beta}^\top\underline{e}_m+\underline{\beta}^\top\sum_{k=0}^{K-1}\Big\{\underline{e}_m^\top\underline{e}_k\Big\}\,\underline{e}_k\underline{e}_k^\top\underline{\beta}+(\underline{Z}_i^\top\underline{\beta}_Z)^2\Big]\mathbf{1}_{\underline{\mathbf{X}}_\mathbf{i}=\mathbf{m}}\\
=\;& -\frac{n}{2}\log(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}\sum_{m=0}^{K-1}\Big[y_i^2-2y_i(\beta_m+\underline{Z}_i^\top\underline{\beta}_Z)+2\underline{Z}_i^\top\beta_Z\underline{\beta}_m+\beta_m^2+(\underline{Z}_i^\top\underline{\beta}_Z)^2\Big]\mathbf{1}_{\underline{\mathbf{X}}_\mathbf{i}=\mathbf{m}}\\
=\;& -\frac{n}{2}\log(2\pi\sigma^2)-(2\sigma^2)^{-1}\sum_{i=1}^{n}(y_i-\underline{X}_i^\top\underline{\beta}-\underline{Z}_i^\top\underline{\beta}_Z)^2\\
=\;& \ell(\underline{\beta},\underline{\beta}_Z,\underline{Y},\mathbf{X},\mathbf{Z}).\end{aligned}$$

So ℓ^* is a corrected log likelihood function and $\begin{pmatrix}\frac{\partial\ell^*}{\partial\underline{\beta}}\\ \frac{\partial\ell^*}{\partial\underline{\beta}_Z}\end{pmatrix}$ is a corrected score function for $(\,\underline{\beta}^\top\quad\underline{\beta}_Z^\top\,)^\top$. If we solve

$$\begin{aligned}&\begin{pmatrix}\frac{\partial\ell^*}{\partial\underline{\beta}}\\ \frac{\partial\ell^*}{\partial\underline{\beta}_Z}\end{pmatrix}\\
=\;& -(2\sigma^2)^{-1}\begin{pmatrix}\sum_{i=1}^{n}\Big[-2y_i\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i+2\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i\underline{Z}_i^\top\underline{\beta}_Z+2\sum_{k=0}^{K-1}\Big\{(\mathbf{P}_{\underline{Z}_i}^{-1}\underline{W}_i)^\top\underline{e}_k\Big\}\,\underline{e}_k\underline{e}_k^\top\underline{\beta}\Big]\\ \sum_{i=1}^{n}\Big[-2y_i\underline{Z}_i+2\underline{Z}_i\underline{W}_i^\top(\mathbf{P}_{\underline{Z}_i}^{-1})^\top\underline{\beta}+2\underline{Z}_i\underline{Z}_i^\top\underline{\beta}_Z\Big]\end{pmatrix}\\
=\;& \underline{0},\end{aligned}$$

i.e. $\begin{pmatrix} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \left\{ (\mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i)^{\mathsf{T}} \underline{e}_k \right\} \underline{e}_k \underline{e}_k^{\mathsf{T}} & \sum_{i=1}^{n} \mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i \underline{Z}_i^{\mathsf{T}} \\ \sum_{i=1}^{n} \underline{Z}_i \underline{W}_i^{\mathsf{T}} (\mathbf{P}_{\underline{Z}_i}^{-1})^{\mathsf{T}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i \\ \mathbf{Z}^{\mathsf{T}} \underline{Y} \end{pmatrix}.$

Then we have

$$= \begin{pmatrix} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \left\{ (\mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i)^{\mathsf{T}} \underline{e}_k \right\} \underline{e}_k \underline{e}_k^{\mathsf{T}} & \sum_{i=1}^{n} \mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i \underline{Z}_i^{\mathsf{T}} \\ \sum_{i=1}^{n} \underline{Z}_i \underline{W}_i^{\mathsf{T}} (\mathbf{P}_{\underline{Z}_i}^{-1})^{\mathsf{T}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i \\ \mathbf{Z}^{\mathsf{T}} \underline{Y} \end{pmatrix}$$

as a consistent estimator for $\begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}$. Without knowing the correlation between $\underline{X}$ and $\underline{Z}$, we cannot get $E\left[\begin{pmatrix} \underline{\widehat{\beta}}_s \\ \underline{\widehat{\beta}}_{Z_s} \end{pmatrix}\right]$ like in Section 4.3. Therefore, we do not know if $\begin{pmatrix} \underline{\widehat{\beta}}_s \\ \underline{\widehat{\beta}}_{Z_s} \end{pmatrix}$ is an unbiased estimator or not for $\begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}$.

The corrected estimator looks complicated, but if we use $\mathbf{P}_{\underline{Z}_i}$ to impute $\underline{\widehat{X}}_i = \mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i$ and $\widehat{\mathbf{X}} = \begin{pmatrix} \underline{\widehat{X}}_1 \\ \vdots \\ \underline{\widehat{X}}_n \end{pmatrix}$, then it can be rewritten as

$$\begin{pmatrix} \underline{\widehat{\beta}}_s \\ \underline{\widehat{\beta}}_{Z_s} \end{pmatrix} = \begin{pmatrix} n\mathbf{D}_{\widehat{\underline{\pi}}} & \widehat{\mathbf{X}}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\widehat{\mathbf{X}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\mathbf{X}}^{\mathsf{T}}\underline{Y} \\ \mathbf{Z}^{\mathsf{T}}\underline{Y} \end{pmatrix}$$

with $\widehat{\underline{\pi}} = \sum_{i=1}^{n} \mathbf{P}_{\underline{Z}_i}^{-1} \underline{W}_i / n$.

Notice $n\mathbf{D}_{\widehat{\underline{\pi}}}$ in the corrected score estimator is an estimate of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$.

If $\mathbf{P}_{\underline{Z}} = \mathbf{P}$ for all values of $\underline{Z}$, the above calculations will be much simpler, and $\underline{W}$, $\underline{Z}$ will be independent given $\underline{X}$. We should note that if $\mathbf{P}_{\underline{Z}} = \mathbf{P}$, then it does not necessarily follow that $\mathbf{Q}_{\underline{Z}} = \mathbf{Q}$.

Differentiating the corrected log likelihood function ℓ^* with respect to σ^2, we will have an estimator for σ^2,

$$\widehat{\sigma}_*^2 = \sum_{i=1}^{n} \left\{ y_i^2 - 2y_i(\underline{\widehat{X}}_i^{\mathsf{T}} \underline{\widehat{\beta}}_s + \underline{Z}_i^{\mathsf{T}} \underline{\widehat{\beta}}_Z) + 2\underline{Z}_i^{\mathsf{T}} \underline{\widehat{\beta}}_{Z_s} \underline{\widehat{X}}_i^{\mathsf{T}} \underline{\widehat{\beta}}_s + \underline{\widehat{\beta}}_s^{\mathsf{T}} \mathbf{D}_{\underline{\widehat{X}}_i} \underline{\widehat{\beta}}_s + (\underline{Z}_i^{\mathsf{T}} \underline{\widehat{\beta}}_{Z_s})^2 \right\} / n.$$

Next, we want to find an estimated variance-covariance matrix for $\begin{pmatrix} \underline{\widehat{\beta}}_s \\ \underline{\widehat{\beta}}_{Z_s} \end{pmatrix}$.

The corrected observed information matrix $\mathbf{I}^*$ is

$$\mathbf{I}^*(\underline{\beta}, \underline{\beta}_Z, \underline{Y}, \mathbf{W}, \mathbf{Z}) = \begin{pmatrix} \sum_{i=1}^{n} \mathbf{D}_{\underline{\widehat{X}}_i} & \sum_{i=1}^{n} \underline{\widehat{X}}_i \underline{Z}_i^{\mathsf{T}} \\ \sum_{i=1}^{n} \underline{Z}_i \underline{\widehat{X}}_i^{\mathsf{T}} & \sum_{i=1}^{n} \underline{Z}_i \underline{Z}_i^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{\underline{\widehat{\pi}}} & \widehat{\mathbf{X}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\widehat{\mathbf{X}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix} / \widehat{\sigma}_*^2,$$

and let

$$\begin{aligned} &\mathbf{V}(\underline{\beta}, \underline{\beta}_Z, \underline{Y}, \mathbf{W}, \mathbf{Z}) \\ = &\sum_{i=1}^{n} \begin{pmatrix} -y_i \underline{\widehat{X}}_i + \underline{\widehat{X}}_i \underline{Z}_i^{\mathsf{T}} \underline{\beta}_Z + \mathbf{D}_{\underline{\widehat{X}}_i} \underline{\beta} \\ -y_i \underline{Z}_i + \underline{Z}_i \underline{\widehat{X}}_i^{\mathsf{T}} \underline{\beta} + \underline{Z}_i \underline{Z}_i^{\mathsf{T}} \underline{\beta}_Z \end{pmatrix} \begin{pmatrix} -y_i \underline{\widehat{X}}_i + \underline{\widehat{X}}_i \underline{Z}_i^{\mathsf{T}} \underline{\beta}_Z + \mathbf{D}_{\underline{\widehat{X}}_i} \underline{\beta} \\ -y_i \underline{Z}_i + \underline{Z}_i \underline{\widehat{X}}_i^{\mathsf{T}} \underline{\beta} + \underline{Z}_i \underline{Z}_i^{\mathsf{T}} \underline{\beta}_Z \end{pmatrix}^{\mathsf{T}} / \widehat{\sigma}_*^4. \end{aligned}$$

As in Section 4.3, we will first find $\omega(\underline{\beta}, \underline{\beta}_Z, y, \underline{W}, \underline{Z})$ such that

$$E[\omega(\underline{\beta}, \underline{\beta}_Z, y, \underline{W}, \underline{Z})|y, \underline{X}, \underline{X}] = S(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z}) S(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z})^{\mathsf{T}}$$

with $S(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z}) = \begin{pmatrix} \frac{\partial \ell(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z})}{\partial \underline{\beta}} \\ \frac{\partial \ell(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z})}{\underline{\beta}_Z} \end{pmatrix}$.

For our case,

$$\begin{aligned} S(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z}) &= (y - \underline{X}^{\mathsf{T}} \underline{\beta} - \underline{Z}^{\mathsf{T}} \underline{\beta}_Z) \begin{pmatrix} \underline{X} \\ \underline{Z} \end{pmatrix} / \sigma^2 \\ S(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z}) S(\underline{\beta}, \underline{\beta}_Z, y, \underline{X}, \underline{Z})^{\mathsf{T}} &= (y - \underline{X}^{\mathsf{T}} \underline{\beta} - \underline{Z}^{\mathsf{T}} \underline{\beta}_Z)^2 \begin{pmatrix} \underline{X}\underline{X}^{\mathsf{T}} & \underline{X}\underline{Z}^{\mathsf{T}} \\ \underline{Z}\underline{X}^{\mathsf{T}} & \underline{Z}\underline{Z}^{\mathsf{T}} \end{pmatrix} / \sigma^4. \end{aligned}$$

Let

$$\begin{aligned} \omega((\underline{\beta}, \underline{\beta}_Z, y, \underline{W}, \underline{Z}, \mathbf{P}_{\underline{Z}}) = &\sum_{k=0}^{K-1} \Big\{ y^2 - 2y(\underline{\beta}^{\mathsf{T}} \underline{e}_k + \underline{Z}^{\mathsf{T}} \underline{\beta}_Z) + 2\underline{\beta}^{\mathsf{T}} \underline{e}_k \underline{Z}^{\mathsf{T}} \underline{\beta}_Z + (\underline{\beta}^{\mathsf{T}} \underline{e}_k)^2 \\ &+ (\underline{Z}^{\mathsf{T}} \underline{\beta}_Z)^2 \Big\} \begin{pmatrix} \left\{ (\mathbf{P}_{\underline{Z}}^{-1} \underline{W})^{\mathsf{T}} \underline{e}_k \right\} \underline{e}_k \underline{e}_k^{\mathsf{T}} & \mathbf{P}_{\underline{Z}}^{-1} \underline{W}\underline{Z}^{\mathsf{T}} \\ \underline{Z}(\mathbf{P}_{\underline{Z}}^{-1} \underline{W})^{\mathsf{T}} & \underline{Z}\underline{Z}^{\mathsf{T}} \end{pmatrix} / \sigma^4, \end{aligned}$$

then $[E[\omega(\underline{\beta}, \underline{\beta}_Z, y, \underline{W}, \underline{Z}, \mathbf{P}_{\underline{Z}})|y, \underline{X}, \underline{X}] = (y - \underline{X}^{\mathsf{T}} \underline{\beta} - \underline{Z}^{\mathsf{T}} \underline{\beta}_Z)^2 \begin{pmatrix} \underline{X}\underline{X}^{\mathsf{T}} & \underline{X}\underline{Z}^{\mathsf{T}} \\ \underline{Z}\underline{X}^{\mathsf{T}} & \underline{Z}\underline{Z}^{\mathsf{T}} \end{pmatrix} / \sigma^4$, and we can use

$$\begin{aligned} \widehat{\mathrm{Var}_2} \begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix} = &\mathbf{I}^*(\widehat{\underline{\beta}}_s, \widehat{\underline{\beta}}_{Z_s}, \underline{Y}, \mathbf{W}, \mathbf{Z})^{-1} \Big\{ \mathbf{V}(\widehat{\underline{\beta}}_s, \widehat{\underline{\beta}}_{Z_s}, \underline{Y}, \mathbf{W}, \mathbf{Z}) \\ &- \sum_{i=1}^{n} \omega(\widehat{\underline{\beta}}_s, \widehat{\underline{\beta}}_{Z_s}, y_i, \underline{W}_i, \underline{Z}_i, \mathbf{P}_{\underline{Z}_i}) \Big\} \mathbf{I}^*(\widehat{\underline{\beta}}_s, \widehat{\underline{\beta}}_{Z_s}, \underline{Y}, \mathbf{W}, \mathbf{Z})^{-1} \\ &+ \mathbf{I}^*(\widehat{\underline{\beta}}_s, \widehat{\underline{\beta}}_{Z_s}, \underline{Y}, \mathbf{W}, \mathbf{Z})^{-1} \end{aligned}$$

as an asymptotic covariance matrix for $\begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix}$.

4.5.2.1 When the Misclassification Model is Estimated

In this section, we will assume the misclassification model is estimated from external data, like in section 4.3.1, and we investigate how the asymptotic covariance matrix of $\begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix}$ changes.

