Distribution Agreement

Multiple Roots in Logistic Regression With Errors-in-Covariates

By

Jian Chen

Doctor of Philosophy

Biostatistics

John J. Hanfelt, Ph.D.
Advisor

Eugene Huang, Ph.D.
Advisor

Andrew N. Hill, Ph.D.
Committee Member

Qi Long, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Multiple Roots in Logistic Regression With Errors-in-Covariates

By

Jian Chen

B.E., Nanjing University of Science and Technology, 1994

M.S., University of Toledo, 2004

Advisor: John J. Hanfelt, Ph.D.

Advisor: Eugene Huang, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2009

Abstract

Multiple Roots in Logistic Regression With Errors-in-Covariates

By Jian Chen

The unbiased estimating function method is a flexible approach to estimate and make inferences on the parameters of interest. However, special problems arise when covariates are measured with error.

Measurement errors arise in public health studies when some covariates are not measured precisely. We focus on the important case where the outcome is a binary variable and the interest is in coefficients from a logistic regression model. Two widely used estimating function methods for logistic regression with errors-in-covariates are the conditional score (Stefanski & Carroll 1987) and the parametric-correction estimation procedure (Huang & Wang 2001). The conditional score can have multiple-roots and not all of them are consistent, whereas the parametric-correction estimation only generate consistent roots. On the other hand, the conditional score in theory has an efficiency advantage in that its consistent estimator is asymptotically locally efficient. Despite the multiple-roots problem, the conditional approach is regarded as the standard method.

In this dissertation research, we aim to resolve the multiple-roots problem of the conditional score in logistic regression with errors-in-covariates. We investigate the root behaviors of the conditional score in finite samples and demonstrate the existence and seriousness of the problem posed by multiple roots, which have not been studied adequately in literature.

We propose two methods to achieve our research goal. In the first approach, we develop a weighted-correction estimating function that only yields consistent estimators and combine it with the conditional score using empirical likelihood. We prove that, asymptotically, the proposed approach admits only consistent estimators and is locally efficient.

In the second approach, we construct objective functions based on the weighted-correction estimating function and use them to distinguish among multiple roots from the conditional

score.

In addition to developing the large sample theories of the proposed methods, we investigate their finite-sample properties through an extensive simulation. The simulation studies show that the proposed methods work well in finite samples and outperform existing methods in many situations. Finally, the proposed methods are applied to data presented in Hammer et al. (1996) and Pan et al. (1990).

Multiple Roots in Logistic Regression With Errors-in-Covariates

By

Jian Chen

B.E., Nanjing University of Science and Technology, 1994

M.S., University of Toledo, 2004

Advisor: John J. Hanfelt, Ph.D.

Advisor: Eugene Huang, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2009

Acknowledgments

I would like to thank my advisors, Dr. John J. Hanfelt and Eugene Huang, for their guidance and great patience through my research journey. I would like to thank my committee members Dr. Qi Long and Dr. Andrew N. Hill for their constructive comments and suggestions. I would like to say a special thanks to Mr. Kirk A. Easley who offered me opportunities to deal with practical projects, and Dr. Robert H. Lyles who introduced me to work part-time at CDC. I would also like to thank all the faculty, staff, and students in the Department of Biostatistics and Bioinformatics for their help through those years.

Finally, I would like to deeply thank my parents for their never-ending support and unconditional love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Biomedical research often involves studying the effects of certain covariates on some binary outcome variable, and logistic regression is the most used binary regression to investigate those mechanisms. When some covariates are subject to measurement errors, special methods are needed to correct for the effects of measurement errors. This dissertation research is motivated by two real studies where covariates are measured with errors.

The AIDS Clinical Trials group (ACTG) 175 study (Hammer et al. 1996) is a randomized, double-blind, placebo-controlled trial to compare treatment with either a single nucleoside or two nucleosides in adults infected with human immunodeficiency virus type 1 (HIV-1) whose CD4 cell counts were from 200 to 500 per cubic millimeter and had no history of an AIDS-defining illness. A total of 2467 HIV-1-infected patients were recruited from 43 AIDS Clinical Trials Units and 9 National Hemophilia Foundation sites in the United States and Puerto Rico. A particular research question is to assess the effect of the true baseline CD4 count on the symptomatic HIV infection defined as candidiasis, oral hairy leukoplakia, or herpes zoster reported within 30 days before randomization in antiretroviral-naive patients (Huang & Wang 2000, 2001). A single covariate logistic regression model that treats the symptomatic HIV infection as outcome and the true baseline log(CD4) count as the covariate can be fitted to assess their relationship. Since the true baseline CD4 count is unobservable, the screening baseline CD4 count is usually used as a substitute as the true baseline CD4 count. As a fact, the screening baseline CD4 count is subject to both instrumental error and biological diurnal fluctuation. Therefore, the true baseline CD4 count is measured with errors.

The second example is the blood pressure study in Pan et al. (1990). In this study, the relationship between 24-hour urinary sodium chloride and blood pressure were investigated in 397 middle-aged Chinese men living in Taipei. A logistic regression model with high systolic blood pressure as outcome and the 24-hour urinary sodium chloride measurement, plus age and body mass index (BMI) as covariates can be applied here. A single (the most recent) urinary sodium chloride measurement can be used as a surrogate of the unobservable true 24-hour urinary sodium chloride measurement in the model, which induces the measurement errors.

When covariates are measured with errors, the naive approach that ignores the measurement errors by treating the surrogates as the true unobservable covariates would lead to inconsistent results. Therefore, correction methods must be applied to the naive approach to correct for the effects of measurement errors. Two widely used methods for logistic regression with errors-in-covariates are the conditional score (Stefanski & Carroll 1987) and the parametric-correction estimation procedure (Huang & Wang 2001).

1.2 The problems of existing methods

The conditional score (Stefanski & Carroll 1987) is regarded as the default functional method for logistic regression with errors-in-covariates. The main advantage of the conditional score over other approaches is its theoretical efficiency property: it is asymptotically locally efficient as described in Stefanski & Carroll (1987). Stefanski & Carroll (1987) showed that , however, the conditional score can have multiple-roots and not all of them are consistent. In the example of the blood pressure study, the conditional score yields three solutions for the coefficient of urinary sodium chloride : -4.36, 0.61, and 6.63. Unfortunately, the conditional score does not indicate which one is appropriate.

In the presence of multiple solutions, Stefanski & Carroll (1987) suggest choosing the solution closest to the naive estimator obtained by ignoring the measurement error. To implement this strategy, they suggest iterating from the naive estimator to find the conditional score estimator using a standard numerical method such as Newton-Raphson. However, this heuristic and theoretically unjustifiable approach can break down when measurement error is large (Stefanski & Carroll 1987). Hanfelt & Liang (1997) proposed a conditional quasi-likelihood function to distinguish between the multiple roots of the conditional score. This approach appears to work well, but, the conditional quasi-likelihood function is formed by a, in general, path-dependent line integral and therefore may not be unique.

To secure consistent estimation, Huang & Wang (2001) developed a parametric-correction estimation procedure. Instead of finding roots of estimating functions, Huang & Wang (2001) construct an objective function by combining two estimating functions and searching for the global minimizer of the objective function. Each estimating function has the same dimension as the parameters and only admits consistent roots. Compared with the conditional score, the parametric-correction estimation procedure does not yield inconsistent estimators. However, this approach is generally less efficient than the conditional score, especially when the measurement error is large (Huang & Wang 2001).

In summary, both the conditional score (Stefanski & Carroll 1987) and the parametric-correction estimation procedure (Huang & Wang 2001) have advantages and disadvantages. The conditional score is in general preferred in that its consistent estimator is asymptotically locally efficient.

Even though the multiple-roots problem of the conditional score has been reported in the literature for a long time (Stefanski & Carroll 1987, Hanfelt & Liang 1997, Small et al. 2000), its seriousness in practice remains unclear. The investigations of multiple roots of conditional score in finite samples are very limited so far. The potential practical difficulties due to multiple roots have not been examined adequately.

1.3 Objectives

In this dissertation research, the goal is to resolve the multiple-roots problem of the conditional score in in Logistic regression with errors-in-covariates. This can be achieved by two approaches.

Sparked by the idea of combining estimating functions (Huang & Wang 2001), we can combine the conditional score and the two estimating functions in the parametric-correction estimating procedure (Huang & Wang 2001) to produce consistent estimators. The inconsistent roots of the conditional score are eliminated since those two other estimating functions do not generate inconsistent roots. On the other hand, the local optimality property of the conditional score will be preserved if a proper combining technique is chosen. As a result, the new procedure has the advantages and gets rid of disadvantages from each individual approach. However, practical concerns arise since we need to estimate P-dimensional parameters using $3P$-dimensional estimating functions, which could cause numerical difficulties in small or medium sized samples. A better approach is to develop a P-dimensional estimating function that has no multiple-roots problem and combine it with the conditional score.

The second approach is to build objective functions to distinguish among multiple roots of the conditional score. As long as the consistent root of the conditional score can be identified, the asymptotical local efficiency is secured.

The objectives of this dissertation are summarized as follows:

1. Investigate the root behaviors of the conditional score (Stefanski & Carroll 1987) and the parametric-correction estimation procedure (Huang & Wang 2001) in finite samples.

2. Develop an estimation procedure that is guaranteed to produces asymptotically locally efficient estimator by combining the conditional score Stefanski & Carroll (1987) with a new estimating function that has no multiple roots in large samples. Develop the large sample theory of the proposed estimator.

3. Develop objective functions to distinguish among multiple roots of the conditional score (Stefanski & Carroll 1987).

4. Investigate the finite-sample properties of the two proposed methods through simulations and apply them to two real examples (Hammer et al. 1996, Pan et al. 1990)

Chapter 2

Background

2.1 Functional methods for logistic regression with errors-in-covariates

2.1.1 Introduction

Measurement error models arise in making inference on the relationship of a response variable and predictor variables that may be subject to measurement errors. When a covariate is measured with error, the corresponding maximum likelihood estimator is biased and therefore, adjustments must be made to account for the effects of measurement errors.

Let Y be the response variable and X be the covariate vector. Often, a component of X is measured with error. Instead of observing X, we observe a surrogate for X, W. The model $p(W \mid X)$ is called the measurement error model (MEM) and typically assumed to be known. The most commonly used MEM is the classical unbiased additive normal error model:

$$W = X + U, \tag{2.1}$$

where the error $U \sim N(0, \Sigma_{uu})$ and it is independent of (Y, X). In other words, $E(Y|X, W) = E(Y|X)$: the measurement errors are nondifferential.

The effects of measurement error can be illustrated via a simple linear measurement model. Consider the model

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

where the unobservable X has variance of σ_x^2. The nondifferential measurement error U has a variance of σ_u^2. Fuller (1987) showed that the naive approach that ignores the measurement error and regress Y on W would yield inconsistent estimator of β. Indeed, the naive approach consistently estimates $\beta_* = \lambda\beta$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}.$$

It is clear that $\lambda < 1$. Therefore, the naive estimator is biased toward zero in large samples, a phenomenon called attenuation. In addition, $\text{var}(Y|W) > \text{var}(Y|X)$. That is, the measurement error also makes the data more noisy. It should be noticed that attenuation is the natural consequence only for simple linear model with classical additive measurement error model.

To make consistent inference on β, one needs to correct the bias induced by measurement error. In the simple linear model, the correction can be performed easily by method-of-moments method. If $\hat{\lambda}$ is a consistent estimator of λ, then β is consistently estimated by $\hat{\beta} = \hat{\beta}_*/\hat{\lambda}$. The method-of-moments method can be extended to multiple linear regression models with covariates measured with errors.

For nonlinear models, the effects of measurement error are more complicated. The above method-of-moments method is inadequate to correct the bias induced by the measurement errors. General approaches include the approximately consistent methods: Regression calibration and SIMEX, and consistent functional methods: conditional score and corrected score.

Regression calibration was first presented by Prentice (1982), and was later extended by B.Armstrong (1985) and Rosner et al. (1989, 1990), among others. It is a popular method to estimated the coefficients in models with one or more continuous covariates measured with an error. It can be easily implemented into the models and the computation is straightforward and stable. Indeed, regression calibration is often regarded as a benchmark in evaluating error-correction methods. The nondifferential assumption is required to apply regression calibration. Suppose that we have a model with covariates X measured with error and covariates Z measured precisely. First, $E(X|Z, W)$ needs to be estimated using replication, validation or instrumental data. Then the unobservable X is replaced by the estimate of $E(X|Z, W)$ in the original model so that the standard analysis can be applied to estimate the parameters. SIMEX (Cook & Stefanski 1995) is a simulation-based method and is computationally intensive. Both methods are fully consistent for linear regression models. However, they are only approximately consistent for nonlinear models.

In the classical functional method approach, the unobservable true covariates X are regarded as fixed but unknown constant and hence are nuisance parameters. The functional method approach usually assume the classical additive measurement error model $W = X + U$. The conditional score (Stefanski & Carroll 1987), and the corrected score (Nakamura 1990, Stefanski 1989) are two major functional method methods. For the special case of logistic regression with errors-in-covariates, Stefanski (1989) showed that the corrected score does not exist. Huang & Wang (2001) developed a corrected-type parametric-correction estimation procedure to logistic regression with errors-in-covariates.

2.1.2 Logistic regression with errors-in-covariates

The logistic regression model takes the form

$$\text{Logit}(\Pr(Y = 1 \mid X, Z)) = \alpha + \beta_x^{\mathrm{T}} X + \beta_z^{\mathrm{T}} Z \tag{2.2}$$

where Y is the binary outcome and X and Z are covariates. Straightforward calculations show that the score function for $\theta = (\alpha, \beta_x^{\mathrm{T}}, \beta_z^{\mathrm{T}})^{\mathrm{T}}$ is

$$S(\theta) = \{Y - F(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X)\} \begin{pmatrix} 1 \\ Z \\ X \end{pmatrix}, \tag{2.3}$$

where $F(t) \equiv \{1 + e^{-t}\}^{-1}$. Assume that X is measured with error and Z is measured precisely . In stead of observing X, we observe a surrogate of X, W. In the presence of measurement error, (2.3) can no longer be used to obtain consistent estimates of parameters.

2.1.3 The conditional score

Stefanski & Carroll (1987) adopted the conditional score by Lindsay (1982) and considered the generalized linear model in canonical form where some covariates are measured with independent normal error.

Assume that the distribution of Y given (X, Z) takes the form of the canonical generalized linear models (McCullagh & Nelder 1989)

$$f(Y|X, Z, \theta) = \exp\left\{\frac{Y\eta - b(\eta)}{\phi} + c(Y, \phi)\right\}, \tag{2.4}$$

where the natural parameter $\eta = \beta_0 + \beta_x^{\mathrm{T}} X + \beta_z^{\mathrm{T}} Z$, and the parameter to be estimated is $\theta = (\beta_0, \beta_x^{\mathrm{T}}, \beta_z^{\mathrm{T}}, \phi)^{\mathrm{T}}$. The independent random variables (Y, W) has the joint density

$$f(Y, W|X, Z, \theta) = f(Y|X, Z, \theta) f(W|X). \tag{2.5}$$

Under (2.1), $W|X \sim N(X, \Sigma_{uu})$. (2.5) is also a canonical generalized linear model. If the unobservable X is regarded as unknown parameters and all other parameters are fixed in

(2.5), a complete and sufficient statistic for X is

$$\Delta = W + Y\Sigma_{uu}\beta_x/\phi.$$

Let $H = (\partial/\partial\theta)\log f(y, w|x, z, \theta)$ be the θ-score. The conditional score takes the form

$$\Psi_{cs}(Y, W, Z, \theta) = H - E(H|\Delta).$$

Taking advantage of the fact that $f(y, w|x, z, \theta)$ is a canonical generalized linear model, one can show that

$$\Psi_{cs}(Y, W, Z, \theta) = \begin{bmatrix} \dfrac{\{Y - E(Y|Z, \Delta)\}}{\phi} \begin{pmatrix} 1 \\ Z \\ X \end{pmatrix} \\ \left(\dfrac{n-p}{n}\right)\phi - \dfrac{\{Y - E(Y|Z, \Delta)\}^2}{\text{var}(Y|Z, \Delta)/\phi} \end{bmatrix} \tag{2.6}$$

The unknown nuisance parameters X appear in (2.6) only as a weight. One can substitute any estimator of X into (2.6) and the unbiasedness of $\Psi_{cs}(Y, W, Z, \theta)$ remains unchanged. Since Δ is a sufficient and complete statistic for X, any real function of Δ: $t(\Delta)$ would be a candidate. Intuitively, the choice of $t(\Delta)$ should be close to X.

Logistic model belongs to canonical generalized linear models (2.4) with $\phi = 1$ and $c(y, \phi) = 0$. Therefore, the conditional score (2.6) reduces to

$$\Psi_{cs}(\theta) = \{Y - F(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}}(W + (Y - 1/2)\Sigma_{uu}\beta_x)\} \begin{pmatrix} 1 \\ Z \\ t(\Delta) \end{pmatrix}, \tag{2.7}$$

where $t(\cdot)$ is some known function of Δ. Then the conditional score estimator $\hat{\theta}_{cs}$ can be obtained by solving

$$\sum_{i=1}^{n} \Psi_{cs}(y_i, w_i, z_i, \hat{\theta}_{cs}) = 0.$$

Under certain circumstances, Ψ_{cs} can be fully efficient (Stefanski & Carroll 1987). Treating X as independent and identically distributed random variables with unknown distribution, Stefanski & Carroll (1987) derived the efficient score for θ that has the same expression of (2.6) with X replaced by $E(X|Z,\Delta)$. The efficient score achieves the so-called semiparametric efficiency bound if $E(X|Z,\Delta)$ is correctly modeled. Since the efficiency of the efficient score depends on the unknown density of X, the efficient score is called semiparametric locally efficient. Therefore, the choice of $t(\Delta) = E(X|Z,\Delta)$ leads the conditional score (2.6) to be the locally optimal efficient score. However, $E(X|Z,\Delta)$ is typically complicated and not feasible to model. Stefanski & Carroll (1987) consider to choose $t(\Delta)$ that is linear in (Z,Δ). Then the conditional score would be optimal if $E(X|Z,\Delta)$ is also linear in (Z,Δ). Stefanski & Carroll (1987) present the conditions for the density of X that would lead to a linear $E(X|\Delta = \delta)$ in (Z,Δ).

The efficiency advantages of the conditional score have been investigated by Stefanski (1989) and Huang & Wang (2001), among others. The results show that the conditional score is in general more efficient than other functional consistent methods.

2.1.4 The parametric-correction estimation procedure

The conditional score (Stefanski & Carroll 1987) removes biases by conditioning on a complete and sufficient statistic of the unobservable X . Its unbiasedness is based on the crucial additive normal error assumption. The corrected score (Nakamura 1990, Stefanski 1989) removes biases by correcting the error-contaminated score function. In the corrected score approach, the measurement error is not restricted to be normally distributed. However, the classical additive error model is usually assumed.

Let $l(\theta; Y, X, Z)$ be the loglikelihood function based on (Y, X, Z). If we observe X, we can construct the score function $S(Y, X, Z; \theta) = \partial l(\theta; Y, X, Z)/\partial\theta$. It is well-known that the score function produce consistent and efficient estimator. Since we are not able to observe X, we can not use $S(Y, X, Z; \theta)$ to estimate parameters. Let $l^*(\theta; Y, W, Z)$ be the corrected loglikelihood function based on the observed data that satisfies

$$E\{l^*(\theta; Y, W, Z)|(Y, X, Z)\} = l(\theta; Y, X, Z).$$

Then the corrected score function is

$$S^*(Y, W, Z; \theta) = \frac{\partial l^*(\theta; Y, W, Z)}{\partial \theta}. \tag{2.8}$$

It is easily to show that

$$E\{S^*(Y, W, Z; \theta)|(Y, X, Z)\} = S(Y, X, Z; \theta).$$

Therefore, $S^*(Y, W, Z; \theta)$ is also unbiased and the estimating equations

$$\sum_{i=1}^{n} S^*(Y, W, Z; \theta) = 0$$

yields consistent estimators.

However, Stefanski (1989) showed that the corrected score (2.8) does not exist for logistic regression. Huang & Wang (2001) observed that the correction is not limited to the original score (2.3) only. They argued that the corrections can be performed on a class of positive weighted estimating function function

$$S_w(\theta) = w(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X)\{Y - F(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X)\}\begin{pmatrix} 1 \\ Z \\ X \end{pmatrix}, \tag{2.9}$$

where $w(\cdot)$ is some positive weight. It is clear that $S_w(\theta)$ satisfies $E_{\theta_0}[S_w(\theta)] = 0$ only at $\theta = \theta_0$ as $S(\theta)$ does. Huang & Wang (2001) chose a pair of weights, $1+\exp(-\alpha-\beta_z^{\mathrm{T}} Z-\beta_x^{\mathrm{T}} X)$

and $1 + \exp(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X)$, to form a pair of correction-amenable estimating functions:

$$\Phi_{-}(\theta) = \{Y - 1 + Y\exp(-\alpha - \beta_z^{\mathrm{T}} Z - \beta_x^{\mathrm{T}} X)\}\begin{pmatrix} 1 \\ Z \\ X \end{pmatrix}, \tag{2.10}$$

$$\Phi_{+}(\theta) = \{Y + (Y - 1)\exp(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X)\}\begin{pmatrix} 1 \\ Z \\ X \end{pmatrix}. \tag{2.11}$$

(2.10) and (2.11) are served as bases on which all the corrections are performed. One of the advantages of choosing these two particular weights is that recoding the outcome event leads to the opposite sign in the coefficients, a property of the original score function.

Under the additive measurement error assumption, the parametric-correction estimation procedure based on (2.10) and (2.11) includes two pairs of estimating functions

$$\begin{aligned} \overline{\Phi}_{-}(\theta) \;=\; & (Y - 1)\begin{pmatrix} 1 \\ Z \\ W - E(U) \end{pmatrix} + \\ & \frac{Y\exp(-\alpha - \beta_z^{\mathrm{T}} Z - \beta_x^{\mathrm{T}} W)}{E(\exp(-\beta_x^{\mathrm{T}} U))}\begin{pmatrix} 1 \\ Z \\ W - \dfrac{E(\exp(-\beta_x^{\mathrm{T}} U) U)}{E(\exp(-\beta_x^{\mathrm{T}} U))} \end{pmatrix}, \end{aligned} \tag{2.12}$$

$$\overline{\Phi}_{+}(\theta) = Y\begin{pmatrix} 1 \\ Z \\ W - E(U) \end{pmatrix} + \frac{(Y-1)\exp(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} W)}{E(\exp(\beta_x^{\mathrm{T}} U))}\begin{pmatrix} 1 \\ Z \\ W - \dfrac{E(\exp(\beta_x^{\mathrm{T}} U)U)}{E(\exp(\beta_x^{\mathrm{T}} U))} \end{pmatrix}. \tag{2.13}$$

More flexible than the conditional score, the parametric-correction estimation procedure does not require the normal error assumption. When $U \sim N(0, \Sigma_{uu})$, $E(\exp(\beta_x^{\mathrm{T}} U)) = \exp(\beta_x^{\mathrm{T}} \Sigma_{uu} \beta_x / 2)$ and $E(\exp(\beta_x^{\mathrm{T}} U)U) = \exp(\beta_x^{\mathrm{T}} \Sigma_{uu} \beta_x / 2)\Sigma_{uu}\beta_x$. Then (2.12) and (2.13) reduces to

$$\overline{\Phi}_{-}(\theta) = (Y-1)\begin{pmatrix} 1 \\ Z \\ W \end{pmatrix} + Y\exp(-\alpha - \beta_z^{\mathrm{T}} Z - \beta_x^{\mathrm{T}} W - \beta_x^{\mathrm{T}} \Sigma_{uu} \beta_x / 2)\begin{pmatrix} 1 \\ Z \\ W + \Sigma_{uu}\beta_x \end{pmatrix}, \tag{2.14}$$

$$\overline{\Phi}_{+}(\theta) = Y\begin{pmatrix} 1 \\ Z \\ W \end{pmatrix} + (Y - 1)\exp(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} W - \beta_x^{\mathrm{T}} \Sigma_{uu} \beta_x / 2)\begin{pmatrix} 1 \\ Z \\ W - \Sigma_{uu}\beta_x \end{pmatrix}. \tag{2.15}$$

Since there are more estimating functions than parameters, Huang & Wang (2001) suggested using the generalized method-of-moments (GMM) method to combine $\overline{\Phi}_{-}(\theta)$ and $\overline{\Phi}_{+}(\theta)$ to obtain the consistent estimator that is efficient in the class of estimators based on $\overline{\Phi}_{-}(\theta)$ and $\overline{\Phi}_{+}(\theta)$. According to their results, the parametric-correction estimation procedure estimator has a very good efficiency performance that is almost compared to the asymptotically locally efficient conditional score estimator (Stefanski & Carroll 1987), with or without measurement error. Its finite-sample performance has been investigated through an extensive simulation and the results show that it is comparable to the consistent conditional score estimator in terms of bias and efficiency.

2.2 Multiple Roots

2.2.1 Introduction

In point estimation, consistent estimators are often obtained by solving unbiased estimating functions. Crowder (1986) proves that an unbiased estimating function is guaranteed to have a consistent root under mild regularity. However, Crowder's theory does not rule out the possibility that an unbiased estimating function could have inconsistent roots. in finite samples , multiple solutions of an unbiased estimating function are often encountered. If solutions are far apart, distinguishing among them could be a major challenge in applications. In this section, we give a brief review of the mechanisms of multiple roots in the unbiased estimating function approach.

2.2.2 Unbiased estimating functions and multiple roots

Let $y_1, \ldots, y_n$ be observations of random variables with density function $f(y;\theta)$, where $\theta \in \Theta$, a compact parameter space. The estimates of the true value θ_0 are typically obtained by solving

$$\sum_{i=1}^{n} g(y_i, \hat{\theta}) = 0, \tag{2.16}$$

where $g(y_i, \theta)$ is some unbiased estimating function satisfying

$$E_{\theta_0}[g(Y, \theta)] = 0. \tag{2.17}$$

The estimator $\hat{\theta}$ is often referred as M-estimator. Under mild regularity, a consistent $\hat{\theta}$ exists, i.e., $\hat{\theta} \xrightarrow{p} \theta_0$. Nevertheless, the definition of unbiased estimating functions in (2.17) does not rule out the possible existence of inconsistent roots .

According to Small & Wang (2003), $g(Y, \theta)$ can be categorized into two classes: the class of regular unbiased estimating functions that satisfies (2.17) if and only if $\theta = \theta_0$, and the class of irregular unbiased estimating functions that may have

$$E_{\theta_0}[g(Y,\theta_1)] = 0, \quad \text{for some} \quad \theta_1 \neq \theta_0. \tag{2.18}$$

The class of regular unbiased estimating functions has no inconsistent roots . For example, the score functions of exponential families are regular. On the other hand, the class of irregular unbiased estimating functions admit inconsistent roots and the multiple roots arise in large samples. For example, the conditional score (Stefanski & Carroll 1987) is irregular .