When $\mathbf{P}_{\underline{Z}}$ are estimated from external data, and $\widehat{\mathbf{P}}_{\underline{Z}}$ is a consistent estimator for $\mathbf{P}_{\underline{Z}}$ for every $\underline{Z}$, then $\begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix}$ is a consistent estimator if the validation size for each $\underline{Z}$ and main study sample size all go to infinity. Assume Σ_K is the asymptotic covariance matrix of $\begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix}$, if $\widehat{\mathbf{P}}_{\underline{Z}_i}$ are assumed to be known (see section 4.3.1). We can estimate the covariance of $\begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix}$:

$$\widehat{\text{Var}} \begin{pmatrix} \widehat{\underline{\beta}}_s \\ \widehat{\underline{\beta}}_{Z_s} \end{pmatrix} = \Sigma_k + \sum_{i=1}^{n} H_{11\underline{Z}_i}^{-1} H_{12\underline{Z}_i} \Sigma_{\underline{Z}_i} H_{12\underline{Z}_i}^{\mathsf{T}} (H_{11\underline{Z}_i}^{-1})^{\mathsf{T}}, \text{ where}$$

$$\begin{aligned}
\Sigma_{\underline{Z}_i} &= \begin{pmatrix} \Sigma_{0\underline{Z}_i} & \mathbf{0} & \ldots & \mathbf{0} \\ \underline{0} & \Sigma_{1\underline{Z}_i} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \Sigma_{(K-1)\underline{Z}_i} \end{pmatrix}, \text{ and} \\
\Sigma_{k\underline{Z}_i} &= \frac{1}{N_{.k\underline{Z}_i}} \begin{pmatrix} \widehat{\theta}_{0k\underline{Z}_i}(1-\widehat{\theta}_{0k\underline{Z}_i}) & -\widehat{\theta}_{0k\underline{Z}_i}\widehat{\theta}_{1k\underline{Z}_i} & \ldots & -\widehat{\theta}_{0k\underline{Z}_i}\widehat{\theta}_{(K-2)k\underline{Z}_i} \\ -\widehat{\theta}_{1k\underline{Z}_i}\widehat{\theta}_{0k\underline{Z}_i} & \widehat{\theta}_{1k\underline{Z}_i}(1-\widehat{\theta}_{1k\underline{Z}_i}) & \ldots & -\widehat{\theta}_{1k\underline{Z}_i}\widehat{\theta}_{(K-2)k\underline{Z}_i} \\ \ldots & \ldots & \ldots & \ldots \\ -\widehat{\theta}_{(K-2)k\underline{Z}_i}\widehat{\theta}_{0k\underline{Z}_i} & -\widehat{\theta}_{(K-2)k\underline{Z}_i}\widehat{\theta}_{1k\underline{Z}_i} & \ldots & \widehat{\theta}_{(K-2)k\underline{Z}_i}(1-\widehat{\theta}_{(K-2)k\underline{Z}_i}) \end{pmatrix} \\
H_{11\underline{Z}_i} &= (\widehat{\sigma}^2)^{-1} \begin{pmatrix} \mathbf{D}_{\widehat{\underline{X}}_i} & \widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\underline{W}_i\underline{Z}_i^{\mathsf{T}} \\ \underline{Z}_i\underline{W}_i^{\mathsf{T}}(\widehat{\mathbf{P}}_{\underline{Z}_i}^{\mathsf{T}})^{-1} & \underline{Z}_i^{\mathsf{T}}\underline{Z}_i \end{pmatrix} \\
H_{12\underline{Z}_i} &= (\underline{h}_{00\underline{Z}_i}, \underline{h}_{10\underline{Z}_i}, \ldots, \underline{h}_{(K-2)0\underline{Z}_i}, \underline{h}_{01\underline{Z}_i}, \ldots, \underline{h}_{(K-2)1\underline{Z}_i}, \ldots, \underline{h}_{0(K-1)\underline{Z}_i}, \ldots, \underline{h}_{(K-2)(K-1)\underline{Z}_i}) \text{ where} \\
\underline{h}_{ml\underline{Z}_i} &= (\widehat{\sigma}_2)^{-1} \begin{pmatrix} -y_i\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\mathbf{M}_{ml}\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\underline{W}_i + \widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\mathbf{M}_{ml}\mathbf{D}_{\widehat{\underline{X}}_i}\widehat{\underline{\beta}}_s + \underline{Z}_i^{\mathsf{T}}\widehat{\underline{\beta}}_{\mathbf{Z}_s}\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\mathbf{M}_{ml}\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\underline{W}_i \\ \underline{W}_i^{\mathsf{T}}(\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\mathbf{M}_{ml}\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1})^{\mathsf{T}}\underline{\beta}_s\underline{Z}_i \end{pmatrix} \text{ with} \\
\mathbf{D}_{\widehat{\underline{X}}_i} &= \sum_{k=0}^{K-1} \left\{ (\widehat{\mathbf{P}}_{\underline{Z}_i}^{-1}\underline{W}_i)^{\mathsf{T}}\underline{e}_k \right\} \underline{e}_k\underline{e}_k^{\mathsf{T}}.
\end{aligned}$$

4.5.3 Reclassification Case

Now, we will consider how to use a reclassification model to get a corrected estimator for regression with misclassified covariates and perfectly measured covariates. Just like in section 4.5.2, if $P(\underline{X}|\underline{W})$ depends on $\underline{Z}$.

Define

$$P(\underline{X}=i|\underline{W}=j,\underline{Z})=\gamma_{ij\underline{Z}} \text{ and } \mathbf{Q}_{\underline{Z}}=\begin{pmatrix} \gamma_{00\underline{Z}} & \gamma_{01\underline{Z}} & \cdots & \gamma_{0(K-1)\underline{Z}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{(K-1)0\underline{Z}} & \gamma_{(K-1)1\underline{Z}} & \cdots & \gamma_{(K-1)(K-1)\underline{Z}} \end{pmatrix}$$

for each $\underline{Z}$. Let $\widehat{\underline{X}}_i = \mathbf{Q}_{\underline{Z}_i}\underline{W}_i$ and $\widehat{\mathbf{X}} = (\widehat{\underline{X}}_1 \ \ \ldots \ \ \widehat{\underline{X}}_n)^{\mathsf{T}}$. From Section 4.4, we could regress y_i on $\widehat{\underline{X}}_i, \underline{Z}_i$, then use the least squares method to get a corrected estimator:

$$\begin{pmatrix} \widehat{\underline{\beta}}_R \\ \widehat{\underline{\beta}}_{Z_R} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{X}}^{\mathsf{T}}\widehat{\mathbf{X}} & \widehat{\mathbf{X}}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\widehat{\mathbf{X}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\mathbf{X}}^{\mathsf{T}}\underline{Y} \\ \mathbf{Z}^{\mathsf{T}}\underline{Y} \end{pmatrix}.$$

If $\mathbf{Q}_{\underline{Z}} = \mathbf{Q}$ for all $\underline{Z}$, then

$$\begin{aligned} \begin{pmatrix} \widehat{\underline{\beta}}_{\mathbf{Q}} \\ \widehat{\underline{\beta}}_{Z_{\mathbf{Q}}} \end{pmatrix} &= \begin{pmatrix} \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}} & \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Q}\mathbf{W}^{\mathsf{T}}\underline{Y} \\ \mathbf{Z}^{\mathsf{T}}\underline{Y} \end{pmatrix} \\ &= \left\{ \begin{pmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{W}^{\mathsf{T}}\mathbf{W} & \mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{Q}^{\mathsf{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \right\}^{-1} \begin{pmatrix} \mathbf{Q}\mathbf{W}^{\mathsf{T}}\underline{Y} \\ \mathbf{Z}^{\mathsf{T}}\underline{Y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix}, \end{aligned}$$

and from the proof of Lemma 4.5.1, $E\left[\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix}\right] = \begin{pmatrix} \mathbf{Q}\underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}$. As a result,

$$E\left[\begin{pmatrix} \widehat{\underline{\beta}}_{\mathbf{Q}} \\ \widehat{\underline{\beta}}_{Z_{\mathbf{Q}}} \end{pmatrix}\right] = \begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}.$$

Again, this method assumes $\mathbf{Q}_{\underline{Z}}$ does depend on $\underline{Z}$.

The asymptotic covariance matrix of $\begin{pmatrix} \widehat{\underline{\beta}}_{\mathbf{Q}} \\ \widehat{\underline{\beta}}_{Z_{\mathbf{Q}}} \end{pmatrix}$ is

$$\widehat{\text{Var}}\begin{pmatrix} \widehat{\underline{\beta}}_{\mathbf{Q}} \\ \widehat{\underline{\beta}}_{Z_{\mathbf{Q}}} \end{pmatrix} = \begin{pmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \widehat{\text{Var}}\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} \begin{pmatrix} \mathbf{Q}^{\mathsf{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1},$$

where

$$\widehat{\text{Var}}\begin{pmatrix} \widehat{\underline{\beta}}_W \\ \widehat{\underline{\beta}}_{Z_W} \end{pmatrix} = \widehat{\sigma}^2_W \begin{pmatrix} \mathbf{W}^{\mathsf{T}}\mathbf{W} & \mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}^{-1}$$

and $\widehat{\sigma}^2_W = (\underline{Y} - \mathbf{W}\underline{\beta}_W - \mathbf{Z}\underline{\beta}_{Z_W})^{\mathsf{T}}(\underline{Y} - \mathbf{W}\underline{\beta}_W - \mathbf{Z}\underline{\beta}_{Z_W})/n$.

$\widehat{\sigma}^2_W$ is not an estimator for σ^2. From

$$\begin{aligned} \text{Var}(y) &= E\{\,\text{Var}(y|\underline{X},\underline{Z})\} + \text{ Var}\{E(y|\underline{X},\underline{Z})\} \\ &= \sigma^2 + \text{ Var}(\underline{X}^{\mathsf{T}}\underline{\beta} + \underline{Z}^{\mathsf{T}}\underline{\beta}_Z), \end{aligned}$$

without knowing the correlation between $\underline{X}$ and $\underline{Z}$, we can not estimate σ^2. If $\underline{X}$ and $\underline{Z}$ are uncorrelated, then we could get an estimator for σ^2 :

$$\widehat{\sigma}^2 = \widehat{\text{Var}(y)} - \widehat{\underline{\beta}}_Q^{\mathsf{T}}\,\widehat{\text{Var}}(\underline{X})\widehat{\underline{\beta}}_Q - \widehat{\underline{\beta}}_{Z_Q}^{\mathsf{T}}\,\widehat{\text{Var}}(\underline{Z})\widehat{\underline{\beta}}_{Z_Q}.$$

Sometimes the reclassification model is not available, and we need to estimate it from external data. If $\mathbf{Q}$ is estimated from external data, and $\widehat{\mathbf{Q}}$ is a consistent estimator for $\mathbf{Q}$, then $(\,\widehat{\underline{\beta}}_{\widehat{\mathbf{Q}}}^{\mathsf{T}} \quad \widehat{\underline{\beta}}_{Z_{\widehat{\mathbf{Q}}}}^{\mathsf{T}}\,)^{\mathsf{T}}$ is a consistent estimator for $(\,\underline{\beta}^{\mathsf{T}} \quad \underline{\beta}_Z^{\mathsf{T}}\,)^{\mathsf{T}}$ if the validation sample size and the main study sample size both go to infinity.

The estimating equation for $\begin{pmatrix} \widehat{\underline{\beta}}_{\mathbf{Q}} \\ \widehat{\underline{\beta}}_{Z_{\mathbf{Q}}} \end{pmatrix}$ is

$$S(\underline{\beta},\underline{\gamma}) = \begin{pmatrix} \mathbf{Q}\mathbf{W}^{\mathsf{T}} \\ \mathbf{Z}^{\mathsf{T}} \end{pmatrix}\underline{Y} - \begin{pmatrix} \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}} & \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix}\begin{pmatrix} \underline{\beta} \\ \underline{\beta}_Z \end{pmatrix}.$$

Using the same notation as in Section 4.4 and applying Parke's (1986) method, the asymptotic covariance matrix of $\begin{pmatrix} \widehat{\underline{\beta}}_{\widehat{\mathbf{Q}}} \\ \widehat{\underline{\beta}}_{Z_{\widehat{\mathbf{Q}}}} \end{pmatrix}$ is

$$\widehat{\text{Var}}\begin{pmatrix} \widehat{\underline{\beta}}_{\widehat{\mathbf{Q}}} \\ \widehat{\underline{\beta}}_{Z_{\widehat{\mathbf{Q}}}} \end{pmatrix} = \Sigma_K + H_{11}^{-1}H_{12}\Sigma_{\underline{\gamma}}H_{12}^{\mathsf{T}}(H_{11}^{-1})^{\mathsf{T}} \text{ where}$$

$$\begin{aligned} H_{11} &= -\begin{pmatrix} \widehat{\mathbf{Q}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\widehat{\mathbf{Q}}^{\mathsf{T}} & \widehat{\mathbf{Q}}\mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W}\widehat{\mathbf{Q}}^{\mathsf{T}} & \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \end{pmatrix} \\ H_{12} &= (\underline{h}_{00}, \underline{h}_{10}, \ldots, \underline{h}_{(K-2)0}, \underline{h}_{01}, \ldots, \underline{h}_{(K-2)1}, \ldots, \underline{h}_{0(K-1)}, \ldots, \underline{h}_{(K-2)(K-1)}) \text{ with} \\ h_{ml} &= \begin{pmatrix} \mathbf{M}_{ml}\mathbf{W}^{\mathsf{T}}\underline{Y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{M}_{ml}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{Q}^{\mathsf{T}} + \mathbf{Q}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{M}_{ml}^{\mathsf{T}} & \mathbf{M}_{ml}\mathbf{W}^{\mathsf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{M}_{ml}^{\mathsf{T}} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \widehat{\underline{\beta}}_{\widehat{\mathbf{Q}}} \\ \widehat{\underline{\beta}}_{Z_{\widehat{\mathbf{Q}}}} \end{pmatrix}, \end{aligned}$$

and Σ_K is the covariance estimate of $\begin{pmatrix} \widehat{\underline{\beta}}_{\widehat{\mathbf{Q}}} \\ \widehat{\underline{\beta}}_{Z_{\widehat{\mathbf{Q}}}} \end{pmatrix}$ if we assume $\widehat{\mathbf{Q}}$ is known.

CHAPTER 5

SIMULATION

In this chapter, we use computer simulation to generate data, compare and evaluate the performance of some of the methods described in this dissertation. The order of simulations is not necessarily the same as the order in which the methods were presented. We use equal validation size for each category in all the simulation.

5.1 Bias Reduced Estimator, Partially Corrected Estimator

The first simulation addresses the problem of estimating a single proportion. We introduce the bias reduced estimator in Section 2.1, and reduced mean square error estimator/partial correction estimator in Section 2.2 for proportion of interest. In this section, we will use simulation to demonstrate the performance of these two estimators. We should make a note that not all naive estimators are biased. From Section 2.2, we know that the bias of a naive estimator is $(\theta_{00} + \theta_{11} - 2)\pi + (1 - \theta_{00})$. Also from Figure 1, we can see that there is a linear equation of $(\theta_{00}, \theta_{11})$, for every π, such that the naive estimator has bias 0. For this reason, we have chosen values for θ_{00}, θ_{11} to give different levels of bias in naive estimators. Table 1 summarizes the levels of θ_{00} and θ_{11} we will use in this section.

We will denote bias 0.2 as a high level of bias , an absolute bias of 0.075 as the medium level of bias and bias 0 as the unbiased level. Along with the bias levels, the main study sample sizes are $n = 50, 100, 1000,$ and validation sizes are either half, same, or

π	bias	θ_{00}	θ_{11}
0.2	0	0.85	0.4
	0.075	0.85	0.775
	0.2	0.75	1
0.5	0	0.75	0.75
	-0.075	0.9	0.75
	0.2	0.55	0.95

Table 1. Parameter Settings for the First Simulation Experiment that Compares Estimators of Prevalence

double the main study size. We also include the case where the misclassification model is known. For each combination of parameters settings, we repeat 2000 times.

Figure 3 compares the Monte Carlo estimate of the root of mean square error (RMSE) of $\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}, \widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$. Figure 4 contains the absolute values of the Monte Carlo estimate of bias for $\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}$,$\widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$. Table 2 and Table 3 contain the same information in tabular form as Figures 3 and 4 respectively. From these figures, we can see that the RMSE of $\widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$ are smaller than or close to the RMSE of $\widehat{\pi}_{PlugIn}$ for the high and medium bias levels and when the validation sizes are half or the same as the main study size. The $\widehat{\pi}_{pc}$ performed much better than $\widehat{\pi}_{PlugIn}$ in terms of both absolute bias and RMSE when the naive estimators is unbiased. The absolute biases of $\widehat{\pi}_{corrected,PI}$ are about the same as $\widehat{\pi}_{PlugIn}$ for the high and medium bias levels. The performance of $\widehat{\pi}_{naive}$ depends on the level of bias. By looking at the tables, for $\pi = 0.5$, $\widehat{\pi}_{Corrected,PI}$ has smaller RMSE than $\widehat{\pi}_{PlugIn}$, and has absolute bias smaller than or almost equal to the absolute bias of $\widehat{\pi}_{PlugIn}$. For $\pi = 0.2, n = 50$, $\widehat{\pi}_{Corrected,PI}$ has not performed as expected for the unbiased parameter setting. We suspect this is because of large remainder terms in the asymptotic expansion on which the estimator is based.