Even though regular unbiased estimating functions do not suffer from multiple roots in large samples, they can have multiple solutions in finite samples. That is, the root advantage of regular unbiased estimating functions in large samples becomes vague in finite samples. To see when an estimating function can have multiple roots in finite samples, we should study the Hessian matrix. Let

$$\dot{g}(\theta) = \frac{\partial g(\theta)}{\partial \theta^{\mathrm{T}}}$$

be the Hessian matrix. When $\hat{\theta}$ exists , its uniqueness is guaranteed if $\dot{g}(\theta)$ is negative definite for all $\theta \in \Theta$. This is generally the case for the exponential family score function. Within exponential family, the loglikelihood is strictly concave for all $\theta \in \Theta$. As a result, the MLE of θ uniquely exists in Θ. However, for most of the regular unbiased estimating functions, the Hessian matrix is not always negative definite and attentions should be paid in finite samples root searchings since there may be more than one solution satisfying $\sum_{i=1}^{n} g(\hat{\theta}) = 0$ and some of them may not be consistent. For examples, the corrected-type estimating functions (Nakamura 1990, Stefanski 1989, Huang & Wang 2001) are regular unbiased estimating functions whose Hessian matrixes are not always negative definite. As a result, the multiple roots could arise in finite samples, even though they would become arbitrarily close to each other for sufficiently large sample sizes. For irregular estimating functions, it is clear that the Hessian can not be always negative definite by the fact that irregular estimating functions admit inconsistent roots even for large sample.

In practice, a major task is to distinguish among multiple roots in finite samples. The root uniqueness only happen to special cases. In the presence of multiple solutions, a good strategy should be able to distinguish good solutions from bad solutions.

2.2.3 Choosing from multiple roots

When $\dot{g}(\theta)$ fails to be negative definite for all θ, a rational solution is to find all the possible roots and compare them to select the good one. There are two popular approaches to detect all the roots. The first approach is to search for roots from random starting points (RSPs) (Robbins 1968, Thode et al. 1987, Finch et al. 1989). The basic idea is to start from r RSPs to search for roots. When r is big enough, the probability of finding a new root not found by the first $r-1$ RSPs will be very small. Another approach is a bootstrap root search proposed by Markatou, Basu & Lindsay (1998) in their weighted likelihood equations analysis. The idea is that all the possible roots are supported by some subsets of the data . To implement this idea, one need to draw $m \leq n$ bootstrap samples from the original data. For each bootstrap sample, the root is found by solving the unbiased estimating function based on that bootstrap sample. For big enough m, all the possible roots would be found. Both approaches have their own merits and are good methods to detect multiple roots in practice.

If $\dot{g}(\theta)$ is symmetric,i.e.,

$$\frac{\partial g_i(y,\theta)}{\partial \theta_j^{\mathrm{T}}} = \frac{\partial g_j(y,\theta)}{\partial \theta_i^{\mathrm{T}}}, \tag{2.19}$$

for $i \neq j$, the vector field of $g(Y,\theta)$ is conservative and there exists an objective function that has $g(Y,\theta)$ as its derivative. In this case, different roots can be compared at their corresponding objective function values. Likelihood score function $\mathrm{S}(\theta)$ satisfies (2.19) and therefore all the stationary points satisfying $\sum_{i=1}^{n} \mathrm{S}(\hat{\theta}) = 0$ can be compared at their loglikelihood values. Huzurbazar (1948) proved that the consistent root of score function is asymptotically unique and corresponds to a local maximum of the likelihood function. Another example is the trivial case where θ is a scalar. Then $g(Y, \theta)$ is always conservative. Therefore we can always intergrade $g(Y, \theta)$ to a scalar objective function and compare the roots. Unfortunately, in general, unbiased estimating functions do not have the property of (2.19) and can not be expressed as derivatives of some scalar objective functions. Therefore, the distinguishing among roots is problematic.

When the explicit forms of roots are available, the roots can be compared with respect to their asymptotic properties (Heyde & Morton 1998). Their methods suggest choosing the correct root $\hat{\theta}$ that is:

1. $\sqrt{n}$-consistent;

2. $\dot{g}(\theta)$ behaves asymptotically as $E_{\theta}[\dot{g}(\theta)]$ at $\theta = \hat{\theta}$;

3. using a least square or goodness-of-fit criterion to select the best root.

Their approaches are useful if the analytic formulas of roots can be derived, which are often not the case for nonlinear models. For example, for logistic regression with errors-in-covariates.

Gan & Jiang (1999) and Biernacki (2005) constructed a test statistics to test for global maximum of the likelihood to choose among multiple roots of a likelihood score function. Some other root selection methods include iterating from consistent estimators (Rao 1973, Lehmann 1983), bootstrap method and selecting roots based on information criterion. Those approaches, however, are not very useful for logistic regression with errors-in-covariates.

Among all the methods to distinguish among multiple roots, building an objective function appears to be attractive and draws great attention and has been applied to distinguish among roots of the conditional score (Hanfelt & Liang 1997). We will briefly review the relevant methodologies in the next section.

2.2.4 Artificial likelihood functions

A big drawback of general unbiased estimating functions compared to likelihood score function is that in general there does not exist a scalar objective function that has derivative as the estimating function as score function does. In the case that score function has multiple roots, the relative likelihood principle would prefer a root having a bigger likelihood than others. On the other hand, it is not clear how to distinguish among multiple roots for a general unbiased estimating function that is not derived by maximizing or minimizing some objective functions.

Recently attentions have been turn to construct objective functions based on optimal estimating functions . The term optimal estimating function refer to the one whose consistent root has the smallest asymptotical variance among the class of estimating functions. McCullagh & Nelder (1989) shows that the optimal estimating function based on $g(Y, \theta)$ takes the form

$$U(Y, \theta) = D^{\mathrm{T}} V^{-1} g(Y, \theta) \tag{2.20}$$

where

$$D = E[-\dot{g}(Y, \theta)], \quad V = \mathrm{Cov}[g(Y, \theta)].$$

The asymptotic variance of estimator $\hat{\theta}$ is $D^{\mathrm{T}} V^{-1} D$. (2.20) reduces to the score function if $g(Y, \theta)$ is the score function. Also, when $g(Y, \theta) = Y - \mu(\theta)$, (2.20) reduces to the quasi-score:

$$q(Y, \theta) = D^{\mathrm{T}} V^{-1} (Y - \mu(\theta)). \tag{2.21}$$

(2.20) shares many nice properties with the score function. For example, it is unbiased and information unbiased.

Wedderburn (1974) obtains the quasi-likelihood function by performing line integral on quasi-score (2.21)

$$Q(\theta, y) = \int_{y}^{\mu} V^{-1}(y - t)\, \mathrm{d}t. \tag{2.22}$$

The quasi-likelihood $Q(\theta, y)$ should behave like a real loglikelihood at least in a neighborhood of the true value of θ. Then, different roots of quasi-score $q(Y, \theta)$ can be compared using the quasi-likelihood ratio (Wedderburn 1974)

$$Q(\theta, \eta) = \int_{\mu_\eta}^{\mu_\theta} V^{-1}(y - t)\, \mathrm{d}t, \tag{2.23}$$

which is analogous to the real loglikelihood ratio. If the variance component V is a diagonal matrix, the line integral in (2.22) is path-independent. Therefore, $Q(\theta, y)$ is unique and has the quasi-score as its gradient vector. However, in general, V is not diagonal and the derivative matrix of quasi-score is not symmetric such as

$$\frac{\partial q_i}{\partial \theta_j^{\mathrm{T}}} \neq \frac{\partial q_j}{\partial \theta_i^{\mathrm{T}}} \qquad \text{for} \quad i \neq j.$$

Therefore, the vector field of quasi-score is in general not conservative and the line integral in (2.22) is path-dependent. Consequently, no quasi-likelihood $Q(\theta, y)$ that has quasi-score as its gradient vector exists. Therefore, the quasi-score estimators can not be viewed as the stationary points of scalar objective functions and using quasi-likelihood to distinguish among multiple roots would be problematic when $q(Y, \theta)$ is not conservative.

Notice that optimal estimating function (2.20) has

$$E[\dot{U}(Y, \theta)] = D^{\mathrm{T}} V^{-1} D,$$

where $\dot{U}(Y, \theta) = \partial U / \partial \theta^{\mathrm{T}}$. That is, even though $\dot{U}(Y, \theta)$ may not be symmetric, $E[\dot{U}(Y, \theta)]$ is always symmetric. Therefore, optimal estimating function is E-conservative. in large samples, it should be very close to conservative. Therefore, an approximate objective function can be constructed by performing a path-dependent line integral on optimal estimating function. Hanfelt & Liang (1995) generalizes the quasi-likelihood to consider other optimal estimating functions. Hanfelt & Liang (1995, 1997) uses the artificial likelihood as an objective function to distinguish among multiple roots in the conditional score approach (Stefanski & Carroll 1987).

Artificial likelihoods may also be constructed through projection. Li (1993) and Hanfelt & Liang (1995) suggested projecting the centred likelihood ratio

$$R(\theta, \eta) = \frac{L(\theta)}{L(\eta)} - 1$$

into a space spanned by the unbiased elementary estimating function $g(Y, \eta)$:

$$\mathscr{L}_\eta = \{r = a^t g(y, \eta) + b\},$$

where r is square integrable. The resulting projection only depends on the first two moments of g:

$$D_\eta(\theta, \eta) = \{C(\theta, \eta)\}^t {V_\eta}^{-1} g(y, \eta),$$

where $C(\theta, \eta) = E_\theta\{g(y, \eta)\}$ and $V_\eta = \mathrm{Cov}_\eta\{g(y, \eta)\}$. Similarly, one can define the reverse projection:

$$D_\theta(\eta, \theta) = \{C(\eta, \theta)\}^t {V_\theta}^{-1} g(y, \theta).$$

The generalized linear projected likelihood ratio (Li 1993, Hanfelt & Liang 1995) takes the form

$$\begin{aligned} D(\theta, \eta) &= \frac{1}{2} D_\theta(\eta, \theta) - \frac{1}{2} D_\eta(\theta, \eta) \\ &= \frac{1}{2}\{C(\theta, \eta)\}^t {V_\eta}^{-1} g(y, \eta) - \frac{1}{2}\{C(\eta, \theta)\}^t {V_\theta}^{-1} g(y, \theta), \end{aligned}$$

which approximates the true loglikelihood ratio between θ and η. Li (1993) considered a special case that takes $g = y - \mu$ and constructed a projected likelihood ratio for the quasi-score.

Other methods to build artificial likelihood as objective function include McCullagh

(1991), Li & McCullagh (1994) and Small & Wang (2003), among others.

2.3 Empirical Likelihood

2.3.1 Overview

The parametric likelihood method is the most widely used technique in statistical inference. When the model is specified correctly, the Wilks's theorem (Wilks 1938) states that the log likelihood ratio statistics follows an asymptotic chisquare distribution. This result can be used to perform hypothesis testings and construct likelihood ratio based confidence intervals. Empirical likelihood (EL), introduced by Owen (1988, 1990, 1991), offers likelihood ratio statistics that shares many properties with the parametric parallel without making any distribution assumptions.

Let $x_1, x_2, \ldots, x_n$ be a random sample from an d -variate unknown distribution function F with mean μ_0 and finite covariance matrix Σ of rank $r > 0$. The empirical likelihood is defined as the nonparametric likelihood of the distribution function F

$$L(F) = \prod_{i=1}^{n} dF(x_i) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p_i,$$

which is maximized by the empirical distribution function of F

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

The empirical likelihood ratio is defined as

$$\begin{aligned} R(F) &= \frac{L(F)}{L(F_n)} \\ &= \prod_{i=1}^{n} np_i \end{aligned}$$

Suppose that we are interested in making inference on some parameter $\theta = \theta(F)$. For example, μ of F. The profile empirical likelihood ratio function for μ is

$$l_E(\mu) = \sup\left\{\prod_{i=1}^{n} np_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x_i = \mu\right\}.$$

$l_E(\mu)$ is uniquely well defined provided that μ is in the convex hull of $\{x_i, i = 1, \ldots, n\}$. Owen (1988, 1990) showed that the Wilks's theorem (Wilks 1938) holds for $l_E(\mu_0)$. That is, under regularity conditions,

$$-2l_E(\mu_0) \xrightarrow{d} \chi_r^2 \quad \text{as} \quad n \to \infty,$$

which parallels the results for the parametric likelihood in Wilks (1938). This asymptotic result allows us to test $H_0 : \mu = \mu_0$ and construct empirical likelihood confidence region for the mean. An approximate α-level confidence region for μ is given by

$$\Re_c = \{\mu : -2l_E(\mu) \leq \chi_r^2(1-\alpha)\}.$$

The accuracy of this empirical likelihood ratio confidence region is of order n^{-1}:

$$P(\mu_0 \in \Re_c) = 1 - \alpha + O(n^{-1}),$$

which is the same as the accuracy of the parametric likelihood ratio confidence regions. Similar asymptotic results can be extended to more general smooth functions of means (Owen 2001).

2.3.2 EL for estimating equations

Qin & Lawless (1994) extended the methods by Owen (1988, 1990, 1991) to link the empirical likelihood and estimating equations together to develop methods of combining parameter information through unbiased estimating functions. Let $\theta \in \mathbb{R}^p$ be the parameter of interest of the population F, and the information about θ can be summarized in terms of functional independent unbiased estimating functions $g(x, \theta) \in \mathbb{R}^s$

$$E_F\{g(x, \theta)\} = 0.$$

When $s = p$, the true value θ_0 is usually estimated by solving

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i, \hat{\theta}) = 0$$

for $\hat{\theta}$. Under appropriate regularity conditions, an $\sqrt{n}$-consistent $\hat{\theta}$ exists.

However, when $s > p$, one may not be able to find a solution that solves the unbiased estimating equations. To obtain a consistent point estimator for this over-determined case, Qin & Lawless (1994) developed method to combine estimating equations through empirical likelihood. The profile empirical likelihood function for θ is defined as

$$L(\theta) = \sup\left\{\prod_{i=1}^{n} p_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(x_i, \theta) = 0\right\}.$$

A unique maximum of $L(\theta)$ exists provided that 0 is in the convex hull of $\{g(x_i, \theta), i = 1, \ldots, n\}$.

To obtain the estimator, one needs to maximize the empirical likelihood function $L(\theta)$. It is generally more convenient to maximize the logarithm of the empirical likelihood: $\sum_{i=1}^{n} \log(p_i)$. This constrained optimization process can be conducted using Lagrange multipliers. Let

$$\mathcal{L} = \sum_{i=1}^{n} \log(p_i) + \gamma(1 - \sum_{i=1}^{n} p_i) - n\lambda^{\mathrm{T}} \sum_{i=1}^{n} p_i g(x_i, \theta),$$

where γ and $\lambda = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ are Lagrange multipliers. Straightforward calculous yields

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_i} &= \frac{1}{p_i} - \gamma - n\lambda^{\mathrm{T}} g(x_i, \theta) = 0, \\
\sum_{i=1}^{n} p_i \frac{\partial \mathcal{L}}{\partial p_i} &= n - \gamma = 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\gamma &= n \\
p_i &= \frac{1}{n(1 + \lambda^{\mathrm{T}} g(x_i, \theta))},
\end{aligned}$$

where the Lagrange multiplier $\lambda = \lambda(\theta) \in \mathbb{R}^s$ solves

$$\frac{1}{n} \sum_{i=1}^{n} \frac{g(x_i, \theta)}{1 + \lambda^{\mathrm{T}} g(x_i, \theta)} = 0.$$

It is understood that $\prod_{i=1}^{n} p_i$ is maximized at $p_i = n^{-1}$ if there are no parametric constraints. Therefore, the empirical loglikelihood ratio for θ is

$$l_E(\theta) = -\sum_{i=1}^{n} \log\{1 + \lambda^{\mathrm{T}} g(x_i, \theta)\}.$$

The maximum empirical likelihood estimator (MELE) $\tilde{\theta}$ of θ_0 can be obtained by maximizing $l_E(\theta)$. The maximization involves two stages. At the first stage, one minimize $l_E(\theta)$ with respect to λ for a given θ. Then $l_E(\theta)$ is maximized with respect to θ. That is

$$\tilde{\theta} = \arg\max_{\theta \in \mathbb{R}^p} l_E(\theta) = \arg\max_{\theta \in \mathbb{R}^p} \min_{\lambda \in \mathbb{R}^s} -\sum_{i=1}^{n} \log(1 + \lambda^{\mathrm{T}} g(x_i, \theta)). \tag{2.24}$$

Therefore the estimator for p_i is

$$\tilde{p}_i = \frac{1}{n(1 + \tilde{\lambda}^{\mathrm{T}} g(x_i, \tilde{\theta}))},$$

and the empirical likelihood estimator for F is

$$\tilde{F}_n(x) = \sum_{i=1}^{n} \tilde{p}_i I(x_i < x).$$

Notice that if we set $g(x_i, \mu) = x_i - \mu$, the above results can be applied to the empirical likelihood for the mean μ automatically. When $s = p$, the MELE $\tilde{\theta} = \hat{\theta}$, the solution to $n^{-1} \sum_{i=1}^{n} g(x_i, \hat{\theta}) = 0$ and $\tilde{p}_i = n^{-1}$. In other words, the empirical likelihood would be maximized at zero-crossings of estimating functions. As a special case, the MELE for μ_0 is $\tilde{\mu} = \hat{\mu} = \bar{x} = n^{-1} \sum_{i=1}^{n} x_i$, the sample mean.

Chapter 3

Root Behaviors

3.1 Introduction

The multiple-roots problem of the conditional score in large samples has been reported by many researchers (Stefanski & Carroll 1987, Hanfelt & Liang 1995, 1997). Nevertheless, whether or not multiple solutions arise frequently in practice is still not very clear. The seriousness of multiple roots in practice have not been investigated adequately.

Compared with the conditional score for logistic regression with errors-in-covariates, the parametric-correction estimation procedure (Huang & Wang 2001) does not have multiple-roots problem in large samples. However, similar to the corrected score (Nakamura 1990), they can also have multiple solutions in finite samples. So far, limited research has been done to study the multiple solutions in finite samples for corrected-type estimating functions either.

Indeed, what really matters is the multiple-roots phenomena in finite samples. We believe that investigations on root behaviors are needed to better understand both approaches in finite samples. In this chapter, We study the finite-sample root behaviors of each approach through graphic illustrations with a focus on the conditional score.

3.2 Root Behaviors in finite samples

For simplicity, we consider a logistic regression model with single covariate $X \sim N(0, 1)$ that is measured with errors. The true values for (α_0, β_0) are (0,1). The measurement error $U \sim N(0, \sigma_u^2)$. For the conditional score (5.2), we choose $t(\Delta) = W + (Y - 1/2)\sigma_u^2\beta_x$. To investigate the influences of the magnitude of the measurement error and sample size on the appearance of the multiple roots, we consider four scenarios: (a) $n = 200$ and $\sigma_u^2 = 0.5$, (b) $n = 500$ and $\sigma_u^2 = 0.5$, (c) $n = 200$ and $\sigma_u^2 = 1$, and (d) $n = 500$ and $\sigma_u^2 = 1$. In the simulated data, the naive estimators for these four scenarios are 0.58, 0.86, 0.38, and 0.70, respectively. We plot out the conditional score in those four scenarios. For illustration purpose, we profile the intercept out and only draw the plots with respect to slope β only. The range for β we considered is [-8, 8], which is adequate for our purposes.

Figure 3.1 shows that when the variance of the measurement error is half of the variance of the true covariate, the conditional score performs well and generate a single root at both sample sizes. As the measurement error increases, the multiple roots appear. When the measurement error and the true covariate have the same variance, three and two roots were found for $n = 200$ and $n = 500$, respectively. Therefore, both the the magnitude of the measurement error and the sample size play roles on the appearance of multiple roots. As the measurement decreases or the sample size increases, the multiple roots would be less likely to occur. Table 3.1 shows the mean number roots for each scenario based on 1000 simulations. It appears that the magnitude of the measurement error has a bigger influence on the appearance of the multiple roots than the sample size. When the measurement error and the true covariate have equal variances, the conditional score usually generates multiple roots even at a large sample size of 500. Plot (c) of Figure 3.1 represents a typical three-roots scenario for a single covariate model with large measurement errors and small sample size. The distribution of roots is usually: a correct root at the middle, two false roots at each side of the correct root. And three roots are far apart. Similar phenomena was found in the blood pressure study (Pan et al. 1990). The conditional score yields three

Table 3.1: *The mean and range of number of roots of the conditional score in a single and error-prone covariate logistic regression model*

Scenario	Size	σ_u^2	Mean	Range
a	200	0.5	1.14	1-2
b	500	0.5	1.01	1-2
c	200	1.0	2.94	1-3
d	500	1.0	2.29	1-3

Note: $X \sim N(0, 1)$. Measurement error $U \sim N(0, \sigma_u^2)$. Results are based on 1000 replications.

roots for the unobservable true urinary sodium chloride : -4.36, 0.61, and 6.63. Out of the three roots, the root 0.61 appears to be the correct one (Hanfelt & Liang 1997). By using replicated measurements to reduce the measurement error, a single root (0.65) is generated by the conditional score (Hanfelt & Liang 1997).

Our findings are based on a single covariate model. If there are multiple covariates measured with errors, the multiple roots problem of the conditional score would be more severe (Carroll et al. 2006). Hanfelt & Liang (1997) found an average of seven conditional score roots in a simulation study of multiple covariates subject to error logistic model.

Our study has confirmed that the existence of the multiple roots for the conditional score. Typically, those roots are far apart and how to distinguish from them are of challenged. When the measurement error is large, the multiple roots are quite frequent. In the case of (c) in Figure 3.1, 3 roots were found in 983 out of 1000 simulations.

The parametric-correction estimation procedure (Huang & Wang 2001) contains two corrected-type estimating functions $\overline{\Phi}_{-}(\theta)$ (2.14) and $\overline{\Phi}_{+}(\theta)$ (2.15). To investigate their root behaviors in finite samples, we use the same set up and data as used in Figure 3.1. to obtain the profiled plots for both $\overline{\Phi}_{-}(\theta)$ (Figure 3.2) and $\overline{\Phi}_{+}(\theta)$ (Figure 3.3).

As seen, both $\overline{\Phi}_{-}(\theta)$ and $\overline{\Phi}_{+}(\theta)$ experiences multiple roots in finite samples. It appears that their multiple roots are even severer than the conditional score in finite samples. For the plots (d) in Figure 3.2 and (c) and (d) in Figure 3.3 , both $\overline{\Phi}_{-}(\theta)$ and $\overline{\Phi}_{+}(\theta)$ fail to generate a root reasonably close to the truth. Even though they do not have multiple roots in large samples, this theoretical advantage may not be of significantly meaningful in practice. In stead of searching roots of either $\overline{\Phi}_{-}(\theta)$ or $\overline{\Phi}_{+}(\theta)$ that might not behave well enough in finite samples, Huang & Wang (2001) successfully overcame the dilemma by combining $\overline{\Phi}_{-}(\theta)$ and $\overline{\Phi}_{+}(\theta)$ together as an objective function to be minimized, which performs well in finite samples.

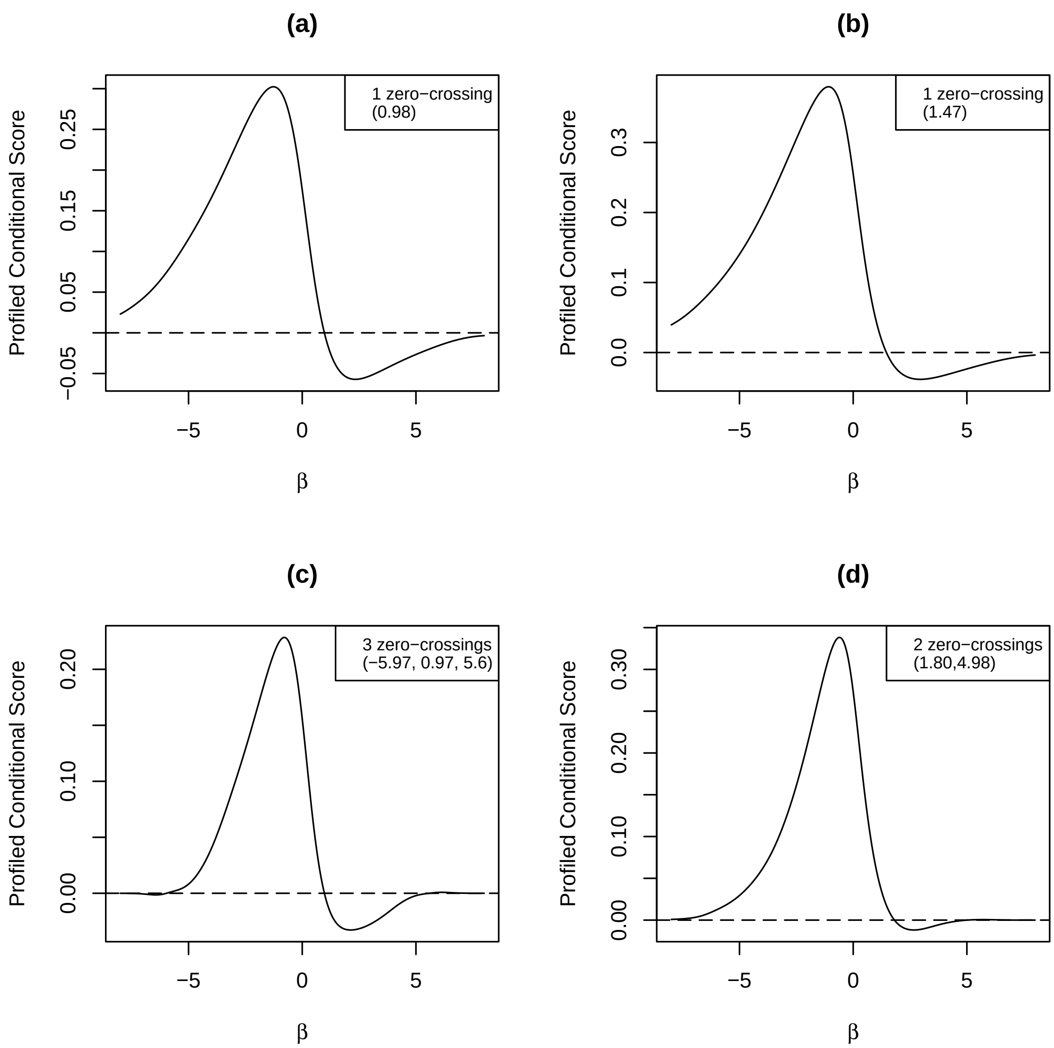

Figure 3.1: *Multiple roots of conditional score of logistic regression model with single covariate measured with error.* $(\alpha_0, \beta_0) = (0, 1)$. *The true covariate follows a standard normal distribution. The measurement error is normal distributed with mean zero and variance* σ_u^2.*(a)* $n = 200$ *and* $\sigma_u^2 = 0.5$ *(b)* $n = 500$ *and* $\sigma_u^2 = 0.5$ *(c)* $n = 200$ *and* $\sigma_u^2 = 1$ *(d)* $n = 500$ *and* $\sigma_u^2 = 1$