5.1.1 Overall Comparison with Different Misclassification Probabilities

In this section we compare $\widehat{\pi}_{naive}, \widehat{\pi}_{PlutIn}, \widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$ for different πs. We use two misclassification models: one is $\theta_{11} = 0.90, \theta_{00} = 0.95$, and the other is $\theta_{11} =$

Table 2. Root of Mean Square Error

π	N	L	$\widehat{\pi}_{naive}$	$\widehat{\pi}_{PlugIn}$	$\widehat{\pi}_{C,PI}$	$\widehat{\pi}_{pc}$	$\widehat{\pi}_{naive}$	$\widehat{\pi}_{PlugIn}$	$\widehat{\pi}_{C,PI}$	$\widehat{\pi}_{pc}$	$\widehat{\pi}_{naive}$	$\widehat{\pi}_{PlugIn}$	$\widehat{\pi}_{C,PI}$	$\widehat{\pi}_{pc}$
			$n = 50$				$n = 100$				$n = 1000$			
0.2	0.5	H	0.211	0.111	0.111	0.117	0.205	0.090	0.089	0.090	0.201	0.029	0.029	0.029
0.2	0.5	M	0.097	0.116	0.112	0.098	0.087	0.093	0.090	0.082	0.076	0.031	0.030	0.033
0.2	0.5	N	0.056	0.227	0.316	0.122	0.040	0.180	0.227	0.109	0.012	0.073	0.072	0.048
0.2	1.0	H	0.212	0.103	0.103	0.107	0.205	0.078	0.078	0.078	0.201	0.025	0.025	0.025
0.2	1.0	M	0.100	0.113	0.110	0.102	0.087	0.083	0.082	0.081	0.076	0.027	0.027	0.028
0.2	1.0	N	0.057	0.219	0.245	0.130	0.039	0.161	0.156	0.101	0.013	0.062	0.061	0.042
0.2	2.0	H	0.212	0.098	0.098	0.100	0.208	0.073	0.072	0.073	0.201	0.023	0.023	0.023
0.2	2.0	M	0.099	0.108	0.107	0.103	0.089	0.078	0.078	0.083	0.076	0.025	0.025	0.025
0.2	2.0	N	0.058	0.203	0.198	0.128	0.041	0.153	0.149	0.098	0.013	0.058	0.058	0.040
0.2	Inf	H	0.209	0.087	0.087	0.090	0.206	0.067	0.067	0.068	0.201	0.021	0.021	0.021
0.2	Inf	M	0.097	0.095	0.095	0.106	0.087	0.070	0.070	0.080	0.076	0.023	0.023	0.023
0.2	Inf	N	0.057	0.176	0.176	0.111	0.039	0.135	0.135	0.085	0.013	0.051	0.051	0.036
0.5	0.5	H	0.210	0.165	0.154	0.155	0.205	0.124	0.118	0.122	0.200	0.040	0.039	0.040
0.5	0.5	M	0.101	0.142	0.133	0.114	0.092	0.099	0.097	0.091	0.077	0.031	0.031	0.033
0.5	0.5	N	0.070	0.188	0.170	0.119	0.052	0.140	0.133	0.096	0.016	0.043	0.043	0.032
0.5	1.0	H	0.211	0.158	0.152	0.159	0.206	0.111	0.109	0.115	0.201	0.034	0.034	0.034
0.5	1.0	M	0.102	0.127	0.125	0.114	0.089	0.088	0.087	0.087	0.077	0.028	0.028	0.029
0.5	1.0	N	0.070	0.173	0.166	0.119	0.050	0.119	0.117	0.085	0.016	0.038	0.038	0.028
0.5	2.0	H	0.210	0.139	0.137	0.147	0.207	0.101	0.101	0.107	0.200	0.032	0.032	0.032
0.5	2.0	M	0.099	0.115	0.114	0.110	0.089	0.082	0.081	0.086	0.077	0.026	0.026	0.027
0.5	2.0	N	0.068	0.152	0.149	0.105	0.050	0.111	0.110	0.080	0.016	0.035	0.035	0.025
0.5	Inf	H	0.214	0.130	0.130	0.142	0.205	0.089	0.089	0.094	0.201	0.029	0.029	0.030
0.5	Inf	M	0.100	0.107	0.107	0.115	0.088	0.074	0.074	0.082	0.076	0.025	0.025	0.025
0.5	Inf	N	0.070	0.140	0.140	0.097	0.048	0.096	0.096	0.065	0.016	0.032	0.032	0.022

Table 3. Absolute Value of Bias

π	N	L	$\widehat{\pi}_{naive}$	$\widehat{\pi}_{PlugIn}$	$\widehat{\pi}_{C,PI}$	$\widehat{\pi}_{pc}$	$\widehat{\pi}_{naive}$	$\widehat{\pi}_{PlugIn}$	$\widehat{\pi}_{C,PI}$	$\widehat{\pi}_{pc}$	$\widehat{\pi}_{naive}$	$\widehat{\pi}_{PlugIn}$	$\widehat{\pi}_{C,PI}$	$\widehat{\pi}_{pc}$
			$n = 50$				$n = 100$				$n = 1000$			
0.2	0.5	H	0.199	0.020	0.032	0.064	0.199	0.007	0.012	0.033	0.200	0.001	0.000	0.003
0.2	0.5	M	0.074	0.017	0.026	0.048	0.075	0.003	0.007	0.033	0.074	0.001	0.001	0.008
0.2	0.5	N	0.000	0.080	0.181	0.038	0.001	0.058	0.119	0.032	0.000	0.000	0.002	0.001
0.2	1.0	H	0.200	0.013	0.019	0.043	0.200	0.000	0.002	0.017	0.200	0.000	0.000	0.002
0.2	1.0	M	0.077	0.012	0.015	0.044	0.075	0.001	0.002	0.029	0.075	0.001	0.001	0.005
0.2	1.0	N	0.003	0.075	0.126	0.040	0.001	0.041	0.058	0.024	0.000	0.001	0.000	0.000
0.2	2.0	H	0.200	0.005	0.007	0.026	0.202	0.002	0.003	0.013	0.200	0.000	0.000	0.001
0.2	2.0	M	0.075	0.006	0.008	0.038	0.077	0.005	0.005	0.032	0.075	0.000	0.000	0.004
0.2	2.0	N	0.001	0.060	0.078	0.031	0.002	0.037	0.042	0.022	0.000	0.000	0.001	0.000
0.2	Inf	H	0.197	0.000	0.000	0.011	0.200	0.001	0.001	0.006	0.200	0.000	0.000	0.001
0.2	Inf	M	0.073	0.001	0.001	0.036	0.075	0.000	0.000	0.023	0.075	0.000	0.000	0.003
0.2	Inf	N	0.000	0.044	0.044	0.021	0.001	0.014	0.014	0.008	0.000	0.001	0.001	0.001
0.5	0.5	H	0.201	0.006	0.015	0.068	0.200	0.011	0.002	0.037	0.200	0.001	0.000	0.005
0.5	0.5	M	0.075	0.006	0.001	0.030	0.077	0.002	0.005	0.031	0.076	0.001	0.001	0.008
0.5	0.5	N	0.000	0.001	0.001	0.000	0.002	0.005	0.005	0.004	0.001	0.001	0.001	0.001
0.5	1.0	H	0.201	0.004	0.005	0.056	0.201	0.003	0.001	0.032	0.200	0.001	0.001	0.003
0.5	1.0	M	0.074	0.003	0.000	0.032	0.074	0.001	0.001	0.027	0.075	0.001	0.001	0.005
0.5	1.0	N	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000
0.5	2.0	H	0.201	0.002	0.002	0.048	0.202	0.002	0.004	0.030	0.200	0.001	0.000	0.003
0.5	2.0	M	0.071	0.007	0.005	0.030	0.074	0.002	0.001	0.025	0.075	0.001	0.001	0.004
0.5	2.0	N	0.002	0.003	0.003	0.002	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.5	Inf	H	0.204	0.007	0.007	0.046	0.201	0.001	0.001	0.022	0.200	0.000	0.000	0.002
0.5	Inf	M	0.072	0.004	0.004	0.035	0.074	0.001	0.001	0.023	0.075	0.000	0.000	0.002
0.5	Inf	N	0.003	0.006	0.006	0.004	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000

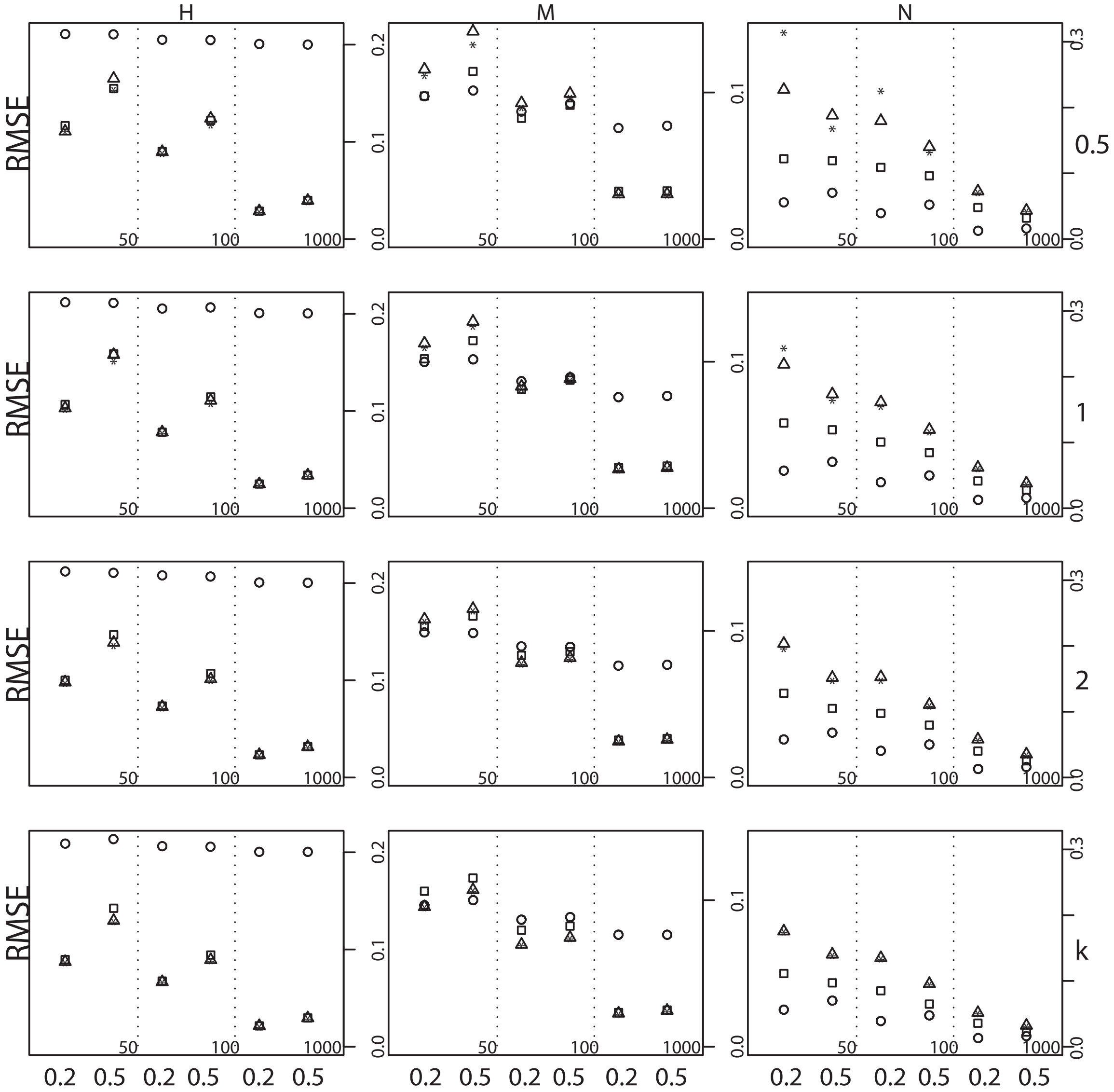


Figure 3. Comparison of Root Mean Square Errors in the First Simulation Experiment (Prevalence): this figure compares the root of mean square error among $\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}, \widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$. Circle is for $\widehat{\pi}_{naive}$, triangle point-up is for $\widehat{\pi}_{PlugIn}$, $'*'$ is for $\widehat{\pi}_{corrected,PI}$, and square is for $\widehat{\pi}_{pc}$. The number in each area is the main study sample size.'H', 'M', 'N' are levels of high, medium, no bias respectively. '0.5','1','2' are validation size which are half, same, double main study size respectively; 'k' means the misclassification is known (please note that the $y-$axes have different scales).

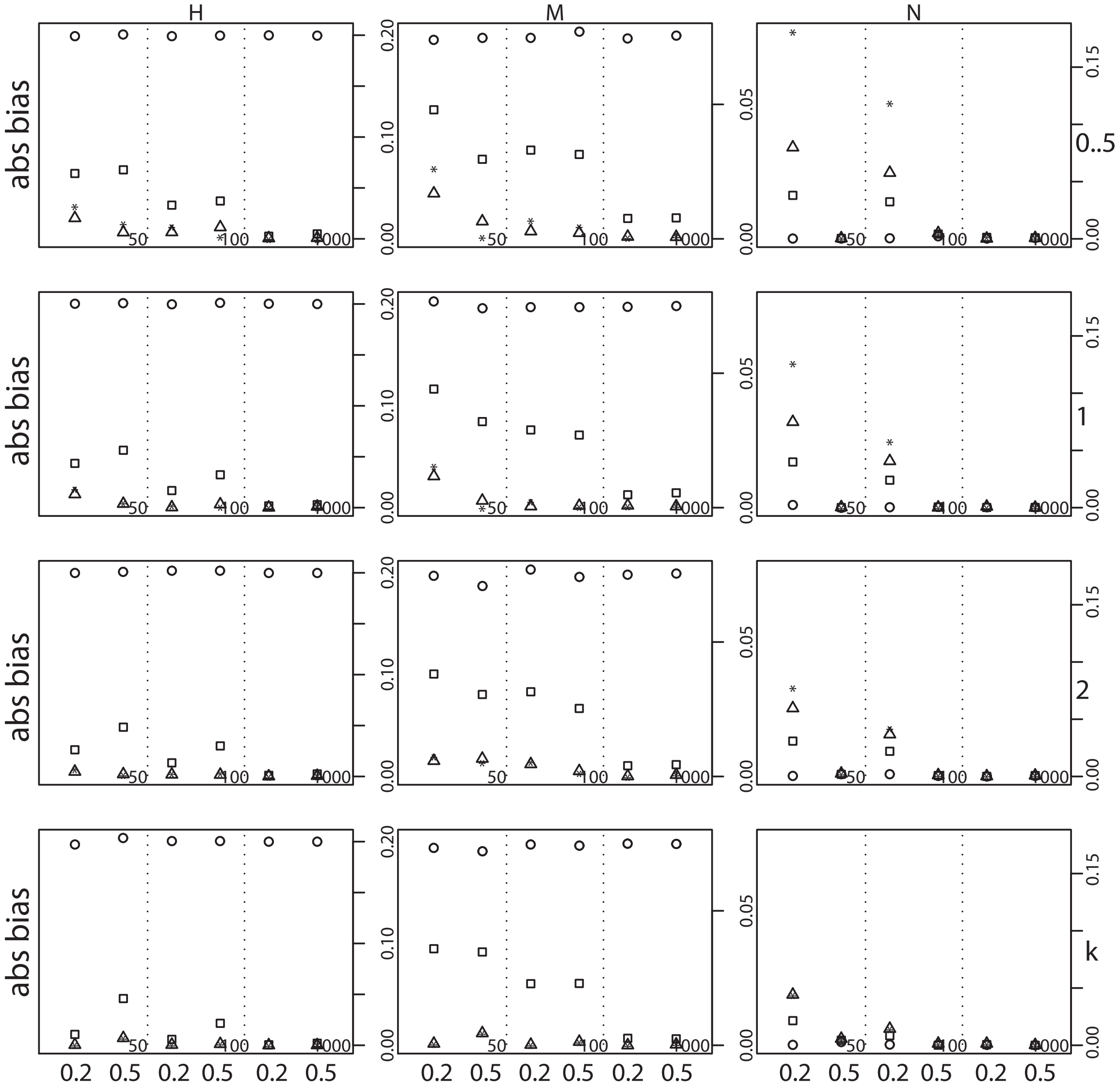


Figure 4. Comparison of Biases in the First Simulation Experiment (Prevalence): this figure compares biases among $\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}, \widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$. Circle is for $\widehat{\pi}_{naive}$, triangle point-up is for $\widehat{\pi}_{PlugIn}$, $'*'$ is for $\widehat{\pi}_{corrected,PI}$, and square is for $\widehat{\pi}_{pc}$. The number in each area is the main study sample size. 'H', 'M', 'N' are levels of high, medium, no bias respectively. '0.5','1','2' are validation size which are half, same, double main study size respectively; 'k' means the misclassification is known (please note that the $y-$axes have different scales).

$0.90, \theta_{00} = 0.75$. For both misclassification models, we use $\pi = (0.01, 0.081, 0.161. \dots, 0.961)$, main study sizes $n = 50, 200, 1000$, and validation sizes $N = 25, 100$. We consider full factorial combinations of these parameters, and the Monte Carlo sample size is 2000.

Figure 5 shows the results for the first misclassification model, and Figure 6 has the results for the second misclassification model. From both figures, we can see that when $N = 25$, there is a wide range of π for which the root of mean square error of $\widehat{\pi}_{pc}$ is smaller than $\widehat{\pi}_{PlugIn}$. When $\theta_{11} = 0.9, \theta_{00} = 0.75, N = 25$, when the main study size increases, and when $\pi = 0.401, 0.481, 0.561$, both RMSE and the absolute bias of $\widehat{\pi}_{corrected,PI}$ are smaller than $\widehat{\pi}_{PlugIn}$.