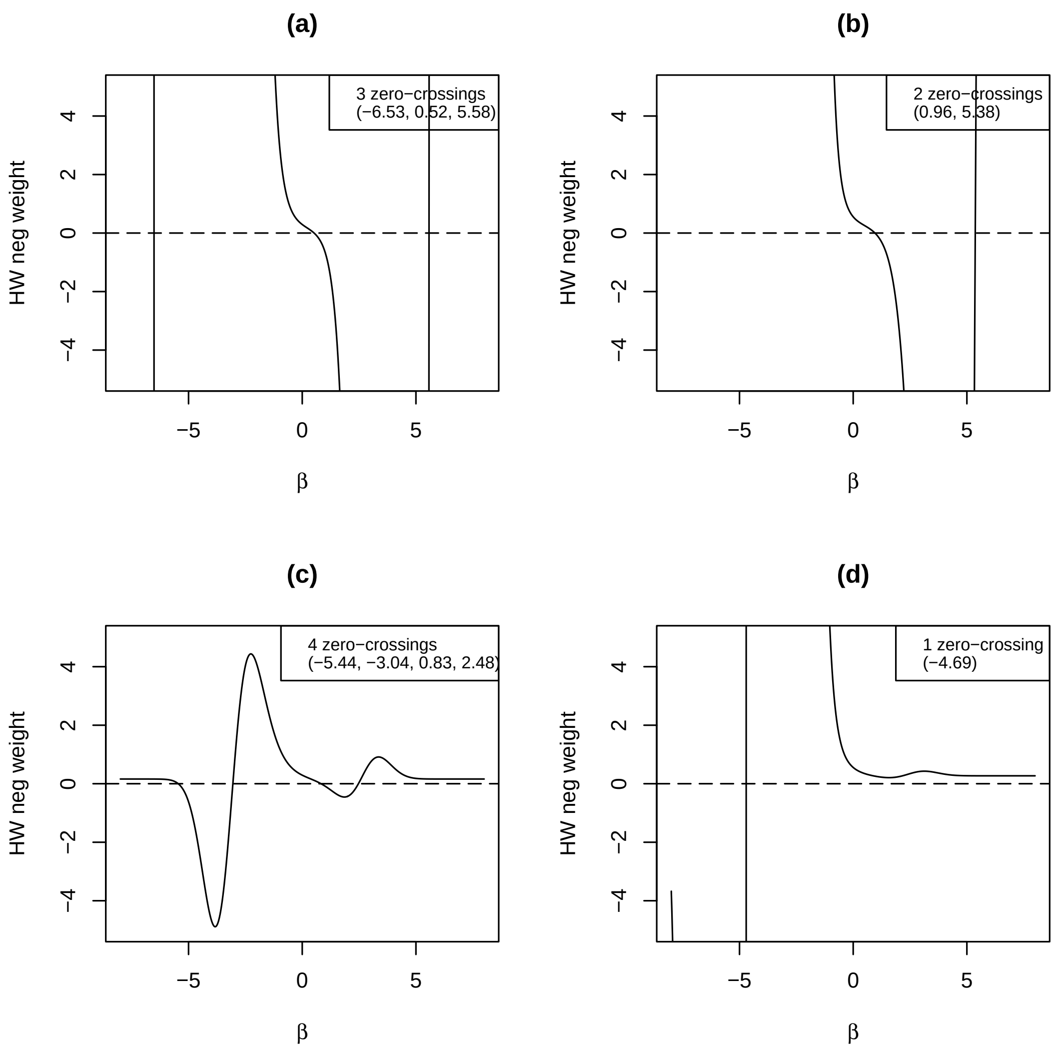

Figure 3.2: *Multiple roots of* $\overline{\Phi}_{-}(\theta)$ *of logistic regression model with single covariate measured with error. In the plot, "HW neg Weight" stands for* $\overline{\Phi}_{-}(\theta)$*.* $(\alpha_0, \beta_0) = (0, 1)$*. The true covariate follows a standard normal distribution. The measurement error is normal distributed with mean zero and variance* σ_u^2*.(a)* $n = 200$ *and* $\sigma_u^2 = 0.5$ *(b)* $n = 500$ *and* $\sigma_u^2 = 0.5$ *(c)* $n = 200$ *and* $\sigma_u^2 = 1$ *(d)* $n = 500$ *and* $\sigma_u^2 = 1$

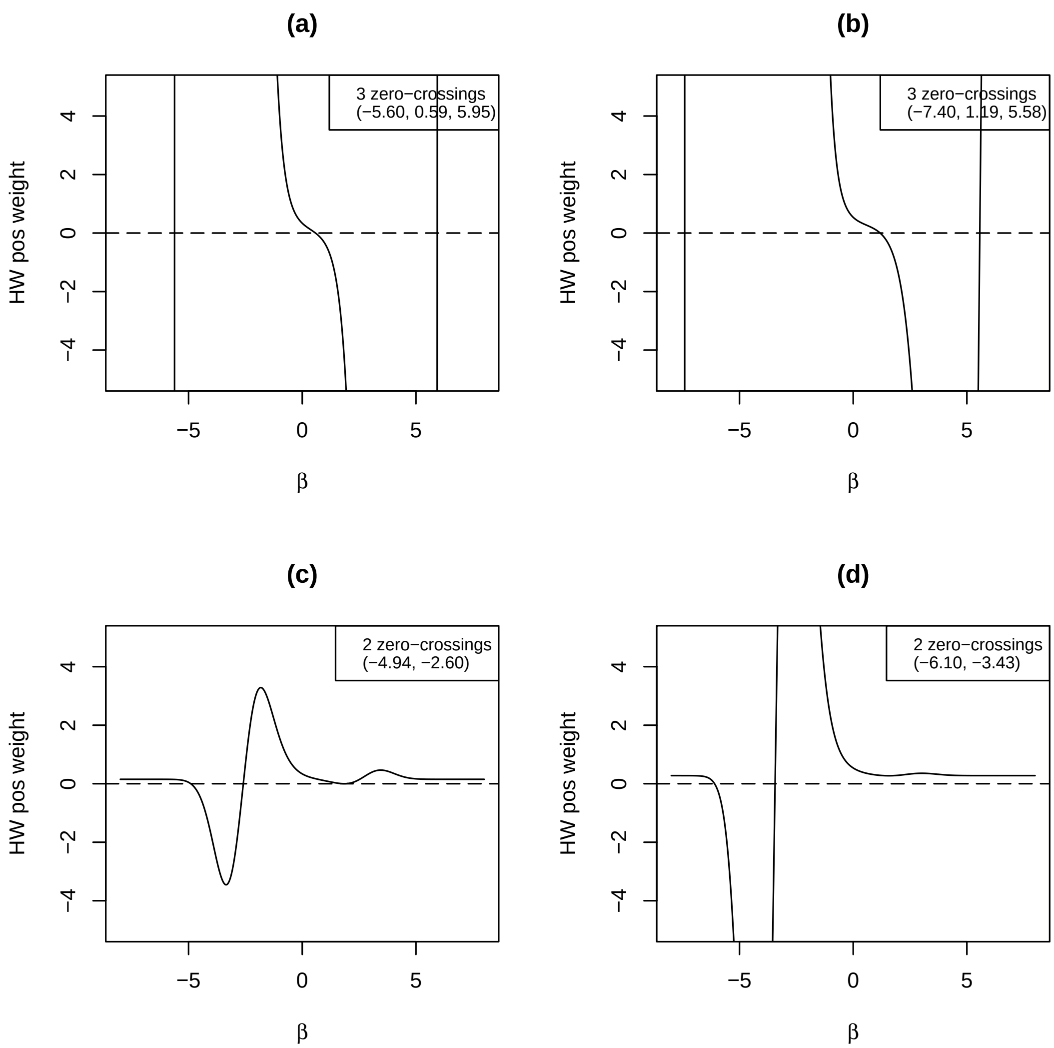

Figure 3.3: *Multiple roots of* $\overline{\Phi}_{+}(\theta)$ *of logistic regression model with single covariate measured with error. In the plot, "HW pos Weight" stands for* $\overline{\Phi}_{+}(\theta)$*.* $(\alpha_0, \beta_0) = (0, 1)$*. The true covariate follows a standard normal distribution. The measurement error is normal distributed with mean zero and variance* σ_u^2*.(a)* $n = 200$ *and* $\sigma_u^2 = 0.5$ *(b)* $n = 500$ *and* $\sigma_u^2 = 0.5$ *(c)* $n = 200$ *and* $\sigma_u^2 = 1$ *(d)* $n = 500$ *and* $\sigma_u^2 = 1$

3.3 Discussions

The study shows that both the conditional score and the corrected-type estimating functions ((2.14) and (2.15)) can have multiple solutions in finite samples, especially when the measurement error is large. The corrected-type estimating functions have theoretical advantage over the conditional score in terms of root consistency in large samples. However, in finite samples, it appears the conditional score has better root behaviors than the existing corrected-type estimating functions, in the logistic regression case.

In the presence of multiple solutions, both Stefanski & Carroll (1987) and Nakamura (1990) suggest an heuristic procedure iterating from the naive estimator to obtain a good root in the numerical practice. This approach appears to work well in practice when the measurement error is small. However, when the measurement error is large, this approach may break down. The plots (d) in Figure 3.2 and (c) and (d) in Figure 3.3 indicates that there are no roots reasonably close to the truth. Conditional score also has this ill-behaved no-good-root scenario, but less frequent. Figure 3.4 shows an example that the conditional score generate a single root that is far from the truth. If one performs a thorough root search to find all the roots, then the heuristic procedure will lead to a false root. However, this no-good-root scenario is rare for the conditional score. Table 3.2 summarizes the percentages of no-good-root scenario for cases considered in Figure 3.1. As seen, Only 1.7% of the samples do not generate a good root for a single case. As the sample size goes larger or the measurement error becomes smaller, this no-good-root scenario would disappear.

Table 3.2: *The percentages of no-good-root scenario for the conditional score in a single and error-prone covariate logistic regression model*

Size	σ_u^2	Percentage
200	0.5	0
200	1.0	1.7%
500	0.5	0
500	1.0	0

Note: $X \sim N(0,1)$. Measurement error $U \sim N(0,\sigma_u^2)$. Results are based on 1000 replications.

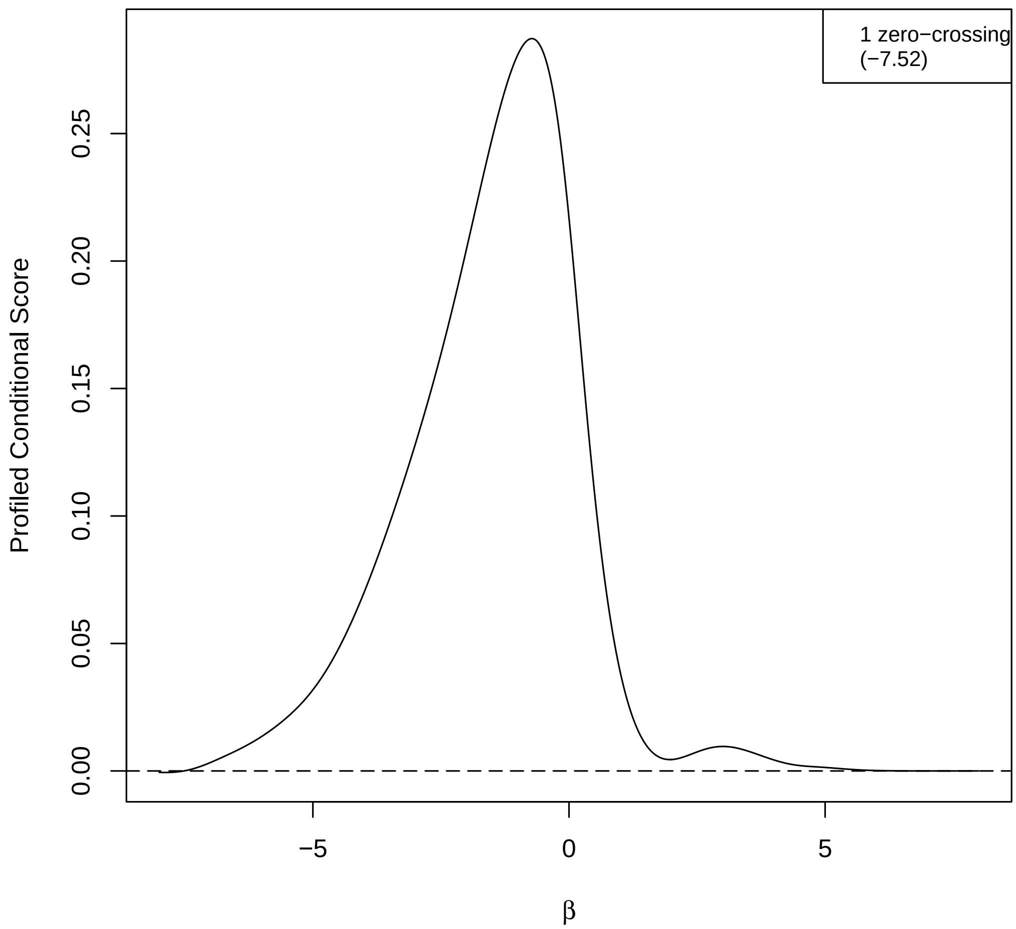

Figure 3.4: *An example of no roots around the truth or naive estimator for logistic regression model with single covariate measured with error. N=200.* $(\alpha_0, \beta_0) = (0, 1)$. *Both the true covariate and the additive measurement error follow a standard normal distribution.* $(\hat{\beta}_{Naive} = 0.54)$

Chapter 4

The combined estimation procedure

4.1 Introduction

To resolve the multiple-roots problem of the conditional score, one could combine the conditional score with the corrected-type estimating functions ((2.14) and (2.15)) to eliminate inconsistent roots of the conditional score in large samples. However, by doing so, we need to estimate P-dimensional parameters using $3P$-dimensional estimating functions. Our simulation shows that the numerical difficulties arise, especially under small or medium sized samples.