5.2 Comparison of Optimization Method and Fieller's Method

In this section, we will present the simulation results for confidence intervals for π, the proportion of interest. We use optimization method and Fieller's methods. We developed these methods in Chapter 3, and we compare them with delta method. The parameter settings are factorial combinations of $\pi = (0, 0.05, \dots, 1)$, misclassification probabilities $\theta_{11} = (0.6, 0.8, 0.95), \theta_{00} = (0.5, 0.58, \dots, 0.98)$, main study sizes $(100, 500, 1000)$, and with equal validation sizes $(10, 20 \dots, 150)$. We use α as 0.05 and $\alpha_i = 1 - (0.95)^{1/3}$. We simulate 1000 Monte Carlo replicates for each parameter setting. This results in 19,845,000 iterations.

In the optimization method, if we use exact or score confidence intervals, the average coverage and confidence interval length will be bigger than Wald's confidence intervals. Since the exact and score confidence interval have coverage levels that are nominal or higher, the joint confidence interval's coverage will be even higher if w use these methods. As a result, we use Wald's confidence interval in optimization programming projection methods. We will refer to this as Wald's interval.

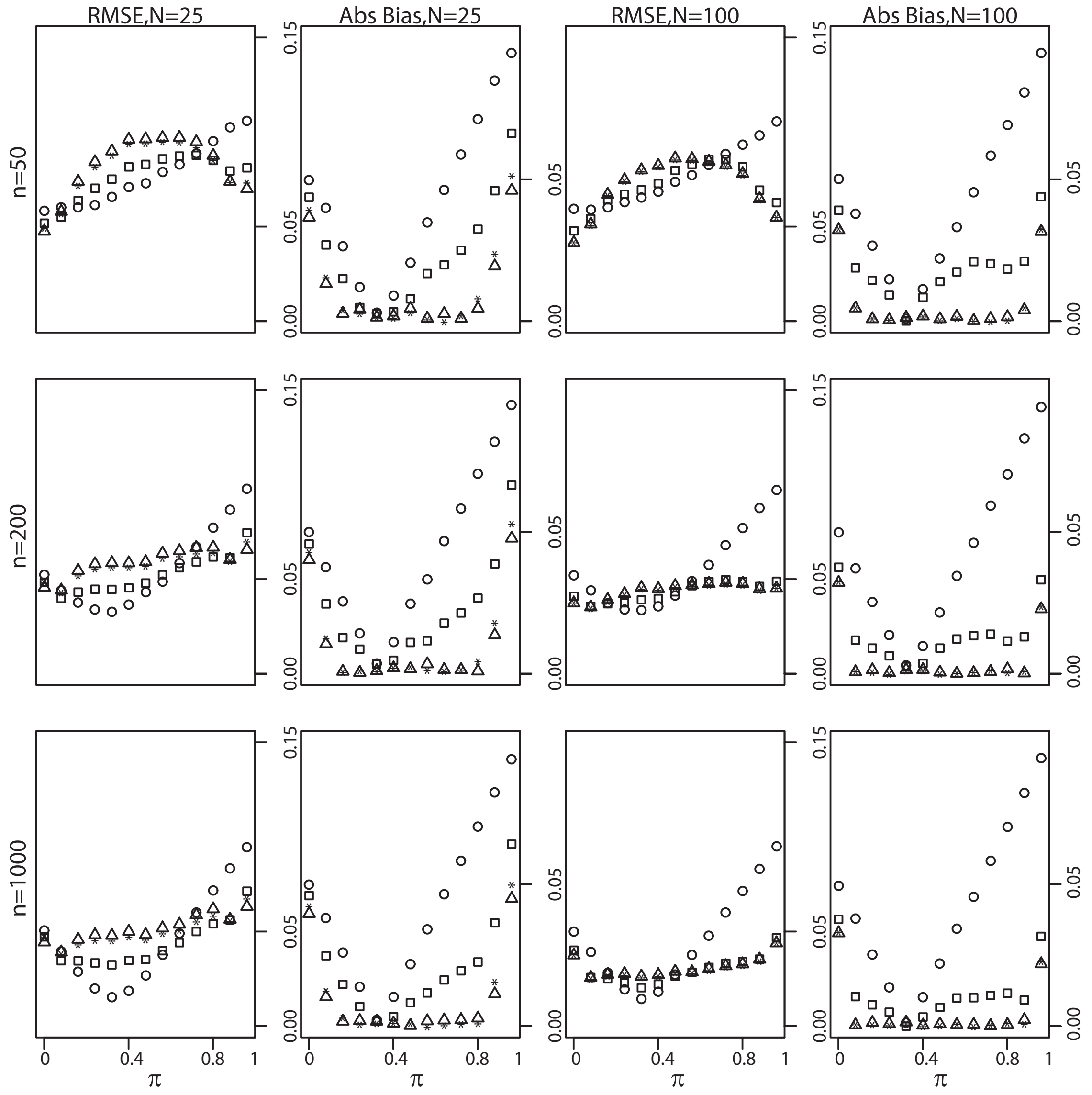


Figure 5. Comparison of Root Mean Square Error and Bias for $\theta_{11} = 0.9, \theta_{00} = .95$ (Prevalence): this figure compares the root of mean square error and absolute of bias when $\theta_{11} = 0.9, \theta_{00} = 0.95$, among $\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}, \widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$. Circle is for $\widehat{\pi}_{naive}$, triangle point-up is for $\widehat{\pi}_{PlugIn}$, $'*'$ is for $\widehat{\pi}_{corrected,PI}$, and square is for $\widehat{\pi}_{pc}$.

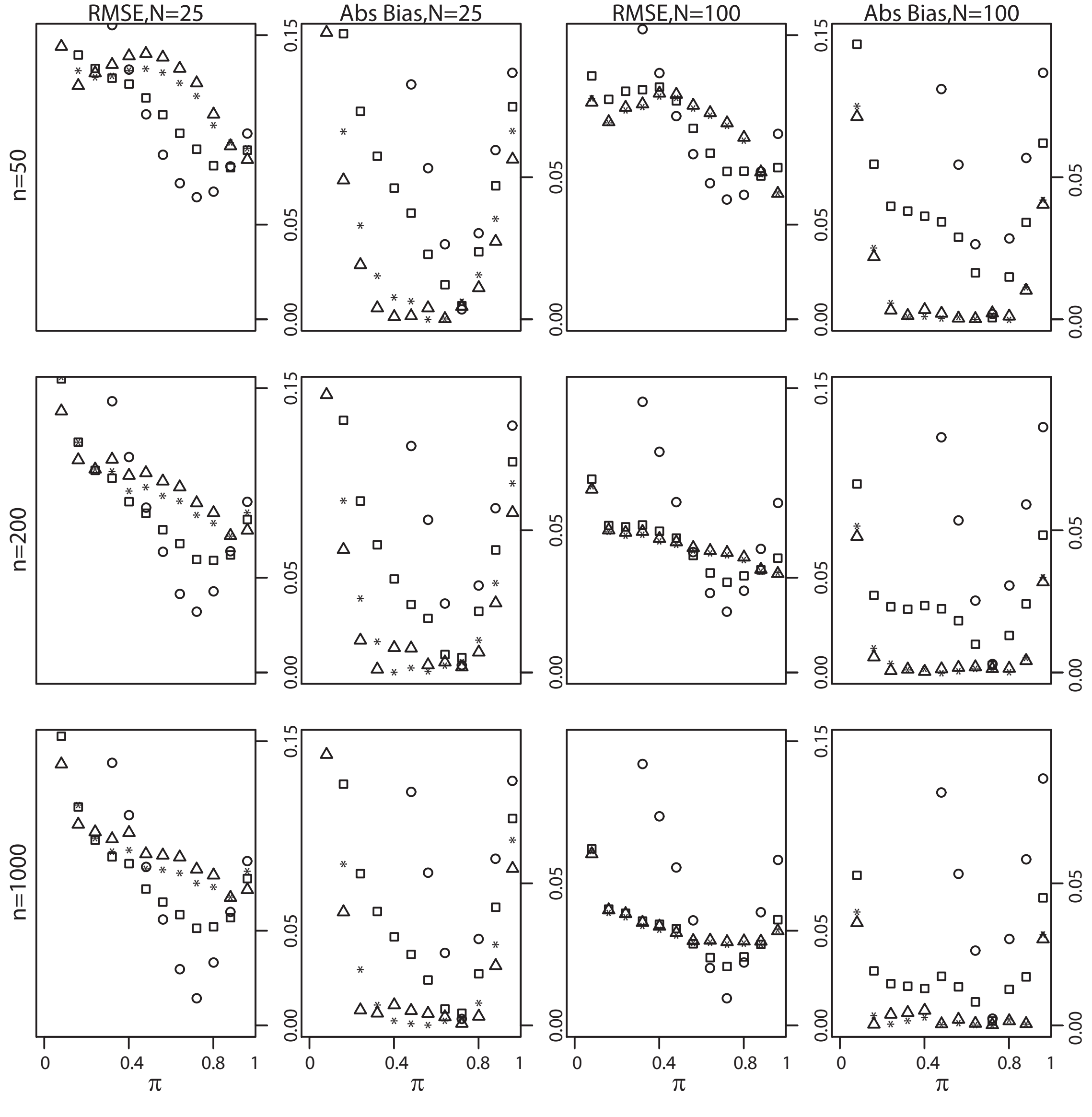


Figure 6. Comparison of Root Mean Square Error and Bias for $\theta_{11} = 0.9, \theta_{00} = .75$ (Prevalence): this figure compares the root of mean square error and absolute of bias when $\theta_{11} = 0.9, \theta_{00} = 0.75$, among $\widehat{\pi}_{naive}, \widehat{\pi}_{PlugIn}, \widehat{\pi}_{corrected,PI}$ and $\widehat{\pi}_{pc}$. Circle is for $\widehat{\pi}_{naive}$, triangle point-up is for $\widehat{\pi}_{PlugIn}$, $'*'$ is for $\widehat{\pi}_{corrected,PI}$, and square is for $\widehat{\pi}_{pc}$.

Using the multivariate delta method we have

$$\begin{aligned}\widehat{\text{Var}}(\widehat{\pi}_{PlugIn}) &= \frac{\widehat{\pi}_{naive}(1-\widehat{\pi}_{naive})}{n(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^2} + \frac{1}{(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)^2}\left[\frac{\widehat{\theta}_{00}(1-\widehat{\theta}_{00})}{N_{.0}} - 2\widehat{\pi}\frac{\widehat{\theta}_{00}(1-\widehat{\theta}_{00})}{N_{.0}}\right. \\ &\quad \left. +\widehat{\pi}^2\left\{\frac{\widehat{\theta}_{00}(1-\widehat{\theta}_{00})}{N_{.0}} + \frac{\widehat{\theta}_{11}(1-\widehat{\theta}_{11})}{N_{.1}}\right\}\right],\end{aligned}$$

and the $100(1-\alpha)\%$ confidence interval for π using delta method is

$$\widehat{\pi}_{PlugIn} \pm z_{\alpha/2}\,\widehat{\text{Var}}(\widehat{\pi}_{PlugIn})^{0.5}.$$

Figure 7 shows the results of this simulation. The average is taken over all possible combinations of parameter settings for each of the following cases: fixed π, fixed validation size, or fixed θ_{00}. We also considered a naive confidence interval as : $\widehat{\pi}_{naive} \pm z_{\alpha/2}\{\widehat{\pi}_{naive}(1-\widehat{\pi}_{naive})/n\}^{1/2}$. The length of that confidence interval is small, but its average coverage is too small to compare with the other methods, so it is not in Figure 7. It will be in Table 4.

From Figure 7, we can draw three conclusions from the simulation. First, the confidence intervals from optimization are too conservative: the coverage is too high and length is too big. Second, the coverage of Fieller's method and the delta method are both very close to the nominal level, and the coverage of delta method is always higher than the coverage of Fieller's method. When the validation size increases to 110, Fieller's method is at the nominal level, but the coverage for the delta method is above the nominal level. Third, the length of confidence intervals using Fieller's method are generally shorter than those of the delta method. When the misclassification probability θ_{00} increases, the coverage drops, and the lengths of the confidence interval decrease significantly. As we mentioned in Chapter 3, Fieller's method is the projection of an elliptical confidence region, so when the sample size and validation size increase we can expect the coverage will be close to the nominal level, since the normal approximation becomes more accurate.

Table 4 compares the average coverage for sample size $n = 500$ and validation size $N = 100$ using naive intervals, Fieller's method, and the delta method. From the ta-

	Fieller			**Delta**			**Naive**		
θ_{11} θ_{00}	0.6	0.8	0.95	0.6	0.8	0.95	0.6	0.8	0.95
0.5	0.9491	0.9498	0.9462	0.9759	0.9629	0.9549	0.0910	0.1070	0.0902
0.58	0.9494	0.9478	0.9453	0.9673	0.9590	0.9506	0.0985	0.1231	0.1109
0.66	0.9520	0.9474	0.9464	0.9623	0.9573	0.9536	0.1096	0.1469	0.1434
0.74	0.9479	0.9493	0.9427	0.9600	0.9560	0.9492	0.1208	0.1753	0.1940
0.82	0.9492	0.9470	0.9475	0.9612	0.9543	0.9503	0.1320	0.2106	0.2821
0.9	0.9500	0.9473	0.9460	0.9569	0.9530	0.9499	0.1323	0.2515	0.4466
0.98	0.9415	0.9449	0.9428	0.9481	0.9481	0.9451	0.0769	0.2197	0.6655

Table 4. Average Coverage for $n = 500, N = 100$

ble, we can see that the coverage for the naive confidence interval will increase as both misclassification probabilities θ_{00}, θ_{11} increase, but when the misclassification probabilities are $\theta_{00} = 0.98, \theta_{11} = 0.95$, the average coverage for the naive interval is only 67%. The delta method also shows some what higher than nominal coverage unless the misclassification parameters θ_{00}, θ_{11} are high. The average coverage for Fieller's method is generally less than the nominal level, but it is very close to it.

5.3 Confidence Interval for the Slope

In this section, we will consider the method of confidence interval using Fieller's method that we developed in Section 4.2 for the slope of simple linear regression. We compare its coverage with the naive confidence interval, Fieller's method's confidence interval with a known misclassification matrix, a SIMEX confidence interval, and the delta method.

Küchenhoff et al (2006) described how to use the SIMEX method to correct the coefficients in a regression with misclassified covariates or response and a known misclassification model. Küchenhoff et al (2007) incorporated the variability of estimated misclassification parameters. The simulation in this section will compare three situa-

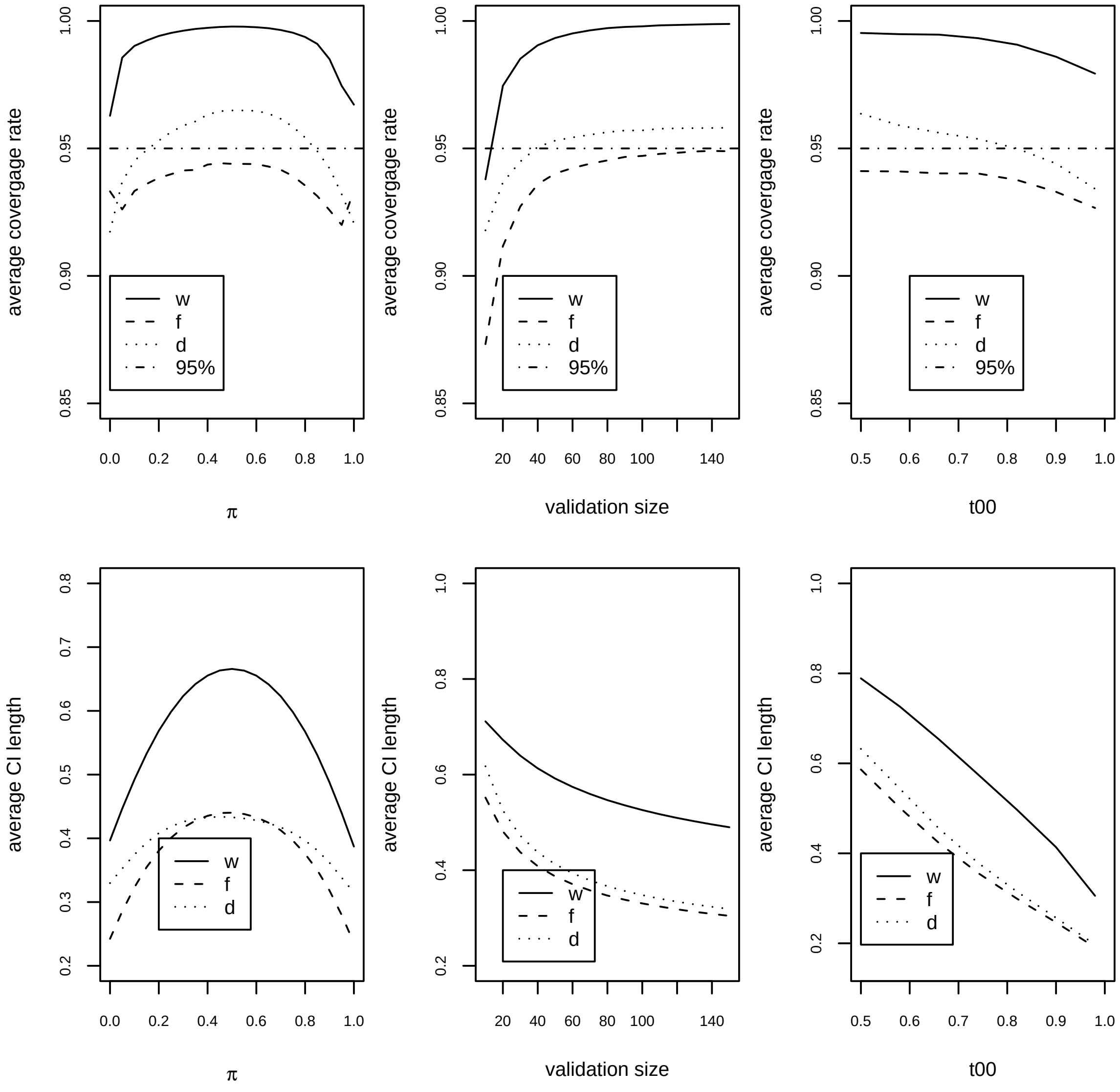


Figure 7. Comparison of Optimization Method, Filler's Method and the Delta Method (Prevalence): this figure compares the average coverage and average length of confidence interval for π using optimization with Wald intervals, Fieller's and delta methods. 'w' is for optimization with Wald intervals, 'f' is for Fieller's method, and 'd' is for delta method.

tions: a known misclassification model, estimated misclassification parameters that are treated as known, and a misclassification model that is estimated and treated as such. The delta method is derived in Appendix C.

The simulation estimates coverage, and confidence interval length for the slope. We consider simple linear regression with misclassified covariates and $K = 2$. Due to the long computation time for the SIMEX procedure, we simulate 250 Monte Carlo replicates of 16 parameter settings. The parameter settings are factorial combinations of $\theta_{11} = (0.7, 0.9)$, $\theta_{00} = (0.8, 0.95)$, validation sizes $(50, 50), (200, 200)$, $\pi = (0.00001, 0.08001, \ldots, 0.96001)$, and Var$(Y|X) = (1, 10)$. The main study size is 1000, true intercept is 10 and true slope is 20.

Figure 8 summarizes the coverage results. Since the coverage for the naive confidence is close to 0, we exclude it from this figure. Also, the case of known misclassification model was dropped for Fieller's and the delta method, since each of those results were very similar to the corresponding case of an estimated misclassifiction model that is treated approximately. When we account for the variability of the estimated misclassification model, Fieller's method can give us coverage very close to nominal. The coverage for the SIMEX method depends on $\pi = P(X = 1), \theta_{00}, \theta_{11}$ and variance of $Y|X$. The coverage of the delta method is lower than nominal level when π is close to 0 or 1. If we assume the estimated misclassification model to be known, the coverage is below the nominal level.

Figure 9 shows length of 95% confidence intervals for the above simulation. When π is close to 0, the length of the delta method's interval is too big, so we do not include that case in the figure. We can see that in general SIMEX can give a smaller confidence interval.

If we know the misclassification matrix, Fieller's method can produce a better confidence interval than the SIMEX method. If the misclassification matrix is estimated, Fieller's method that accounts for the variability of θ_{ij} can achieve the nominal level. In general, the SIMEX method's confidence interval is shorter, but the coverage is low. The coverage of the delta's method is too small when π is close to zero or one. It is very time consuming to use the SIMEX method to do correction. Our simulation took us 10 days. Not surprisingly, if we do not do any correction, the coverage of the naive confidence interval is very close to zero.

5.4 Score Approach and Regression Calibration

We use the score approach in Section 4.3, and the regression calibration approach in Section 4.4.2 to correct the coefficients of regression with misclassified covariates. In this section, we will use simulation to assess the performance of those methods.

We will use binary covariates with $\pi = 0.2, 0.5$, and $\theta_{11} = 0.9, \theta_{00} = 0.75$ as high level of misclassified, $\theta_{11} = 0.9, \theta_{00} = 0.85$ as medium level of misclassified, and $\theta_{11} = 0.9, \theta_{00} = 0.95$ as low level of misclassified. We also use the relationship between misclassification /reclassification models: $\mathbf{Q} = \mathbf{D}_{\underline{\pi}}\mathbf{P}^{\mathsf{T}}\mathbf{D}_{\underline{\lambda}}^{-1}$ to find the corresponding reclassification parameters. Table 5 summarizes the parameters we use.

π	Level	θ_{00}	θ_{11}	γ_{00}	γ_{11}
0.2	H	0.75	0.9	0.968	0.474
	M	0.85	0.9	0.971	0.6
	L	0.95	0.9	0.974	0.818
0.5	H	0.75	0.9	0.882	0.783
	M	0.85	0.9	0.895	0.857
	L	0.95	0.9	0.905	0.947

Table 5. Parameter Settings for the Simulation of Regression

The model for the simulated data is :

$$y|\underline{X} = 0 \sim N(\beta_0, \sigma^2), y|\underline{X} = 1 \sim N(\beta_1, \sigma^2) \text{ with } \beta_0 = 3, \beta_1 = -4 \text{ and } \sigma^2 = 2.$$

The main study sizes are 50, 200, 1000, the validation sizes are 50, 100. We also include the case when the parameters are known. For each unique parameter combina-

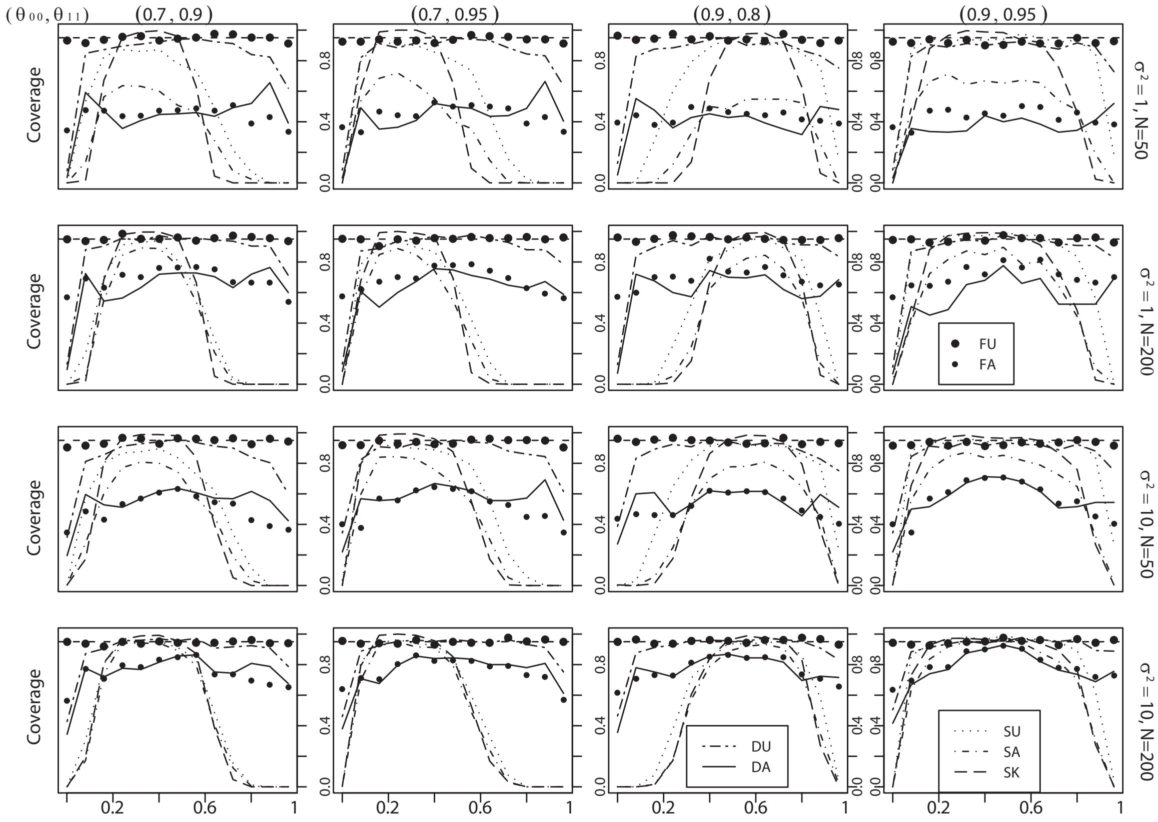


Figure 8. Coverage for Interval Estimates of Regression Slopes: this figure compares the average coverage for slope of simple regression under different settings and using different methods. ′FU′ is Fieller′s method accounting for the estimated MC model;′FA′ is Fieller′s method with the MC model estimated and assumed known; ′SU′ is SIMEX method accounting for the MC model;′SA′ SIMEX method with the MC model estimated and assumed known; ′SK′ is SIMEX method with known MC matrix;′DU′ is the delta′s method accounting the the estimated MC model;′DA′ is the delta′s method with the MC model estimated and assumed known. The first two rows have variance 1 and the last 2 rows variance is 10. Validation size for first and third row is 50, the other 2 rows is 200.

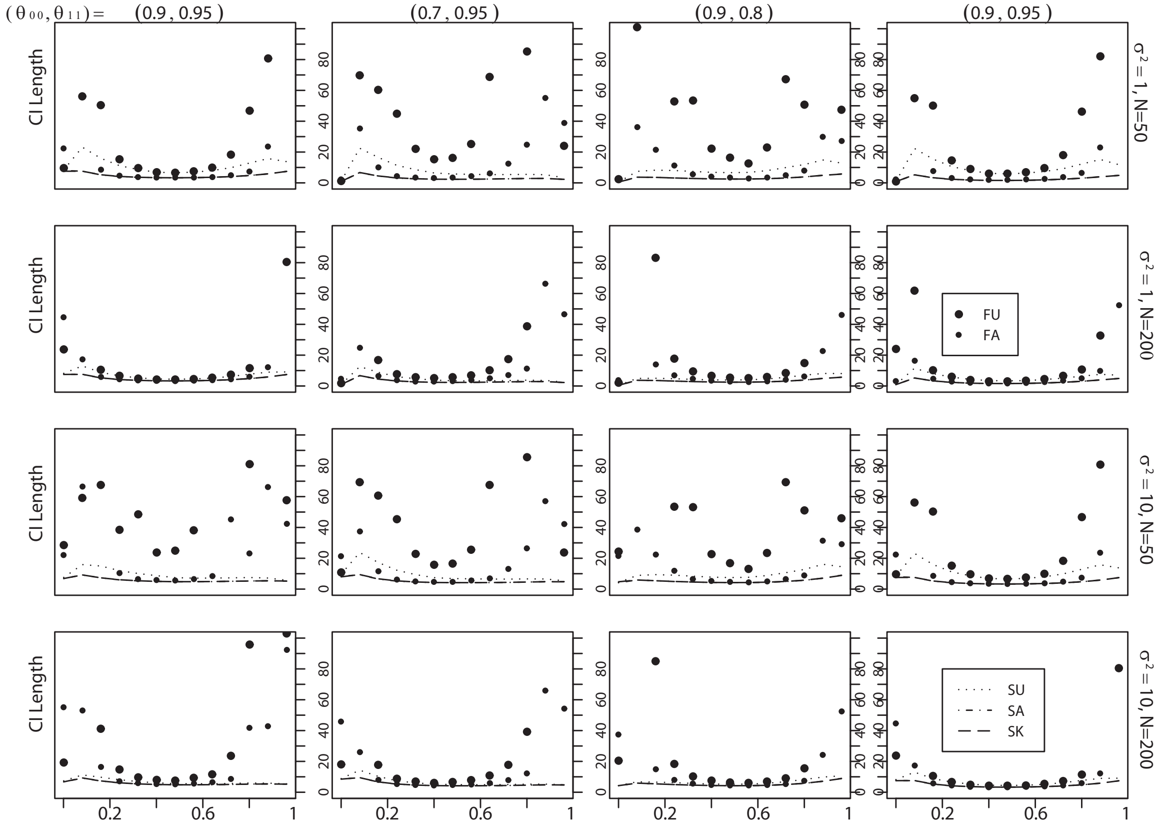


Figure 9. Length of Interval Estimates of Regression Slopes: this figure compares the the length of confidence interval for slope of simple regression under different settings and using different methods. 'FU' is Fieller's method accounting for the estimated MC model;'FA' is Fieller's method with the MC model estimated and assumed known; 'SU' is SIMEX method accounting for the MC model;'SA' SIMEX method with the MC model estimated and assumed known; 'SK' is SIMEX method with known MC matrix. The first two rows have variance 1 and the last 2 rows variance is 10. Validation size for first and third row is 50, the other 2 rows is 200.

tion, we replicate the experiment 4000 times. Tables 6, 7, 8 and 9 contain summaries of the results of the simulation for score corrected estimators. The entry $\widehat{\beta}$ denotes the median of the 4000 observed score corrected estimators, $\sigma\widehat{\beta}$ denotes the observed standard deviation of the 4000 $\widehat{\beta}$'s, and $\widehat{\sigma}_4(\widehat{\beta})$, and $\widehat{\sigma}_5(\widehat{\beta})$ denote the mean of the 4000 standard deviations using Equations (4) and (5) on Nakamura (1990) respectively.

From Table 6, we see if the misclassification parameters are at the high level of misclassification, $n = 50$ and $\pi = 0.2$, then the estimated and the observed standard deviation do not close to each other. This is due to the fact that many (21$\%$) of the $\widehat{\sigma}^2$'s are negative in that case. Since the error of regression with misclassified covariates is not additive, we can not just use the usual methods (see Bock and Peterson,1975, Amemiya,1985) to correct for this problem. Equation (5) of Nakamura (1990) is less biased than Equation (4) of Nakamura (1990) for estimating the standard deviation of the slope when there is a low level of misclassification.

From Tables 7, and 8 we can see that, as might be expected, increasing of the main study size and validation size makes the coefficient estimators closer to the true estimators on average. Note that though that increasing the main study sample size alone does not make inference better. Increasing the validation size is also necessarily to increase the efficiency of inference.

Table 9 is the median of the 4000 variance estimators using the score function approach. As mentioned before, even though there is some chance that the variance estimate will be negative, the estimate of the variance is still close to true variance on average as the validation size and the main study size both increase.

Tables 10 and 11 are regression calibration results using the same simulation data. From the tables, we can see that even when main study and validation size are small, the estimated and the observed standard deviation are very close to each other. Also the regression variance is still negative fairly often, but in this case, inference is unaffected because the regression variance is not necessary in this case. See Section 4.4.2.

π	level	n	$\widehat{\beta}_0$	$\sigma(\widehat{\beta}_0)$	$\widehat{\sigma}_4(\widehat{\beta}_0)$	$\widehat{\sigma}_5(\widehat{\beta}_0)$	$\widehat{\beta}_1$	$\widehat{\sigma}(\widehat{\beta}_1)$	$\widehat{\sigma}_4(\widehat{\beta}_1)$	$\widehat{\sigma}_5(\widehat{\beta}_1)$
0.2	H	50	3.013	0.391	0.37	0.405	-3.919	10.883	16.881	17.192
		200	3.005	0.19	0.186	0.183	-3.987	2.108	1.899	1.923
		1000	3.001	0.083	0.084	0.082	-4.016	0.727	0.705	0.708
	M	50	3.005	0.345	0.33	0.348	-4.029	7.407	7.882	8.151
		200	3.003	0.17	0.166	0.163	-3.984	1.341	1.219	1.236
		1000	3.001	0.075	0.075	0.074	-4.002	0.497	0.501	0.502
	L	50	3.004	0.312	0.3	0.294	-4.112	2.052	1.682	1.842
		200	3.002	0.154	0.151	0.149	-4.018	0.662	0.638	0.651
		1000	3.001	0.068	0.068	0.068	-4.009	0.28	0.281	0.282
0.5	H	50	3.049	0.781	0.757	0.768	-4.051	1.159	1.101	1.123
		200	3.02	0.381	0.372	0.373	-4.005	0.508	0.51	0.512
		1000	3.001	0.169	0.166	0.166	-3.998	0.227	0.225	0.225
	M	50	3.046	0.674	0.656	0.666	-4.044	0.791	0.767	0.785
		200	3.018	0.334	0.326	0.327	-4.006	0.377	0.375	0.377
		1000	3.001	0.148	0.146	0.146	-4	0.17	0.167	0.167
	L	50	3.04	0.597	0.581	0.594	-4.051	0.485	0.457	0.472
		200	3.013	0.299	0.291	0.293	-4.004	0.241	0.235	0.236
		1000	3	0.132	0.13	0.13	-4.001	0.106	0.106	0.106

Table 6. Score Function Approach with Known Misclassification Parameters

5.5 Mixture Method and Regression Calibration Method

In this section, we will use simulation to compare the mixture method and the regression calibration method that we developed in Section 4.4 for regression with misclassified covariates and a known reclassification model. We use the same parameter settings and model assumptions as in Section 5.4, and the Monte Carlo sample size is 4000.