In this chapter, we develop a new corrected-type estimating function named the weighted-correction estimating function for logistic regression with errors-in-covariates. Its asymptotical properties and finite-sample performances will be presented. We propose to combine the conditional score with the weighted-correction estimating function using empirical likelihood. By combining, we eliminate the inconsistent roots induced by the conditional score in large samples. The resulting maximum empirical likelihood estimator is asymptotically locally efficient. Its asymptotical properties and finite-sample performances will be presented. The proposed combined estimation procedure is guaranteed to produce asymptotically locally efficient estimator for logistic regression with errors-in-covariates. The finite-sample simulations show that the proposed combined estimation procedure outperforms existing methods in many situations and could be a more reliable estimation procedure. Applications to the ACTG 175 study (Hammer et al. 1996) and the blood pressure study (Pan et al. 1990) are provided.

4.2 The weighted-correction estimating function

In this section, we develop a new consistent functional method, namely the weighted-correction estimating function, for logistic regression with errors-in-covariates. Similar to the parametric-correction estimation procedure (Huang & Wang 2001), the proposed estimating function only produces consistent estimators in large samples.

Huang & Wang (2001) chose a pair of weights to (2.9) to form the parametric-correction estimation procedure (Section 2.1.4). We consider applying a single weight

$$w(\cdot) = \exp((-\alpha - \beta_z^{\mathrm{T}} Z - \beta_x^{\mathrm{T}} X)/2) + \exp((\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X)/2)$$

to (2.9). Then the resulting correction-amenable estimating function is given by

$$\Psi(\theta) = \left\{ (Y-1)\exp(\frac{\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X}{2}) + Y\exp(-\frac{\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} X}{2}) \right\} \begin{pmatrix} 1 \\ Z \\ X \end{pmatrix}. \tag{4.1}$$

$\Psi(\theta)$ also enjoys the recoding properties of the score function: recoding the outcome event leads to the opposite sign in the coefficients.

To obtain the root of $\Psi(\theta)$, one can solve the unbiased estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(y_i, x_i, z_i, \theta) = 0.$$

Lemma 4.2.1 *Under regularity conditions A1 and A2 in Appendix, the roots of* $\Psi(\theta), \widetilde{\theta}$ *exists in probability and converge to* θ_0*, and* $\sqrt{n}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, V_1)$*, where*

$$V_1 = \left[E\left(\frac{\partial \Psi}{\partial \theta}\right)^{\mathrm{T}} (E(\Psi\Psi^{\mathrm{T}}))^{-1} E\left(\frac{\partial \Psi}{\partial \theta}\right) \right]^{-1}.$$

Under the additive error model $W = X + U$, the weighted-correction estimating function that performs correction based on (4.1) is given by

$$\Psi_{ws}(\theta) = \frac{(Y-1)\exp((\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}W)/2)}{E(\exp(\beta_x^{\mathrm{T}}U/2))}\begin{pmatrix} 1 \\ Z \\ W - \dfrac{E(\exp(\beta_x^{\mathrm{T}}U/2)U)}{E(\exp(\beta_x^{\mathrm{T}}U/2))} \end{pmatrix} + \frac{Y\exp(-(\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}W)/2)}{E(\exp(-\beta_x^{\mathrm{T}}U/2))}\begin{pmatrix} 1 \\ Z \\ W - \dfrac{E(\exp(-\beta_x^{\mathrm{T}}U/2)U)}{E(\exp(-\beta_x^{\mathrm{T}}U/2))} \end{pmatrix}, \tag{4.2}$$

which satisfies

$$E\{\Psi_{ws}(\theta)|(Y,X,Z)\} = \Psi(\theta).$$

To obtain the root of $\Psi(\theta)$, one can solve the unbiased estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_{ws}(y_i,w_i,z_i,\theta) = 0.$$

Theorem 4.2.2 *Under regularity conditions A1, A2 and A3 in the appendix, the roots of* $\Psi_{ws}(\theta)$*,* $\hat\theta$ *exists in probability and converge to* θ_0*, and* $\sqrt{n}(\hat\theta-\theta_0)\xrightarrow{d}N(0,V_2)$*, where*

$$V_2 = \left[E\left(\frac{\partial\Psi_{ws}}{\partial\theta}\right)^{\mathrm{T}}(E(\Psi_{ws}\Psi_{ws}^{\mathrm{T}}))^{-1}E\left(\frac{\partial\Psi_{ws}}{\partial\theta}\right)\right]^{-1}.$$

Similar to the parametric-correction estimation procedure (Huang & Wang 2001), the weighted-correction estimating function does not require the normal error assumption to secure its consistent estimation. It is usually assumed that $U \sim N(0,\Sigma_{uu})$. Then $E(\exp(\beta_x^{\mathrm{T}}U/2)) =$

$\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)$ and $E(\exp(\beta_x^{\mathrm{T}}U/2)U) = \exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)\Sigma_{uu}\beta_x/2$. Therefore, (4.2) becomes

$$\Psi_{ws}(\theta) = (Y-1)\exp((\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}W)/2-\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)\begin{pmatrix}1\\ Z\\ W-\Sigma_{uu}\beta_x/2\end{pmatrix}$$
$$+Y\exp(-(\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}W)/2-\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)\begin{pmatrix}1\\ Z\\ W+\Sigma_{uu}\beta_x/2\end{pmatrix}. \quad (4.3)$$

It is of interest to compare the asymptotical efficiency of the weighted-correction estimating function with the parametric-correction procedure. Without measurement errors, the parametric-correction procedure achieves remarkable efficiency performance compare with the maximum likelihood estimator (Huang & Wang 2001). We perform similar studies to compare the asymptotical relative efficiency of the weighted-correction estimating function estimator $\hat{\theta}$ to the parametric-correction estimation procedure estimator $\bar{\theta}$, in the absence of measurement error. We consider the case that the model only contains a single covariate $X \sim N(0,\sigma^2)$. The asymptotic efficiency depends on $|\alpha_0|$ and $|\beta_0\sigma|$. We investigated several different true parameter values and reported the results in Table 4.1. The relative asymptotical efficiency of $\hat{\theta}$ decreases as $|\alpha_0|$ or $|\beta_0\sigma|$ increases. Compared with $\bar{\theta}$, $\hat{\theta}$ is slightly less efficient.

In the presence of normal measurement errors, the parametric-correction estimator is generally less efficient than the semiparametrically efficient conditional score estimator (Huang & Wang 2001). Since the proposed weighted-correction estimating function also does not have the asymptotical local efficiency property, its estimator is in theory less efficient than the semiparametrically efficient conditional score estimator. Indeed, it is of more interest to compare the weighted-correction estimating function estimator with the parametric-correction estimator in the presence of measurement errors, which will be investigated through simulations.

Table 4.1: *The asymptotic relative efficiency between the weighted-correction estimating function estimator and the parametric-correction estimation procedure estimator (Huang & Wang 2001) in a single covariate ($X \sim N(0,1)$) error-free logistic regression model*

α_0	β_0	$\hat{\alpha}$	$\hat{\beta}$
-1.4	0.7	0.987	0.939
	1	0.968	0.909
	1.4	0.953	0.892
-0.7	0.7	0.993	0.974
	1	0.982	0.951
	1.4	0.975	0.926
0	0.7	0.995	0.987
	1	0.987	0.966
	1.4	0.983	0.937

We conduct several simulation studies to evaluate the finite-sample properties of the weighted-correction estimating function estimator. Suppose we have a logistic model with a single covariate that is measured with error. We consider four distributions for the true covariate X. That is, $X \sim N(0,1)$, $X \sim \text{Unif}(-\sqrt{3}, \sqrt{3})$, $X \sim \exp(1)-1$, and $X \sim (\chi^2_{(1)}-1)/\sqrt{2}$. A single surrogate of X is observed. We consider two substantial amount of measurement errors for each model: $U \sim N(0, 0.25)$ and $U \sim N(0, 0.5)$. The true values for (α_0, β_0) are (0,1). The true variances of the measurement errors are used in the simulations. For each model, we conduct 1000 simulations on sample sizes of 300 and 600. For comparison, the naive estimator, the regression calibration estimator with formulas given by Carroll et al. (2006) in Section 4.4.2, the conditional score estimator with $t(\Delta) = W + (Y - 1/2)\sigma^2_u\beta_x$ in (5.2), the weighted-correction estimating function estimator, and the parametric-correction estimation procedure estimator estimator are also reported in the simulations. The mean bias, the estimated standard deviations and the empirical coverage based 95% Wald type confidence interval are reported. A modified Newton-Raphson procedure (Appendix) starting from the naive estimators was applied to find the roots of the conditional score and weighted-correction estimating function. The parametric-correction estimation procedure estimator (Huang & Wang 2001) was obtained by the standard two-step GMM estimation procedure and the identity matrix was used in the first step to obtain a consistent estimator.

The simulation results for the slope estimators are summarized in Table 4.2. The performance of the naive estimator is unacceptable with large biases and very poor coverage probabilities. This confirms that correction procedures must be employed in the presence of measurement error. The approximately consistent regression calibration estimator performs reasonable well under normal X, as expected. It also performs well when X is uniform distributed. However, it performs poorly when X has a skewed distribution. For the three consistent methods, the bias performances are overall comparable. The estimated standard errors of the weighted-correction estimating function estimator and the conditional score estimator are in general close, which indicates the weighted-correction estimating function achieves good efficiency performance in the presence of measurement error. Both the weighted-correction estimating function and conditional score estimator have excellent coverage probabilities. The HW's estimator appears to have the smallest estimated standard errors among three consistent approaches, but its coverage probabilities tend to be below the nominal level. Huang & Wang (2001) showed that the coverage probabilities could be improved using bootstrap methods. Notice that both the weighted-correction estimating function and conditional score estimator experience root-finding failures when X is skewed and measurement error is large, which indicates that there are no zero-crossings in a neighborhood of naive estimators. HW's estimator does not suffer from the root-finding issues in that it is not defined as a zero-crossing of unbiased estimating functions. Overall, the proposed weighted-correction estimating function estimator has a comparable bias and efficiency performance with the conditional score estimator in finite samples. The HW's estimator performs slightly better in terms of mean bias than the weighted-correction estimating function estimator when the underlying distribution of X is skewed. In general, the weighted-correction estimating function estimator outperforms the HW's estimator in coverage probabilities.

The weighted-correction estimating function estimator has very good finite-sample performance overall, especially when the sample size is moderate large. We believe that it is a valuable approach and can be applied to large-scale epidemiologic studies as an alternative to the conditional score approach.

Table 4.2: *Summary of the performances of Slope estimators in a single and error-prone covariate logistic regression model.*

size	σ_u^2		NV	RC	CS	HW	WS	NV	RC	CS	HW	WS
			$X \sim N(0,1)$					$X \sim \text{Unif}(-\sqrt{3},\ \sqrt{3})$				
300	0.25	Bias	-223.2	-27.2	19.7	-19.8	28.6	-197.6	4.1	12.3	-22.4	17.3
		SD	129.7	162.9	186.1	203.3	198.1	129.1	163.0	169.0	183.4	178.3
		SE	129.5	162.3	185.6	172.4	190.2	125.8	157.5	168.0	153.9	174.3
		EC	55.7	94.3	95.5	85.7	94.9	63.2	94.8	95.6	85.8	94.9
		F	0	0	0	0	0	0	0	0	0	0
	0.5	Bias	-368.3	-47.9	32.7	-23.5	49.8	-338.8	-5.0	20.1	-31.0	31.1
		SD	115.4	177.5	224.6	230.6	248.4	114.7	177.9	197.4	209.3	218.4
		SE	113.9	171.9	221.1	203.4	231.7	112.7	169.8	196.0	180.0	209.1
		EC	13.3	93.3	96.0	85.8	95.5	16.6	94.1	96.4	85.7	95.8
		F	0	0	0	0	0	0	0	0	0	0
600	0.25	Bias	-228	-34.6	9.6	-14.0	13.7	-202.6	-3.1	3.6	-16.0	6.0
		SD	91.7	114.4	129.5	140.5	134.0	91.9	116	119.4	130.4	123.1
		SE	91.1	114.0	129.9	126.9	133.3	88.5	110.6	117.5	114.1	121.9
		EC	28.5	93.9	95.3	89.7	94.4	37.1	93.9	94.3	88.4	94.5
		F	0	0	0	0	0	0	0	0	0	0
	0.5	Bias	-372.4	-57.3	16.6	-22.3	23.4	-343.2	-13.6	7.0	-23.1	11.9
		SD	80.7	122.3	151.2	162.3	159.3	81.8	126.4	136.7	147.1	144.0
		SE	80.1	120.4	152.7	147.9	158.6	79.2	119.1	135.8	132.6	144.3
		EC	0.7	91.4	95.6	89.2	95.8	2.0	93.0	94.8	89.0	94.7
		F	0	0	0	0	0	0	0	0	0	0
			$X \sim \exp(1) - 1$					$X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$				
300	0.25	Bias	-280.9	-98.7	29.0	21.4	50.6	-322.3	-145.3	51.6	47.5	71.4
		SD	130.4	165.1	230.1	253.3	250.5	144.4	184.6	302.9	305.4	308.3
		SE	136.7	171.5	226.2	198.1	224.6	141.8	179.2	273.2	220.3	259.6
		EC	43.5	90.0	95.7	86.5	94.2	36.2	80.6	94.5	84.0	93.3
		F	0	0	0	0	0	0	0	0	0	0
	0.5	Bias	-434.7	-147.9	44.6	1.8	76.9	-476.9	-202.7	88.7	19.5	103.4
		SD	110.3	171.7	290.5	271.6	332.8	119.5	190.2	408.6	318.7	396.5
		SE	116.1	175.4	283.8	313.6	302.1	118.1	180.8	369.6	298.5	362.0
		EC	5.4	84.4	95.9	86.1	96.8	4.1	73	93.8	85.1	94.5
		F	0	0	0.3	0	0.3	0	0	0.3	0	1.0
600	0.25	Bias	-287.6	-108.3	11.2	3.1	22.4	-329.5	-159.2	20.3	18.8	31.6
		SD	91.7	115.4	155.5	171.9	164.9	99.5	126.7	193.8	200.8	194.8
		SE	96.1	120.4	157.0	145.5	156.9	99.5	124.8	184.4	160.9	177.8
		EC	17.2	84.4	95.5	89.7	95.6	11.5	70.9	95.1	88.5	93.7
		F	0	0	0	0	0	0	0	0	0	0
	0.5	Bias	-439.5	-157.1	19.7	-9.5	33.3	-482.2	-218.8	35.5	9.6	47.6
		SD	78.0	120.2	191.6	200.7	204.2	82.5	129.8	241.6	227.6	239.5
		SE	81.7	123.1	191.5	171.2	191.3	82.9	125.2	232.3	191.5	221.7
		EC	0	71.9	96.0	88.5	96.0	0	53.5	95.5	88.9	95.2
		F	0	0	0	0	0	0	0	0	0	0

Note: measurement error $U \sim N(0, \sigma_u^2)$; NV: naive estimator; RC: regression calibration estimator; CS: conditional score estimator; HW: Huang & Wang (2001) parametric-correction estimation procedure estimator; WS: weighted-correction estimating function estimator; Bias: mean bias (× 1000); SD: empirical Monte Carlo standard deviations (× 1000) of estimators; SE: the average of estimated standard errors(× 1000); EC: empirical coverage probability (%) of 95% Wald confidence interval based on SE; F: root finding failure (%).

4.3 The combined estimation procedure

When there are more estimating functions than parameters, a general approach is to combine estimating functions to use as much as possible the information available to estimate the parameters of interest. Among methods to combine estimating functions, empirical likelihood provides a flexible and efficient way to combine information about parameters and distributions. Qin & Lawless (1994) showed that the nonparametric empirical likelihood approach holds many nice parametric properties (Section 2.3.2). For example, the empirical log likelihood ratio follows a limiting distribution of chi-square. Analog to the parametric approach to obtain an efficient estimator, maximizing the empirical likelihood function yields the efficient estimator in the class of estimators based on combined estimating functions (Qin & Lawless 1994, Owen 2001).

We propose to combine the conditional score with the weighted-correction estimating function using the empirical likelihood. The essence of combining two estimating functions is that at the limit, the weighted-correction estimating function and the conditional score only share the unique zero crossing at the true value of the parameter. Therefore, the combined estimation procedure does not have multiple roots problem in large samples. In addition, the combined estimation procedure maintains the asymptotical local optimality of the conditional score since the maximum empirical likelihood estimator is at least as efficient as the estimators from each estimating function combined. Moreover, as a nonparametric method, the empirical likelihood is expected to have satisfactory finite-sample performance. Therefore, we adopt the empirical likelihood method to combine the conditional score and the weighted-correction estimating function.

Let $O = (Y, W, Z)$ be the observed data and

$$g(o, \theta) = \begin{pmatrix} \Psi_{cs}(o, \theta) \\ \Psi_{ws}(o, \theta) \end{pmatrix},$$

where $\Psi_{cs}(\theta)$ and $\Psi_{ws}(\theta)$ are defined in (5.2) and (4.2), respectively. The profile empirical likelihood function for θ is

$$L(\theta) = \sup\left\{\prod_{i=1}^{n} p_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(o_i, \theta) = 0\right\}. \tag{4.4}$$

A unique maximum of $L(\theta)$ exists provided that 0 is in the convex hull of $\{g_i(\theta), i = 1, \ldots, n\}$. (4.4) is a constraint optimization problem. Section 2.3.2 shows that the empirical loglikelihood ratio for θ is

$$l_E(\theta) = -\sum_{i=1}^{n} \log\{1 + \lambda^{\mathrm{T}} g(o_i, \theta)\},$$

and the maximum empirical likelihood estimator for $\theta_0 \in \mathbb{R}^p$ is given by

$$\tilde{\theta} = \arg\max_{\theta \in \mathbb{R}^p} \min_{\lambda \in \mathbb{R}^{2p}} -\sum_{i=1}^{n} \log(1 + \lambda^{\mathrm{T}} g(o_i, \theta)), \tag{4.5}$$

where λ is the Lagrange multiplier.

Theorem 4.3.1 *Assume that $\theta \in \mathbb{R}^p$ and $g(\theta) \in \mathbb{R}^s$. Under regularity conditions B1-B9 in the appendix, then*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

$$-2l_E(\theta_0) \xrightarrow{d} \chi^2_s,$$

$$-2\{l_E(\theta_0) - l_E(\tilde{\theta})\} \xrightarrow{d} \chi^2_p,$$

$$-2l_E(\tilde{\theta}) \xrightarrow{d} \chi^2_q,$$

where $q = s - p$ and

$$V = [E(\frac{\partial g}{\partial \theta})^{\mathrm{T}}(E(gg^{\mathrm{T}}))^{-1}E(\frac{\partial g}{\partial \theta})]^{-1}.$$

Corollary 4.3.2 *Let $\theta^{\mathrm{T}} = (\theta_1, \theta_2)^{\mathrm{T}}$, where $\theta_1 \in \mathbb{R}^k$ and $\theta_2 \in \mathbb{R}^{p-k}$. The empirical loglikeli-*

hood ratio statistic for testing $H_0 : \theta_1 = \theta_1^0$ *is*

$$-2\{l_E(\theta_1^0, \tilde{\theta}_2^0) - l_E(\tilde{\theta}_1, \tilde{\theta}_2)\} \xrightarrow{d} \chi_k^2,$$

where $\tilde{\theta}_2^0$ *is the MELE of* $l_E(\theta_1^0, \theta_2)$.

The empirical loglikelihood ratio statistic is analogous to the parametric loglikelihood ratio statistic in many large sample properties. We can conduct testing and obtain confidence regions for parameters of interest using the empirical likelihood similar to the general parametric likelihood approach.

Corollary 4.3.3 $\tilde{\theta}$ *is asymptotically locally efficient.*

Corollary 4.3.3 is a natural result from Corollary 1, 2 and 3 (Qin & Lawless 1994). Indeed, $\tilde{\theta}$ is the optimal estimator among the linear combinations of $\Psi_{cs}(o, \theta)$ and $\Psi_{ws}(o, \theta)$. The fact that the consistent conditional score estimator is asymptotically locally efficient implies $\tilde{\theta}$ is asymptotically locally efficient.

In addition to showing the large-sample advantages of the combined estimation procedure over either the conditional score or weighted-correction estimating function, we investigate the finite-sample advantages of combining two estimating functions.

we consider a single and error-prone covariate logistic regression model with $X \sim N(0, 1)$. The true values for (α_0,β_0) are (0,1). The measurement error follows a standard normal distribution. The sample size considered is 200. For illustration purpose, we profile the intercept out and only draw the weighted-correction estimating function plots with respect to slope β only. We also plot out the conditional score and the combined estimation procedure. Figure 4.1 and Figure 4.2 are plots of two samples. For the sample in Figure 4.1, we found three roots for both the conditional score and the weighted-correction estimating function. Only the middle one appears to be the correct solution. For the sample in Figure 4.2, the only root for the weighted-correction estimating function is a false root and is far from the truth. We again found three roots for the conditional score. After combining these two estimating functions, we found a global empirical likelihood ratio maximizer around the truth. The empirical likelihoods are maximized at $\hat{\beta} = 0.87$ and 1.17 for those two samples, respectively. These two plots show a very promising picture that combining the conditional score and the weighted-correction estimating function has great potential to produce a unique global maximizer around the truth. In finite samples, the global empirical likelihood maximizer may not be attainable in a neighborhood of the truth. However, starting from the naive estimator, our approach typically identifies a good local empirical likelihood maximizer around the truth. As the measurement error becomes smaller or sample size goes larger, a global maximum of the empirical likelihood can be attainable around the truth in a compact parameter space.

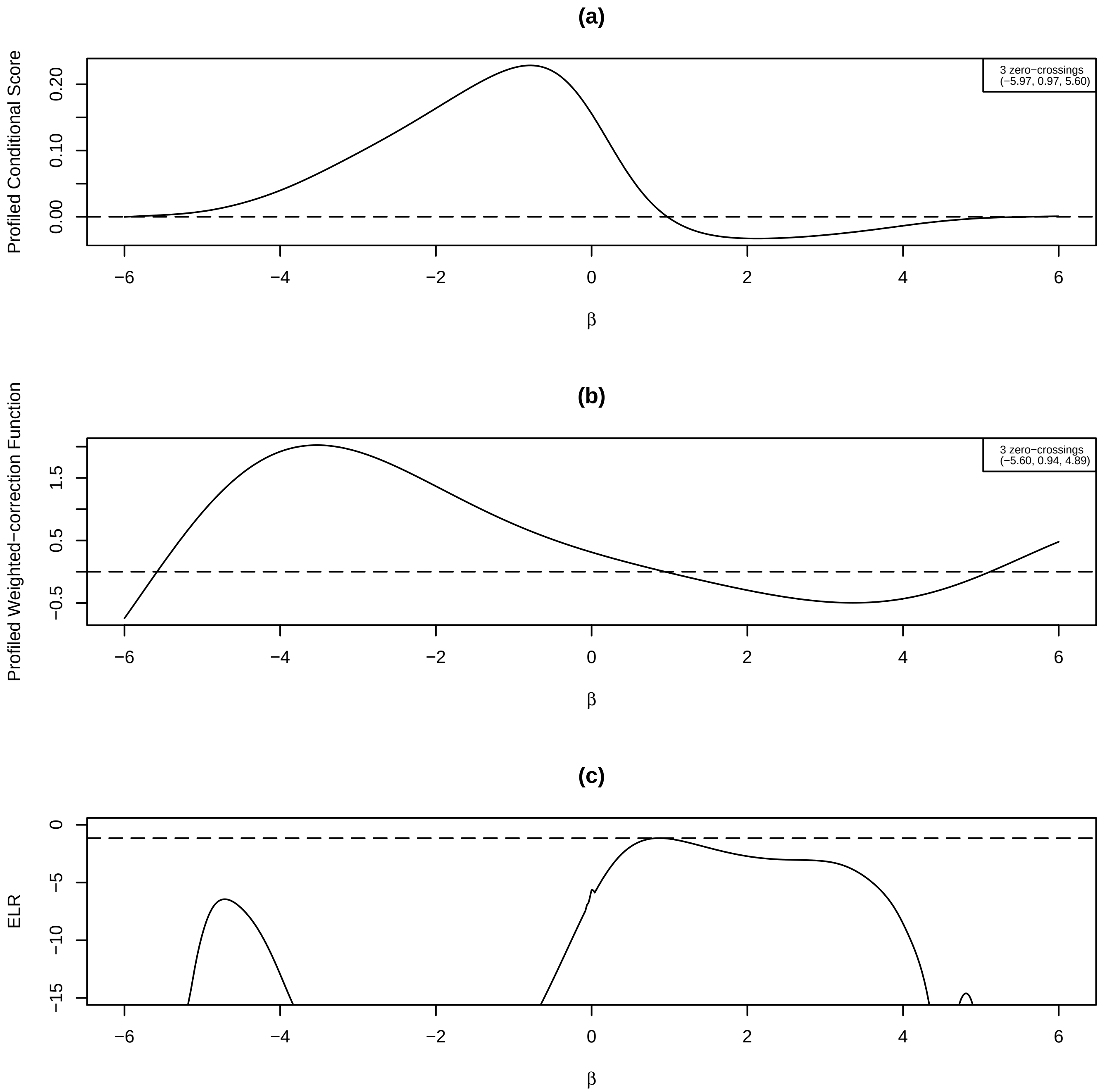

Figure 4.1: *Plots of (a) the profiled conditional score , (b) profiled weighted-correction estimating function , and (c) the empirical likelihood ratio (ELR) based on the combined estimation procedure for a single and error-prone covariate logistic regression model. For this sample data: N=200. True values* $(\alpha_0, \beta_0) = (0, 1)$. *Both the true covariate and the additive measurement error follow a standard normal distribution.* $\hat{\beta}_{naive} = 0.38$. *The ELR is maximized at* $\hat{\beta} = 0.87$.