Table 12 and Table 13 show the results of simulation. The entry $\widehat{\beta}$ in Table 12 denotes the median of the 4000 observed corrected estimators, $\sigma(\widehat{\beta})$ denotes the observed standard deviation of the 4000 values of $\widehat{\beta}$, and $\widehat{\sigma}(\widehat{\beta})$ denotes the mean of the 4000 estimated standard deviations for each $\widehat{\beta}$. Some of the observed standard deviations and the estimated standard deviation are somewhat different. If we use a robust estimator of standard deviation median (median of absolute deviation, MAD), they are very close to each other though. This is not shown in the table. Table 13 contains the median of

π	level	n	$\widehat{\beta}_0$	$\sigma(\widehat{\beta}_0)$	$\widehat{\sigma}_4(\widehat{\beta}_0)$	$\widehat{\sigma}_5(\widehat{\beta}_0)$	$\widehat{\beta}_1$	$\widehat{\sigma}(\widehat{\beta}_1)$	$\widehat{\sigma}_4(\widehat{\beta}_1)$	$\widehat{\sigma}_5(\widehat{\beta}_1)$
0.2	H	50	3.023	0.417	0.387	0.433	-3.544	7.227	9.838	10.038
		200	3.003	0.225	0.202	0.206	-3.768	13.76	12.509	12.641
		1000	2.996	0.147	0.11	0.106	-3.914	19.401	10.023	10.078
	M	50	3.009	0.367	0.342	0.373	-3.846	6.387	6.775	7.015
		200	3.001	0.197	0.179	0.179	-3.97	10.197	6.358	6.46
		1000	2.998	0.127	0.096	0.093	-3.889	17.356	6.566	6.602
	L	50	3.004	0.329	0.308	0.314	-4.036	3.03	2.193	2.363
		200	3	0.176	0.162	0.159	-3.903	2.637	1.098	1.122
		1000	2.997	0.113	0.086	0.084	-3.916	1.601	0.592	0.595
0.5	H	50	3.059	1.158	0.905	0.928	-4.045	2.041	1.442	1.474
		200	2.984	0.711	0.49	0.49	-4.024	1.132	0.67	0.676
		1000	2.981	0.566	0.331	0.33	-3.977	0.887	0.4	0.401
	M	50	3.06	0.923	0.756	0.772	-4.026	1.097	0.871	0.892
		200	2.992	0.592	0.42	0.419	-4.019	0.711	0.459	0.462
		1000	2.982	0.476	0.283	0.282	-3.972	0.562	0.278	0.278
	L	50	3.046	0.78	0.656	0.673	-4.021	0.577	0.491	0.506
		200	2.994	0.511	0.369	0.37	-3.999	0.376	0.268	0.268
		1000	2.988	0.411	0.247	0.247	-3.982	0.293	0.155	0.154

Table 7. Score Function Approach with Validation Size = 50

variance estimates for both methods and for each parameter setting. We should note that even when the reclassification model is known, there is still a small chance that the variance estimate from regression calibration will be negative, but the variance estimate from the mixture method is always non-negative.

5.6 Corrected Score Estimator for Regression with Misclassified Covariates and Perfectly Measured Covariates

In this section, we will simulate data from $Z|X = 0 \sim N(\mu_0, \sigma_Z^2), Z|\ X = 1 \sim N(\mu_1, \sigma_Z^2)$, and $Y|X, Z \sim N\left\{(1 - X\beta_0) + X\beta_1 + Z\beta_Z, \sigma_Y^2\right\}$. From the assumed model, we have

$$\begin{aligned}\text{Var}(Z) &= E\{\ \text{Var}(Z|X)\} +\ \text{Var}\{E(Z|\underline{X})\}\\ &= \sigma_Z^2 +\ \text{Var}\{\mu_0(1-X) + \mu_1 X\},\end{aligned}$$

π	level	n	$\widehat{\beta}_0$	$\sigma(\widehat{\beta}_0)$	$\widehat{\sigma}_4(\widehat{\beta}_0)$	$\widehat{\sigma}_5(\widehat{\beta}_0)$	$\widehat{\beta}_1$	$\widehat{\sigma}(\widehat{\beta}_1)$	$\widehat{\sigma}_4(\widehat{\beta}_1)$	$\widehat{\sigma}_5(\widehat{\beta}_1)$
0.2	H	50	3.036	0.398	0.377	0.421	-3.822	10.953	16.695	17.005
		200	3.003	0.208	0.194	0.196	-3.917	8.507	6.249	6.331
		1000	3	0.118	0.097	0.094	-3.977	10.088	3.633	3.661
	M	50	3.027	0.352	0.335	0.361	-3.992	8.335	8.894	9.173
		200	3.002	0.185	0.173	0.172	-3.976	4.528	2.316	2.364
		1000	2.999	0.104	0.086	0.083	-3.961	2.669	0.977	0.988
	L	50	3.02	0.315	0.303	0.302	-4.083	3.134	2.187	2.363
		200	3.005	0.166	0.157	0.154	-3.985	1.173	0.774	0.792
		1000	2.998	0.093	0.077	0.076	-3.983	0.863	0.44	0.442
0.5	H	50	3.04	0.92	0.825	0.846	-3.99	1.379	1.186	1.208
		200	3.013	0.524	0.431	0.43	-4.017	0.759	0.577	0.581
		1000	2.985	0.387	0.256	0.255	-4.001	0.569	0.315	0.315
	M	50	3.043	0.778	0.706	0.719	-4.009	0.925	0.805	0.825
		200	3.013	0.452	0.374	0.375	-3.998	0.524	0.412	0.415
		1000	2.985	0.332	0.221	0.221	-3.991	0.394	0.226	0.226
	L	50	3.041	0.681	0.622	0.639	-4.023	0.526	0.473	0.487
		200	3.01	0.396	0.331	0.333	-4.009	0.296	0.251	0.251
		1000	2.99	0.291	0.195	0.195	-3.986	0.218	0.132	0.132

Table 8. Score Function Approach with Validation Size = 100

$$\begin{aligned}
&= \sigma_Z^2 + (\mu_1 - \mu_0)^2 \pi(1-\pi) \\
\text{Var}(X) &= \pi(1-\pi), \text{ and} \\
\text{Cov}(Z, X) &= \text{Cov}\left\{E(Z|X), E(X|X)\right\} + E\left\{\text{Cov}(Z, X|X)\right\} \\
&= (\mu_0 - \mu_1)\pi(1-\pi).
\end{aligned}$$

As a result, we have

$$\text{Cor}(Z, X) = \frac{sign(\mu_0 - \mu_1)}{(1 + \dfrac{\sigma_Z^2}{(\mu_0 - \mu_1)^2 \pi(1-\pi)})^{1/2}},$$

a function of $\mu_0 - \mu_1, \pi$ and σ_Z^2.

In this simulation, we will use the same misclassification parameters as before, i.e. $\theta_{11} = 0.9, \theta_{00} = 0.75$ for high level of misclassification, $\theta_{11} = 0.9, \theta_{00} = 0.85$ for medium of misclassification and $\theta_{11} = 0.9, \theta_{00} = 0.95$ for low level of misclassification. We use $\pi = 0.4$, main study sizes 100,200, and 1000 and validation sizes 50 and 100. We also include the case with known misclassification parameters. The regression parameters

level		0.2			0.5		
	N n	∞	50	100	∞	50	100
H	50	1.758	2.233	1.747	1.392	1.081	1.39
	200	1.954	2.145	2.013	1.856	1.683	1.793
	1000	1.947	2.053	1.949	1.983	1.923	1.961
M	50	1.667	1.88	1.63	1.518	1.373	1.525
	200	1.952	1.94	1.948	1.824	1.731	1.871
	1000	1.973	2.07	1.986	1.991	1.972	1.982
L	50	1.63	1.745	1.564	1.654	1.599	1.639
	200	1.93	2.035	1.971	1.884	1.863	1.914
	1000	1.973	2.068	2.02	1.977	2.011	2.026

Table 9. Variance Estimate from Score Function Approach

are $\beta_0 = -3, \beta_1 = 4, \beta_z = 2, \sigma^2 = 1, \mu_0 = 1, \mu_1 = 9$ and $\sigma_Z^2 = 100$. This results in $\mathrm{Cor}(Z, X) = 0.365$. The Monte Carlo sample size is 4000.

Table 14 contains the simulation results. Since Equation (5) of Nakamura (1990) produces too many negative definite matrix estimates in this case, we use Equation (4) of Nakamura (1990) instead. The entry $\widehat{\beta}$ in Table 14 denotes the median of the 4000 observed corrected estimators, $\sigma(\widehat{\beta})$ denotes the observed standard deviation of the 4000 values of $\widehat{\beta}$, $\widehat{\sigma}(\widehat{\beta})$ denotes the mean of the 4000 estimated standard deviations for each $\widehat{\beta}$, and $\widehat{\sigma}^2$ denotes the median of the 4000 observed corrected estimates of the variance. From this table, we can see that, on the average, score corrected coefficients estimators are very close to the true coefficients for any misclassification level, any main study size, and any validation size. But the variance estimator is not as impressive unless there is a medium level of misclassification with a known misclassification model and $n = 1000$, or low level of misclassification, with a known misclassification model or large validation size, $n = 1000$. The estimated and observed standard deviations of $\widehat{\beta}$ are very close to each other when the misclassification is known. When validation size is 100, as main study size increases, the inference worsens. We suspect this is because some of the estimates of $\widehat{\sigma}^2$ are negative.

			H			M			L		
n	N	π	$\widehat{\beta_0}$	$\sigma(\widehat{\beta_0})$	$\widehat{\sigma}(\widehat{\beta_0})$	$\widehat{\beta_0}$	$\sigma(\widehat{\beta_0})$	$\widehat{\sigma}(\widehat{\beta_0})$	$\widehat{\beta_0}$	$\sigma(\widehat{\beta_0})$	$\widehat{\sigma}(\widehat{\beta_0})$
50	∞	0.2	3.001	0.379	0.533	3.003	0.337	0.45	3.002	0.305	0.349
	50		3.004	0.43	0.584	2.996	0.383	0.493	2.997	0.343	0.389
	100		3.02	0.404	0.557	3.013	0.359	0.47	3.008	0.32	0.368
	∞	0.5	3.04	0.704	0.773	3.036	0.618	0.637	3.028	0.551	0.501
	50		3.043	0.814	0.883	3.024	0.722	0.741	3.015	0.65	0.607
	100		3.027	0.749	0.831	3.039	0.663	0.692	3.025	0.592	0.559
200	∞	0.2	3.003	0.186	0.266	3.002	0.166	0.226	3.001	0.15	0.176
	50		2.99	0.282	0.341	2.99	0.247	0.29	2.988	0.223	0.238
	100		2.999	0.238	0.304	2.997	0.214	0.261	2.999	0.192	0.211
	∞	0.5	3.014	0.348	0.386	3.011	0.307	0.319	3.006	0.276	0.253
	50		3.017	0.545	0.564	3.016	0.487	0.487	3.005	0.436	0.418
	100		3.001	0.456	0.478	3.004	0.406	0.409	3.003	0.363	0.344
10^3	∞	0.2	3.002	0.082	0.119	3	0.073	0.101	3	0.066	0.079
	50		2.98	0.232	0.235	2.983	0.204	0.203	2.982	0.183	0.173
	100		2.99	0.168	0.184	2.989	0.149	0.161	2.989	0.135	0.138
	∞	0.5	3.002	0.157	0.172	3.002	0.138	0.143	3.001	0.123	0.113
	50		2.98	0.445	0.434	2.978	0.397	0.384	2.987	0.357	0.344
	100		2.983	0.312	0.324	2.984	0.283	0.287	2.98	0.259	0.255
n	N	π	$\widehat{\beta_0}$	$\sigma(\widehat{\beta_1})$	$\widehat{\sigma}(\widehat{\beta_1})$	$\widehat{\beta_1}$	$\sigma(\widehat{\beta_1})$	$\widehat{\sigma}(\widehat{\beta_1})$	$\widehat{\beta_1}$	$\sigma(\widehat{\beta_1})$	$\widehat{\sigma}(\widehat{\beta_1})$
50	∞	0.2	-3.983	1.998	1.51	-3.994	1.714	1.16	-4.075	1.198	0.803
	50		-3.878	2.432	1.996	-3.979	1.95	1.488	-4.033	1.277	0.951
	100		-3.987	2.255	1.753	-4.02	1.873	1.325	-4.106	1.25	0.869
	∞	0.5	-4.023	0.825	0.77	-4.012	0.672	0.636	-4.043	0.464	0.504
	50		-4.03	1.025	0.983	-4.001	0.814	0.777	-4.009	0.523	0.563
	100		-4.022	0.933	0.877	-4.014	0.749	0.709	-4.019	0.502	0.535
200	∞	0.2	-3.978	0.978	0.753	-3.98	0.826	0.578	-4.012	0.581	0.398
	50		-4.011	1.613	1.437	-3.991	1.224	1.054	-3.997	0.76	0.628
	100		-3.981	1.347	1.12	-3.993	1.068	0.842	-4.013	0.678	0.524
	∞	0.5	-4.002	0.409	0.385	-4.004	0.333	0.319	-4.005	0.234	0.253
	50		-3.996	0.717	0.702	-3.996	0.548	0.539	-3.989	0.332	0.347
	100		-3.994	0.582	0.557	-4.001	0.449	0.439	-4.002	0.288	0.304
10^3	∞	0.2	-4	0.437	0.337	-3.999	0.364	0.259	-4.003	0.251	0.178
	50		-4.029	1.358	1.258	-4.039	1.011	0.923	-3.999	0.563	0.514
	100		-4.008	0.961	0.886	-3.997	0.719	0.658	-3.988	0.426	0.38
	∞	0.5	-3.998	0.186	0.172	-4	0.151	0.143	-4.002	0.103	0.113
	50		-3.989	0.619	0.601	-3.978	0.461	0.447	-3.981	0.262	0.258
	100		-3.992	0.445	0.433	-3.99	0.34	0.33	-3.996	0.197	0.201

Table 10. Regression Calibration Results

level		0.2			0.5		
	N n	∞	50	100	∞	50	100
H	50	1.988	2.123	1.909	1.895	1.937	2.033
	200	1.977	2.091	2.059	1.955	2.048	2.008
	1000	1.969	2.074	2.051	2.003	2.117	2.135
M	50	1.934	2.056	1.865	1.91	2.024	2.059
	200	1.977	2.074	2.037	1.933	2.057	2.029
	1000	1.984	2.064	2.066	2.012	2.209	2.085
L	50	1.898	2.011	1.817	1.935	2.058	2.012
	200	1.996	2.118	2.032	1.958	2.087	1.982
	1000	1.987	2.107	2.058	1.994	2.104	2.061

Table 11. Variance Estimate for Regression Calibration Method

Table 12. Comparison of standard deviation of mixture estimation and regression calibration estimators

			0.2						0.5					
	n	MD	$\widehat{\beta}_0$	$\sigma(\widehat{\beta}_0)$	$\widehat{\sigma}(\widehat{\beta}_0)$	$\widehat{\beta}_1$	$\sigma(\widehat{\beta}_1)$	$\widehat{\sigma}(\widehat{\beta}_1)$	$\widehat{\beta}_0$	$\sigma(\widehat{\beta}_0)$	$\widehat{\sigma}(\widehat{\beta}_0)$	$\widehat{\beta}_1$	$\sigma(\widehat{\beta}_1)$	$\widehat{\sigma}(\widehat{\beta}_1)$
H	50	MIX	3.002	0.232	0.16	-4.009	0.58	0.381	3	0.313	0.211	-3.999	0.312	0.21
		RC	3.012	0.374	0.534	-4.07	1.997	1.511	3.031	0.705	0.777	-4.006	0.837	0.77
	200	MIX	3	0.113	0.08	-4.005	0.238	0.169	2.998	0.152	0.103	-3.998	0.151	0.103
		RC	3.004	0.183	0.266	-4.013	0.962	0.753	3.006	0.346	0.385	-4.008	0.417	0.385
	10^3	MIX	3	0.051	0.036	-3.998	0.104	0.074	2.997	0.066	0.046	-3.999	0.066	0.046
		RC	3.003	0.082	0.119	-4.003	0.428	0.337	3	0.155	0.172	-3.998	0.182	0.172
M	50	MIX	3.001	0.229	0.16	-4.012	0.525	0.38	2.999	0.3	0.21	-4.001	0.297	0.209
		RC	3.009	0.332	0.451	-4.081	1.685	1.157	3.029	0.614	0.636	-4.027	0.676	0.635
	200	MIX	3.001	0.113	0.08	-4.004	0.235	0.168	2.997	0.151	0.103	-3.998	0.15	0.103
		RC	3.005	0.165	0.226	-4.034	0.821	0.578	3.011	0.305	0.318	-4.011	0.338	0.318
	10^3	MIX	3	0.051	0.036	-3.999	0.103	0.074	2.997	0.067	0.046	-3.999	0.067	0.046
		RC	3.002	0.073	0.101	-3.997	0.362	0.259	3	0.136	0.143	-4.001	0.148	0.143
L	50	MIX	3.002	0.227	0.16	-4.008	0.491	0.378	3.002	0.296	0.209	-4.001	0.292	0.208
		RC	3.004	0.301	0.349	-4.099	1.165	0.798	3.027	0.553	0.502	-4.028	0.476	0.503
	200	MIX	3.001	0.113	0.08	-4.003	0.232	0.166	2.997	0.148	0.102	-4	0.148	0.102
		RC	3.003	0.148	0.176	-4.042	0.56	0.398	3.003	0.274	0.252	-4.007	0.233	0.252
	10^3	MIX	2.999	0.051	0.036	-3.998	0.101	0.073	2.998	0.065	0.046	-3.999	0.065	0.046
		RC	3.001	0.067	0.079	-4.001	0.252	0.178	3.001	0.122	0.113	-4.001	0.102	0.113

		MIX			RC		
level	n π	50	200	1000	50	200	1000
H	0.2	1.9	1.97	1.996	1.785	1.925	1.979
	0.5	1.893	1.968	1.998	1.882	1.916	1.992
M	0.2	1.898	1.97	1.996	1.817	1.926	1.979
	0.5	1.894	1.968	2	1.905	1.927	2
L	0.2	1.902	1.97	1.996	1.843	1.941	1.999
	0.5	1.892	1.967	1.999	1.916	1.951	1.995

Table 13. Variance Estimate From Mixture Method and Regression Calibration

N	L	n	$\widehat{\beta}_0$	$\sigma(\widehat{\beta}_0)$	$\widehat{\sigma}(\widehat{\beta}_0)$	$\widehat{\beta}_1$	$\sigma(\widehat{\beta}_1)$	$\widehat{\sigma}(\widehat{\beta}_1)$	$\widehat{\beta}_z$	$\sigma(\widehat{\beta}_z)$	$\widehat{\sigma}(\widehat{\beta}_z)$	$\widehat{\sigma}^2$
∞	H	100	-3.060	0.407	0.405	4.167	1.696	1.204	1.996	0.070	0.053	0.409
		200	-3.022	0.269	0.274	4.042	0.997	0.856	1.999	0.042	0.037	0.8
		10^3	-3.003	0.118	0.117	3.992	0.403	0.395	2.001	0.018	0.017	1.002
	M	100	-3.042	0.338	0.343	4.071	1.006	0.917	1.998	0.045	0.041	0.747
		200	-3.022	0.227	0.234	4.066	0.674	0.648	1.998	0.031	0.029	0.777
		10^3	-3.002	0.101	0.102	4.013	0.288	0.290	1.999	0.013	0.013	0.961
	L	100	-3.019	0.294	0.296	4.077	0.560	0.557	1.997	0.030	0.029	0.783
		200	-3.008	0.203	0.206	4.014	0.388	0.388	1.999	0.021	0.020	0.922
		10^3	-3.002	0.092	0.090	4.004	0.172	0.171	2.000	0.009	0.009	0.991
50	H	100	-3.069	0.517	0.777	4.163	3.216	2.679	1.994	0.110	0.103	0.327
		200	-3.049	0.410	0.620	4.074	1.970	2.223	1.997	0.064	0.079	0.607
		10^3	-3.011	0.324	0.501	4.058	1.548	1.845	1.998	0.043	0.063	0.754
	M	100	-3.065	0.436	0.551	4.193	1.503	1.641	1.996	0.055	0.066	0.358
		200	-3.033	0.342	0.450	4.099	1.198	1.386	1.997	0.041	0.054	0.655
		10^3	-3.025	0.278	0.359	4.063	0.991	1.126	1.998	0.030	0.042	0.735
	L	100	-3.048	0.368	0.417	4.125	0.744	0.841	1.996	0.034	0.043	0.646
		200	-3.027	0.293	0.332	4.117	0.629	0.673	1.997	0.026	0.034	0.688
		10^3	-3.016	0.244	0.245	4.052	0.517	0.502	1.999	0.019	0.025	0.857
100	H	100	-3.073	0.462	0.412	4.189	2.483	1.169	1.995	0.091	0.051	0.365
		200	-3.043	0.342	0.279	4.070	1.543	0.818	2.000	0.053	0.036	0.701
		10^3	-3.010	0.240	0.121	4.051	1.051	0.371	1.999	0.032	0.016	0.840
	M	100	-3.052	0.385	0.346	4.115	1.290	0.904	1.997	0.051	0.041	0.631
		200	-3.025	0.296	0.239	4.078	0.971	0.635	1.998	0.036	0.029	0.710
		10^3	-3.012	0.208	0.104	4.043	0.729	0.283	1.999	0.023	0.013	0.882
	L	100	-3.030	0.332	0.299	4.102	0.662	0.561	1.997	0.032	0.029	0.683
		200	-3.028	0.255	0.207	4.077	0.508	0.389	1.997	0.023	0.020	0.751
		10^3	-3.011	0.182	0.091	4.024	0.398	0.172	1.999	0.015	0.009	0.930

Table 14. Simulation Results for Regression with Misclassified and Perfectly Measured Covariates

CHAPTER 6

CONCLUSION

In this dissertation, we show how to use external validation data to correct estimates of a proportion when the main study data are misclassified, and coefficients of linear regression with misclassified covariates (both with and without perfectly measured covariates).

We introduce two estimators for a proportion, $\widehat{\pi}_{Corrected,PI}$ with reduced bias, and $\widehat{\pi}_{pc}$ with smaller mean square error. Simulation suggests that $\widehat{\pi}_{pc}$ performs quite well in general. We use Fieller's method and optimization techniques to find a confidence interval for proportion of interest. Simulation shows that Fieller's method performs better than the optimization techniques.

For regression with misclassified covariates, we can use a score function approach, regression calibration, or a mixture model method. If a misclassification model is available, known or estimated, we can use the score function approach. If a reclassification model is available, known or estimated, we can use regression calibration or the mixture model method.

In the process of writing this dissertation, we have observed that there are some interesting topics that can be explored further as future research.

- In the simulation, sometimes $\widehat{\sigma}^2$ is negative after correction. It is not clear how to do correction in this case. For a regression model with misclassified covariates, the error model is not additive. In general, under the additive error model, there are methods that can apply to correct in the case when the covariance estimate is non-positive definite. Perhaps these methods can be adapted to our setting.

- We used mixture methods in linear regression with misclassified covariates, and a reclassification model. It would be nice to generalize this approach to address a wider variety of regression models.
- When we use the mixture model, we require information about the reclassification model to use the EM algorithm to do the correction. It is unclear to us under which conditions we can recover the reclassification model from the EM algorithm without a known/estimated reclassification model.
- For a regression model with misclassified covariates and a konwn misclassification/reclassification modelv, the corrected score estimator and the regression calibration estimator are all unbiased. This is not so in the case of regression models with misclassified and perfectly measured covariates. It is not clear that we can find an unbiased estimator for a regression model with misclassified and perfectly measured covariates.
- We would like to extend our methods to generalized linear models.
- When π is of interest, the reclassification model is always more efficient. When one has a misclassification model only, it would be interesting, to evaluate the following procedure:

  1. Use θ_{00}, θ_{11} to evaluate π.
  2. Use $\theta_{00}, \theta_{11}, \widehat{\pi}$ to estimate a reclassification model: γ_{00}, γ_{11}.
  3. Use γ_{00}, γ_{11} to estimate π.

  Steps 2 and 3 could be iterated also. We wonder if, in this way, we could produce a more efficient estimator.

APPENDIX A

Different Types of Proportional Confidence Intervals

Let $\widehat{p}$ denote the sample proportion of a binomial distribution of sample size m for a binomial parameter p. A $100(1-\alpha)\%$ Wald confidence interval for p is

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{m}}$$

where z_α is the $1-\alpha$ quartile of the standard normal distribution.

A $100(1-\alpha)\%$ exact confidence interval for p is

$$\left[\frac{1}{1+\dfrac{m-x+1}{x}F_{2(m-x+1),2x,\alpha/2}}, \frac{\dfrac{x+1}{m-x}F_{2(x+1),2(m-x),\alpha/2}}{1+\dfrac{x+1}{m-x}F_{2(x+1),2(m-x),\alpha/2}}\right]$$

where x is the total number in the sample equal to 1 and $F_{\nu_1,\nu_2,\alpha}$ is the upper α percentile from F distribution with degree ν_1, ν_2. This is an exact interval, not the result of a large sample. It is based on the result of Casella and Berger's Exercise 9.2.1.

A $100(1-\alpha)\%$ score confidence interval for p is

$$\frac{\widehat{p}+\dfrac{z^2_{\alpha/2}}{2m} \pm z_{\alpha/2}\sqrt{[\widehat{p}(1-\widehat{p})+\dfrac{z^2_{\alpha/2}}{4m}]/m}}{1+\frac{z^2_{\alpha/2}}{m}}.$$

The endpoints of the score confidence interval are the solutions p_0 to the equations $\frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)/m}} = \pm z_{\alpha/2}$. The score confidence interval is the inversion of score test for p. See Agresti and Coull (1998). The exact confidence interval will have average coverage too large, Wald confidence interval will have average coverage too small and score confidence interval will have average coverage close to the nominal confidence interval. See Agresti and Coull (1998).

APPENDIX B

Algorithm to Find Upper/Lower Bound for π

In this appendix, we will give an algorithm of using the results of Section 3.1 to find the upper and lower bound for π. f is defined in Section 3.1. Let A be the boundary values (eight of them) of f on this region. Here are the steps that we will follow :

1. If $(L_{00} + L_{11} - 1)(U_{00} + U_{11} - 1) > 0$:

- if length$(A \geq 0) = 0$, then min $=$ max $= 0$
- else min $=$ min$($min$(A \geq 0), 1)$, max $=$ min$($max$(A \geq 0), 1)$
- the interval is $[\text{min}, \text{max}]$.

2. If $(L_{00} + L_{11} - 1)(U_{00} + U_{11} - 1) < 0$ and $(U_\lambda + U_{00} - 1) * (L_\lambda + L_{00} - 1) \geq 0)$ and A has no valid value:

- if length$(A \geq 0) = 0$ min $= 0$ else min=min(min$(A \geq 0), 1)$
- max=1
- the interval is $\{0\} \cup [\text{min}, \text{max}]$

3. If $(L_{00} + L_{11} - 1)(U_{00} + U_{11} - 1) < 0$ and $(U_\lambda + U_{00} - 1) * (L_\lambda + L_{00} - 1) < 0)$ and A has no invalid value:

- relative minimum value will be greater than or equal to 1, the relative maximum value will be less than or equal to 1

- max=relative maximum value
- the interval is $[0, \text{max}] \cup \{0\}$

4. If $U_{00} + U_{11} - 1 = 0$ and $L_\lambda + L_{00} - 1 \geq 0$:

- min=0, max=0, length=0
- the interval is $\{0\}$

5. If $U_{00} + U_{11} - 1 = 0$ and $U_\lambda + U_{00} - 1 \leq 0$:

- min=0 if no relative minimum else min=min(max(relative min,0),1)
- max=1
- the interval is $[\text{min}, \text{max}]$

6. if $U_{00} + U_{11} - 1 = 0$ and $(U_\lambda + U_{00} - 1)(L_\lambda + L_{00} - 1) < 0$:

- min=0, max=1,length=1
- the interval is $[0, 1]$

7. if $L_{00} + L_{11} - 1 = 0$ and $U_\lambda + U_{00} - 1 \leq 0$:

- min=0, max=0, length=0
- the interval is $\{0\}$

8. if $L_{00} + L_{11} - 1 = 0$ and $L_\lambda + L_{00} - 1 \geq 0$:

- min=0 if no relative minimum else min=min(max(relative min,0),1)
- max=1
- the interval is $[\text{min}, \text{max}]$

9. if $L_{00} + L_{11} - 1 = 0$ and $(U_\lambda + U_{00} - 1)(L_\lambda + L_{00} - 1) < 0$:

- min=0, max=1,length=1
- the interval is $[0, 1]$

10. if $L_{00} + U_{11} - 1 = 0$ or $U_{00} + L_{11} - 1 = 0$ and $(U_\lambda + U_{00} - 1)(L_\lambda + L_{00} - 1) \geq 0$:
    - if (length($A > 0$)) $= 0$) min=0 else min=min(min($A > 0$),1)
    - max=1
    - the interval is $[\text{min}, 1] \cup \{0\}$
11. if $L_{00} + U_{11} - 1 = 0$ or $U_{00} + L_{11} - 1 = 0$ and $(U_\lambda + U_{00} - 1)(L_\lambda + L_{00} - 1) < 0$:
    - min=0, max=1,length=1
    - the interval is $[0, 1]$

APPENDIX C

Using the Delta method to Estimate Variance for the Slope

From Corollary 4.2.1,

$$\beta_1 = \frac{\lambda(1-\lambda)(\theta_{00}+\theta_{11}-1)\beta_{w1}}{(\lambda+\theta_{00}-1)(\theta_{11}-\lambda)} = f(\beta_{w1}, \lambda, \theta_{00}, \theta_{11})$$

Using the delta method, we can estimate the variance of $\widehat{\beta}_1$:

$$\widehat{\text{Var}}(\widehat{\beta}_1) = \nabla f(\widehat{\beta}_{w1}, \widehat{\lambda}, \widehat{\theta}_{00}, \widehat{\theta}_{11})^{\mathsf{T}}\, \widehat{\text{Cov}}(\begin{pmatrix} \widehat{\beta}_{w1} \\ \widehat{\lambda} \\ \widehat{\theta}_{00} \\ \widehat{\theta}_{11} \end{pmatrix}) \nabla f(\widehat{\beta}_{w1}, \widehat{\lambda}, \widehat{\theta}_{00}, \widehat{\theta}_{11}).$$

In the following, we will show the detailed computations that we need for the above formula.

Lemma C.0.1 *$Cov(\widehat{\beta}_{w1}, \widehat{\lambda}) = 0.$*

Proof

$$\begin{aligned}
\text{Cov}(\widehat{\beta}_{w1}, \widehat{\lambda}) &= E\left\{\text{Cov}(\widehat{\beta}_{w1}, \widehat{\lambda}|\mathbf{W})\right\} + \text{Cov}\left\{E(\widehat{\beta}_{w1}|\mathbf{W}), E(\widehat{\lambda}|\mathbf{W})\right\} \\
&= \text{Cov}\left[E\left\{\frac{\sum_i y_i(W_i - \widehat{\lambda})}{n\widehat{\lambda}(1-\widehat{\lambda})}|\mathbf{W}\right\}, \widehat{\lambda}\right] \\
&= \text{Cov}\left\{\frac{\sum_i(\beta_{w0} + \beta_{w1}W_i)(W_i - \widehat{\lambda})}{n\widehat{\lambda}(1-\widehat{\lambda})}, \widehat{\lambda}\right\} \\
&= \text{Cov}(\beta_{w1}, \widehat{\lambda}) \\
&= 0.
\end{aligned}$$

Since $(\widehat{\theta}_{00}, \widehat{\theta}_{11})$ are independent from $(\widehat{\beta}_{w1}, \widehat{\lambda})$, and we have

$$\widehat{\text{Cov}}(\begin{pmatrix} \widehat{\beta}_{w1} \\ \widehat{\lambda} \\ \widehat{\theta}_{00} \\ \widehat{\theta}_{11} \end{pmatrix}) = \begin{pmatrix} \widehat{\text{Var}}(\widehat{\beta}_{w1}) & 0 & 0 & 0 \\ 0 & \widehat{\text{Var}}(\widehat{\lambda}) & 0 & 0 \\ 0 & 0 & \widehat{\text{Var}}(\widehat{\theta}_{00}) & 0 \\ 0 & 0 & 0 & \widehat{\text{Var}}(\widehat{\theta}_{11}) \end{pmatrix}.$$

By direct computation, we will have

$$\begin{aligned} &\nabla f(\widehat{\beta}_{w1}, \widehat{\lambda}, \widehat{\theta}_{00}, \widehat{\theta}_{11}) \\ = &\begin{pmatrix} \dfrac{\widehat{\lambda}(1-\widehat{\lambda})}{(\widehat{\lambda}+\widehat{\theta}_{00}-1)(\widehat{\theta}_{11}-\widehat{\lambda})} \\ \dfrac{\widehat{\beta}_{w1}(\widehat{\theta}_{00}+\widehat{\theta}_{11}-1)\left\{-\widehat{\lambda}^2(\widehat{\theta}_{11}-\widehat{\theta}_{00})+\widehat{\theta}_{11}(\widehat{\theta}_{00}-1)(1-2\widehat{\lambda})\right\}}{(\widehat{\lambda}+\widehat{\theta}_{00}-1)^2(\widehat{\lambda}-\widehat{\theta}_{11})^2} \\ -\dfrac{\widehat{\lambda}(1-\widehat{\lambda})\widehat{\beta}_{w1}}{(\widehat{\lambda}+\theta_{00}-1)^2} \\ -\dfrac{\widehat{\lambda}(1-\widehat{\lambda})\widehat{\beta}_{w1}}{(\widehat{\lambda}-\widehat{\theta}_{11})^2} \end{pmatrix}. \end{aligned}$$

BIBLIOGRAPHY

[1] Agresti A, Coull B. (1998) Approximate is better than ’exact’ for interval estimation of binomial proportions, *American Statistician* **52**(2), pp. 119-126.

[2] Akazawa K, Kinukawa N, Nakamura T. (1998) A note on the corrected score function adjusting for misclassification, *J. Japan Statist. Soc.* **28**(1), pp. 115-123.

[3] Amemiya Y. (1985) What should be done when an estimated between-group covariance matrix is not nonnegative definite?, *The American Statistician* **39**(2), pp. 112-117.

[4] Barron BA. (1977) Effects of misclassification on estimation of relative risk, *Biometrics* **33**(2), pp. 414-418.

[5] Bock RD, Petersen AC. (1975) A multivariate correction for attenuation, *Biometrika* **62**(3), pp. 673-678.

[6] Bross I. (1954) Misclassification in 2 x 2 tables, *Biometrics* **10**(4), pp. 478-486.

[7] Brown LD, Cai TT, DasGupta A. (2001) Interval estimation for a binomial proportion, *Statistical Science* **16**(2), pp. 101-133.

[8] Buonaccorsi JP. (1996) A modified estimating equation approach to correcting for measurement error in regression, *Biometrika* **83**(2), pp. 433-440.

[9] Buonaccorsi JP, Laake P, Veierod MB. (2005) On the effect of misclassification on bias of perfectly measured covariates in regression, *Biometrics* **61**(3), pp. 831-836.

[10] Buonnacorsi JP. (2001) Fieller’s theorem, in *Encyclopedia of Environmetrics*, El-Shaarawi AH, Piegorsch WW, eds., Chichester, New York.

[11] Buonnacorsi JP. (2010) *Measurement Error and Misclassification: Models, Methods and Applications*. Chapman & Hall, Boca Raton.

[12] Carroll RJ. (1998) Measurement error in epidemiologic studies, in *Encyclopedia of Biostatistics*, Armitage P, Colton T, eds., Wiley, New York.

[13] Carroll RJ, Gallo P, Gleser LJ. (1985) Comparison of least-squares and errors-in-variables regression, with special reference to randomized analysis of covariance, *Journal of the American Statistical Association* **80**(392), pp. 929-932.

[14] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. (2006) *Measurement Error in Nonlinear Models*. Chapman & Hall, Boca Raton.

[15] Casella G, Berger RL. (1990) *Statistical Inference*. Duxbury Press, Belmont, Ca.

[16] Chen TT. (1989) A review of methods for misclassified categorical-data in epidemiology, *Statistics in Medicine* **8**(9), pp. 1095-1106.

[17] Christopher SR, Kupper LL. (1995) On the effects of predictor misclassification in multiple linear-regression analysis, *Communications in Statistics-Theory and Methods* **24**(1), pp. 13-37.

[18] Cochran WG. (1968) Errors of measurement in statistics, *Technometrics* **10**(4), p. 637.

[19] Cook JR, Stefanski LA. (1994) Simulation-extrapolation estimation in parametric measurement error models, *Journal of the American Statistical Association* **89**(428), pp. 1314-1328.

[20] Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. (1977) Bias due to misclassification in estimation of relative risk, *American Journal of Epidemiology* **105**(5), pp. 488-495.

[21] Dalabehera M, Sahoo LN. (1995) Efficiencies of six almost unbiased ratio estimators under a particular model, *Statistical Papers* **36**(1), pp. 61-67.

[22] Davidov O, Faraggi D, Reiser B. (2003) Misclassification in logistic regression with discrete covariates, *Biometrical Journal* **45**(5), pp. 541-553.

[23] Demidenko E. (2004) *Mixed Models: Theory and Applications*. Wiley, New York.

[24] Dosemeci M, Wacholder S, Lubin JH. (1990) Does nondifferential misclassification of exposure always bias a true effect toward the null value, *American Journal of Epidemiology* **132**(4), pp. 746-748.

[25] Espeland MA, Hui SL. (1987) A general-approach to analyzing epidemiologic data that contain misclassification errors, *Biometrics* **43**(4), pp. 1001-1012.

[26] Fieller EC. (1954) Some problems in interval estimation, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **16**(2), pp. 175-185.

[27] Frost C, Thompson SG. (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable, *Journal of the Royal Statistical Society Series A-Statistics in Society* **163**, pp. 173-189.

[28] Freedson PS, Melanson E, Sirard J. (1998) Calibration of the Computer Science and Applications, Inc. accelerometer, *Medicine and Science in Sports and Exercise* **30**, pp. 777-778.

[29] Frost C, Thompson SG. (2000) Science and Applications, Inc. accelerometer, *Journal of the Royal Statistical Society Series A-Statistics in Society* **163**, pp. 173-189.

[30] Fuller WA. (1987) *Measurement Error Models*. Wiley, New York.

[31] Gladen B, Rogan WJ. (1979) Misclassification and the design of environmental-studies, *American Journal of Epidemiology* **109**(5), pp. 607-616.

[32] Gong G, Samaniego FJ. (1981) Pseudo maximum-likelihood estimation - theory and applications, *Annals of Statistics* **9**(4), pp. 861-869.

[33] Greenland S. (1988) Variance-estimation for epidemiologic effect estimates under misclassification, *Statistics in Medicine* **7**(7), pp. 745-757.

[34] Greenland S. (2008) Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification, *Journal of Statistical Planning and Inference* **138**(2), pp. 528-538.

[35] Guiard V. (1989) Some remarks on the estimation of the ratio of the expectation values of a two-dimensional normal random variable (correction of the theorem of Milliken), *Biometrical Journal* **31**(6), pp. 681-697.

[36] Gustafson P. (2004) *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall, Boca Raton.

[37] Harper D. (1964) Misclassification in epidemiological surveys, *American Journal of Public Health and the Nations Health* **54**(11), pp. 1882-1886.

[38] Harville DA. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer, New York.

[39] Hofler M. (2005) The effect of misclassification on the estimation of association: a review, *International Journal of Methods in Psychiatric Research* **14**(2), pp. 92-101.

[40] Hui SL, Walter SD. (1980) Estimating the error rates of diagnostic-tests, *Biometrics* **36**(1), pp. 167-171.

[41] Hutchison MC. (1971) Monte Carlo comparison of some ratio estimators, *Biometrika* **58**(2), p. 313.

[42] Katz BM, McSweeney M. (1979) Misclassification errors and categorical data-analysis, *Journal of Experimental Education* **47**(4), pp. 331-338.

[43] Kuchenhoff H, Lederer W, Lesaffre E. (2007) Asymptotic variance estimation for the misclassification SIMEX, *Computational Statistics & Data Analysis* **51**(12), pp. 6197-6211.

[44] Kuchenhoff H, Mwalili SM, Lesaffre E. (2006) A general method for dealing with misclassification in regression: The misclassification SIMEX, *Biometrics* **62**(1), pp. 85-96.

[45] Kuha J. (1997) Estimation by data augmentation in regression models with continuous and discrete covariates measured with error, *Statistics in Medicine* **16**(3-Jan), pp. 189-201.

[46] Kuha J, Skinner C, Palmgren J. (1998) Misclassification error, in *Encyclopedia of Biostatistics*, Armitage P, Colton T, eds., Wiley, New York.

[47] Levine JA. (2005) Measurement of energy expenditure, *Public Health Nutrition* **8**, pp. 1123-1132.

[48] Li L, Palta M, Shao J. (2004) A measurement error model with a Poisson distributed surrogate, *Statistics in Medicine* **23**(16), pp. 2527-2536.

[49] Liang KY, Liu XH. (1991) Estimating equations in generalized linear models with measurement error, in *Estimating Functions*, Godambe VP, ed., Oxford University Press, Cambridge.

[50] Lin HM, Lyles RH, Williamson JM, Kunselman AR. (2005) Estimation of the intervention effect in a nonrandomized study with pre- and post-mismeasured binary, *Statistics in Medicine* **24**, pp. 419-435.

[51] Liu XH, Liang KY. (1991) Adjustment for nondifferential misclassification error in the generalized linear-model, *Statistics in Medicine* **10**(8), pp. 1197-1211.

[52] Louis TA. (1982) Finding the observed information matrix when using the EM Algorithm, *Journal of the Royal Statistical Society* **44**(2), pp. 226-233.

[53] Luenberger DG. (1973) *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Reading, Ma.

[54] Lyles RH. (2002) A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure, *Biometrics* **58**(4), pp. 1034-1036.

[55] Marshall RJ. (1990) Validation-study methods for estimating exposure proportions and odds ratios with misclassified data, *Journal of Clinical Epidemiology* **43**(9), pp. 941-947.

[56] McLachlan G, Basford KE. (1988) *Mixture Models*. Marcel Dekker, New York.

[57] McLachlan G, Peel D. (2000) *Finite Mixture Models*. Wiley, New York.

[58] Milliken GA. (1982) On a confidence-interval about a parameter estimated by a ratio of normal random-variables, *Communications in Statistics Part A-Theory and Methods* **11**(17), pp. 1985-1995.

[59] Morrissey MJ, Spiegelman D. (1999) Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons, *Biometrics* **55**(2), pp. 338-344.

[60] Mote VL, Anderson RL. (1965) An investigation of effect of misclassification on properties of x2-tests in analysis of categorical data, *Biometrika* **52**, p. 95.

[61] Nakamura T. (1990) Corrected score function for errors-in-variables models - methodology and application to generalized linear-models, *Biometrika* **77**(1), pp. 127-137.

[62] Newcombe RG. (1998) Two-sided confidence intervals for the singel proportion: comparison of seven methods, *Statistics in Medicine* **17**, pp. 857-872.

[63] Parke WR. (1986) Pseudo maximum-likelihood estimation: The asymptotic distribution, *Annals of Statistics* **14**(1), pp. 355-357.

[64] Pate RR, Pratt M, Blair SN, et al. (1995) Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine, *Journal of the American Medical Association* **273**, pp. 402-407.

[65] Perez CJ, Giron FJ, Martin J, Ruiz M, Rojano C. (2007) Misclassified multinomial data: a Bayesian approach, *Evista de la Real Academia de Ciencias Exactas Fisicas y Naturales* **101**(1), pp. 71-80.

[66] Pober D, Staudenmayer J, Raphael C, Freedson PS. (2006) Development of a novel analytical technique to assess physical activity using accelerometers, *Medicine and Science in Sports and Exercise* **38**(9), pp. 1626-1634.

[67] Prescott GJ, Garthwaite PH. (2002) A simple Bayesian analysis of misclassified binary data with a validation substudy, *Biometrics* **58**(2), pp. 454-458.

[68] Quade D, Lachenbruch PA, Whaley FS, McClish DK, Haley RW. (1980) Effects of misclassifications on statistical inferences in epidemiology, *American Journal of Epidemiology* **111**(5), pp. 503-515.

[69] Rao PSRS, Rao JNK. (1971) Small sample results for ratio estimators, *Biometrika* **58**(3), p. 625.

[70] Reade-Christopher SJ, Kupper LL. (1991) Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data, *Biometrics* **47**, pp. 535-548.

[71] Rosner B, Spiegelman D, Willett WC. (1992) Correction of logistic-regression relative risk estimates and confidence-intervals for random within-person measurement error, *American Journal of Epidemiology* **136**(11), pp. 1400-1413.

[72] Rubin T, Rosenbaum J, Cobb S. (1956) The use of interview data for the detection of associations in field studies, *J Chronic Dis* **4**(3), pp. 253-266.

[73] Schafer DW. (1986) Combining information on measurement error in the errors-in-variables model, *Journal of the American Statistical Association* **81**(393), pp. 181-185.

[74] Schwartz JE. (1985) The neglected problem of measurement error in categorical data, *Sociological Methods & Research* **13**(4), pp. 435-466.

[75] Selen J. (1986) Adjusting for errors in classification and measurement in the analysis of partly and purely categorical-data, *Journal of the American Statistical Association* **81**(393), pp. 75-81.

[76] Shao J. (2003) *Mathematical Statistics*. Springer-Verlag, New York.

[77] Sirard JR, Pate RR. (2001) Physical activity assessment in children and adolescents, *Sports Medicine* **31**, pp. 439-454.

[78] Spiegelman D, Rosner B, Logan R. (2000) Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs, *Journal of the American Statistical Association* **95**(449), pp. 51-61.

[79] Stamey JD, Seaman JW, Young DM. (2007) Bayesian estimation of intervention effect with pre- and post-misclassified binomial data, *Journal of Biopharmaceutical Statistics* **17**(1), pp. 93-108.

[80] Stefanski LA, Carroll RJ. (1987) Conditional scores and optimal scores for generalized linear measurement-error models, *Biometrika* **74**(4), pp. 703-716.

[81] Stephens DA, Dellaportas P. (1992) Bayesian analysis of generalised linear models with covariate measurement error, *Bayesian Statistics* **4**, pp. 813-820.

[82] Tenenbein A. (1970) A double sampling scheme for estimating from binomial data with misclassifications, *Journal of the American Statistical Association* **65**(331), pp. 1350-1361.

[83] Tin M. (1965) Comparison of some ratio estimators, *Journal of the American Statistical Association* **60**(309), pp. 294-307.

[84] van den Hout A, Kooiman P. (2006) Estimating the linear regression model with categorical covariates subject to randomized response, *Computational Statistics & Data Analysis* **50**(11), pp. 3311-3323.

[85] van den Hout A, van der Heijden PGM. (2002) Randomized response, statistical disclosure control and misclassification: a review, *International Statistical Review* **70**(2), pp. 269-288.

[86] Veierod MB, Laake P. (2001) Exposure misclassification: bias in category specific Poisson regression coefficients, *Statistics in Medicine* **20**(5), pp. 771-784.

[87] Vollset SE. (1993) Confidence-intervals for a binomial proportion, *Statistics in Medicine* **12**(9), pp. 809-824.

[88] von Luxburg U, Franz VH. (2009) A geometric approach to confidence sets for ratios; Fieller's theorem, generalizations, and bootstrap, *Statistica Sinica (to appear)* .

[89] Walter SD, Irwig LM. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data - a review, *Journal of Clinical Epidemiology* **41**(9), pp. 923-937.

[90] White I, Frost C, Tokunaga S. (2001) Correcting for measurement error in binary and continuous variables using replicates, *Statistics in Medicine* **20**(22), pp. 3441-3457.

[91] Zelen M, Haitovsky Y. (1991) Testing hypotheses with binary data subject to misclassification errors - analysis and experimental-design, *Biometrika* **78**(4), pp. 857-865.

[92] Zucker DM, Spiegelman D. (2008) Corrected score estimation in the proportional hazards model with misclassified discrete covariates, *Statistics in Medicine* **27**(11), pp. 1911-1933.