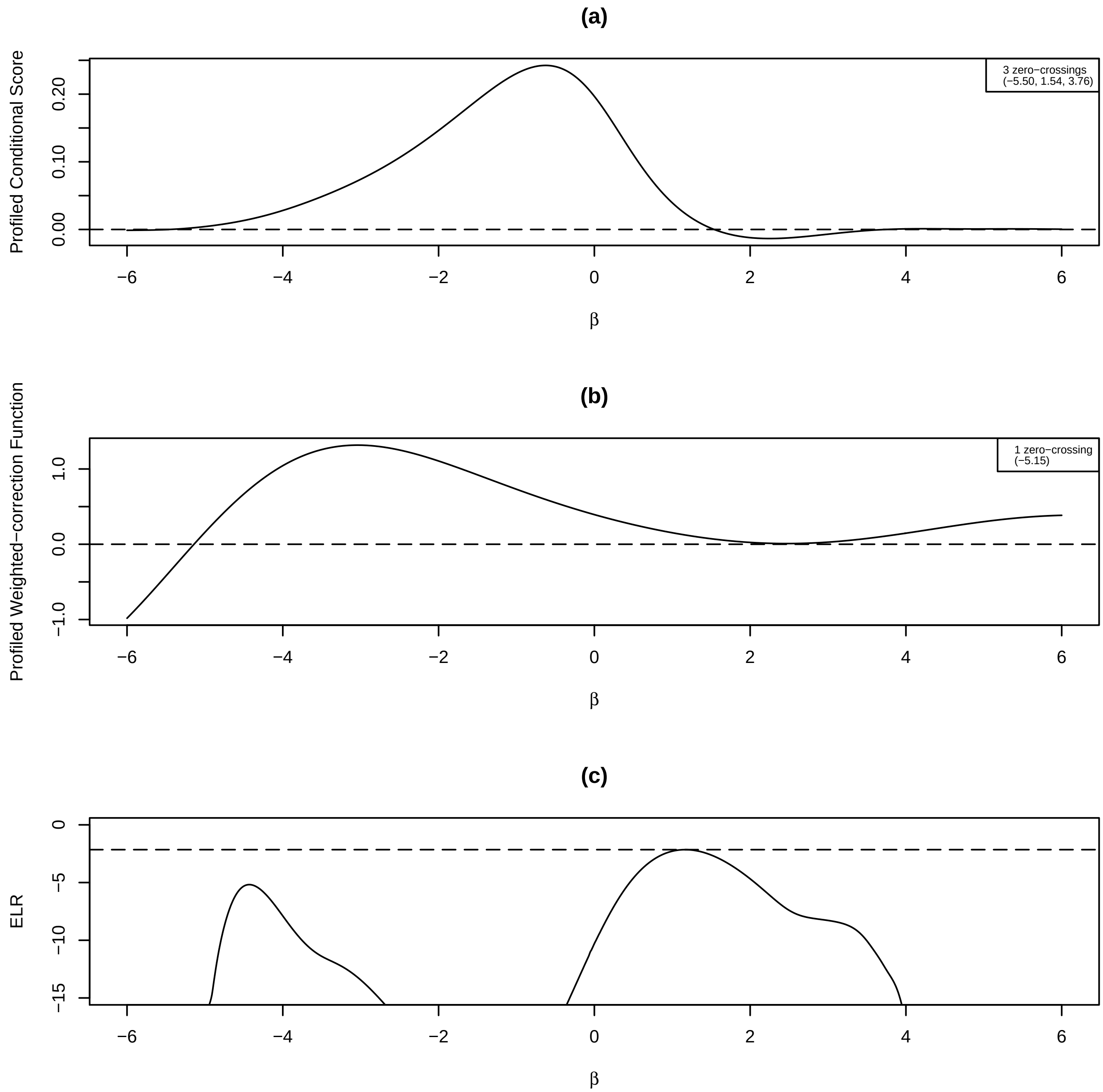

Figure 4.2: *Plots of (a) the profiled conditional score , (b) profiled weighted-correction estimating function , and (c) the empirical likelihood ratio (ELR) based on the combined estimation procedure for a single and error-prone covariate logistic regression model. For this sample data: N=200. True values* $(\alpha_0, \beta_0) = (0, 1)$. *Both the true covariate and the additive measurement error follow a standard normal distribution.* $\hat{\beta}_{naive} = 0.54$. *The ELR is maximized at* $\hat{\beta} = 1.17$.

4.4 Simulations

We conduct simulation studies to evaluate the finite-sample performance of the proposed maximum empirical likelihood estimator. Suppose we have a logistic model with a single covariate that is measured with large measurement error. We consider two distributions for the true covariate X: $X \sim N(0,1)$ and $X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$. A single surrogate of X is observed. We consider a model with large measurement error: $U \sim N(0,1)$. The true values for (α_0, β_0) are (0,1). The true variances of the measurement errors are used in the simulations. For each model, we conduct 1000 simulations on sample sizes of 200 and 500. For comparison, the naive estimator, the regression calibration estimator, the conditional score estimator, the weighted-correction estimating function estimator , and the parametric-correction procedure estimator (Huang & Wang 2001) are also reported in the simulations. Considering that each estimator has some skewness due to the large amount of measurement errors, we report the bias and efficiency performance using the median bias and interquartile divided by 1.349, which are similar to Huang & Wang (2001). The empirical coverage probabilities are evaluated using the 95% bootstrap percentile confidence interval. The bootstrap size considered is 39. A bootstrap size of 39 might not be big enough to obtain very accurate confidence intervals. However, it is suffice to compare different approaches here. The naive estimators are chosen as the starting points to perform all numerical iterations. The simulation results are summarized in Table 4.3.

As seen, the naive estimator has expected large bias. As a natural consequence, its coverage are extremely poor. The regression calibration estimator has good bias performance under normal X. However, its performance becomes very poor under chi-square X. Both the conditional score estimator and the weighted-correction estimating function estimator experience certain amount of root finding failures, especially when sample size is small. Other than that, both estimators have good bias performance on successful root-findings. The bias performance of maximum empirical likelihood estimator is also very good, whereas the HW's estimator tends to have a larger bias than other consistent es-

Table 4.3: *The performance of Slope estimators in a single and error-prone covariate logistic regression model: Summary of simulation studies.*

Size		NV	RC	CS	HW	WS	EL
				$X \sim N(0,1)$			
200	Bias	-541.0	-74.2	31.2	-103.4	16.5	-10.4
	SD	118.9	261.7	349.6	278.9	371.0	392.3
	EC	2.3	88.2	94.1	91.0	95.8	96.6
	F	0	0	1.7	0	8.4	0
500	Bias	-542.2	-89.4	17.8	-50.5	33.2	27.7
	SD	70.1	159.7	228.0	198.5	236.2	234.5
	EC	0	87.3	93.0	91.0	93.3	93.2
	F	0	0	0	0	1.0	0
				$X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$			
200	Bias	-637.9	-267.3	-38.0	-195.1	-72.7	62.8
	SD	108.6	264.4	491.0	307.0	407.3	479.0
	EC	0.8	78.3	92.1	84.1	96.8	95.2
	F	0	0	8.5	0	19.3	0
500	Bias	-639.4	-278.7	19.1	-68.0	28.1	86.7
	SD	71.1	161.4	331.1	244.8	310.9	351.2
	EC	0	59.1	93.0	90.9	93.5	91.0
	F	0	0	1.4	0	4.9	0

Note: measurement error $U \sim N(0,1)$; NV: naive estimator; RC: regression calibration estimator; CS: conditional score estimator; HW: Huang & Wang (2001)parametric-correction estimation procedure estimator; WS: weighted-correction estimating function estimator; EL: the maximum empirical likelihood estimator; Bias: median bias ($\times$ 1000); SD: estimated standard deviations ($\times$ 1000) based on interquartile; EC: empirical coverage probability(%) of 95% bootstrap percentile confidence interval; F: root-finding failure (%).

timators when the sample size is small. The HW's estimator has the smallest estimated standard deviation associated with lower coverage probabilities than other consistent estimators. The the proposed maximum empirical likelihood estimator, the conditional score estimator and the weighted-correction estimating function estimator have similar performance on empirical coverage rates. Similar to HW's estimator, the proposed maximum empirical likelihood estimator does not suffer from the root finding failures by maximizing an objective function. Figures 4.3 and 4.4 show QQ plots for each slope estimator of the four consistent methods. The HW's estimator has the smallest skewness. Both the conditional score and the weighted-correction estimating function estimators suffers from outliers. Consequently, it is not surprised that the proposed maximum empirical likelihood estimator has outstanding outliers. The problems of outliers become more noticeable when X has a modified chi-square distribution.

For the proposed maximum empirical likelihood estimator, there are at least for different ways to obtain the 95% confidence intervals: the Wald type confidence interval, the bootstrap percentile confidence interval, the empirical likelihood ratio confidence interval based on Chi-square approximation (Appendix), and the bootstrap empirical likelihood ratio confidence interval (Appendix). The last two types of confidence intervals are not available to HW's estimator based on the GMM approach. This is an advantage of the empirical likelihood over the GMM approach to combine estimating functions since empirical likelihood provides more ways to obtain confidence intervals. To compare those four types of confidence intervals, we summarize the coverage probabilities for the slop estimator from each confidence interval using the same set up as in Table 4.3. The results are shown in Table 4.4.

Table 4.4: *The comparison of empirical coverage probabilities (%) of four 95% confidence intervals of the maximum empirical likelihood slope estimators in a single and error-prone covariate logistic regression model*

Size	Wald	ELC	BP	BELR
	$X \sim N(0,1)$			
200	75.0	89.5	96.6	98.4
500	91.3	93.1	93.2	95.6
	$X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$			
200	71.1	84.7	95.2	97.2
500	84.1	87.5	91.0	94.6

Note: measurement error $U \sim N(0,1)$; Wald: Wald confidence interval. ELC: Empirical likelihood ratio confidence interval. BP: bootstrap percentile confidence interval with the bootstrap size of 39. BELR: bootstrap empirical likelihood ratio confidence interval with the bootstrap size of 39.

The coverage probabilities of the Wald intervals are consistently below the nominal level. The empirical likelihood ratio confidence interval has better coverage probabilities than the Wald interval. However, it is in general below the nominal level in finite samples. Compared to the bootstrap percentile confidence interval, the bootstrap empirical likelihood ratio confidence interval tends to have a larger coverage probabilities and shows a satisfactory coverage probability at sample size of 500. Theoretically, the bootstrap empirical likelihood ratio confidence interval is $O(n^{-2})$ accurate. We recommend using either the bootstrap percentile method or the bootstrap empirical likelihood ratio method in practice to obtain confidence intervals.

In the previous numerical studies, the measurement error and the true covariate have the same standard deviation. The results show that the proposed estimator has a good performance especially at sample size of 500. Indeed, it is possible that more substantial measurement error would occur in practice due to the difficulties to obtain accurate measurements of the true covariates. To investigate the abilities of the proposed procedure to obtain valid point and interval estimates in heavily contaminated data, we conduct another set of numerical studies with an increased amount of measurement error. We consider a logistic model with single covariate measured with error and four distributions for the true covariate X with standard deviation of 1. That is, $X \sim N(0, 1)$, $X \sim \text{Unif}(-\sqrt{3}, \sqrt{3})$, $X \sim \exp(1) - 1$, and $X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$. A single surrogate of X is observed. We consider a model with very large measurement error: $U \sim N(0, 1.3^2)$. The true values for (α_0, β_0) are (0,1). The empirical coverage probabilities are evaluated using the 95% bootstrap percentile confidence interval. The bootstrap size is 39. The true variances of the measurement errors are used in the simulations. Simulation results based on 1000 simulations are summarized in Table 4.5.

It is not surprised that both the naive estimator and the regression calibration estimator perform poorly in general. Compared to the results in Table 4.3, both the conditional score estimator and the weighted-correction estimating function estimator experience even larger percents of root finding failures, due to a larger measurement error. Other than that, both estimators have good performance on bias and coverage probabilities. The HW's estimator appears to have an undesired median bias performance. In addition, its coverage probabilities are consistently below the nominal level, a natural consequence of its large biases and small estimated standard deviations. The proposed maximum empirical likelihood estimator performs excellent in terms of its small median bias. Its performance on coverage probabilities is slightly better than the conditional score estimator and the weighted-correction estimating function estimator. In general, it outperforms all other estimators.

In summary, both the proposed combined estimation procedure and parametric-correction estimation procedure (Huang & Wang 2001) have the theoretical advantages over the conditional score in that they only admit consistent sequences in large samples. In finite samples, they do not suffer the possible root-finding failures as the conditional score and weighted-correction estimating function does. In general, the proposed estimator has the efficiency advantage over HW's estimator in large samples. In addition, it has a better coverage property than HW's estimator in finite samples. When the measurement error is not too large, HW's estimator works excellent even if the sample size is small. The HW's estimator is also less skewed than the proposed estimator. The proposed estimator appears to work better when the measurement error is large. When the sample size is small, the proposed estimator shows some undesired skewness performance. However, the proposed maximum empirical likelihood estimator has evident advantages over other existing estimators in dealing with large measurement errors at moderate or large-scale studies.

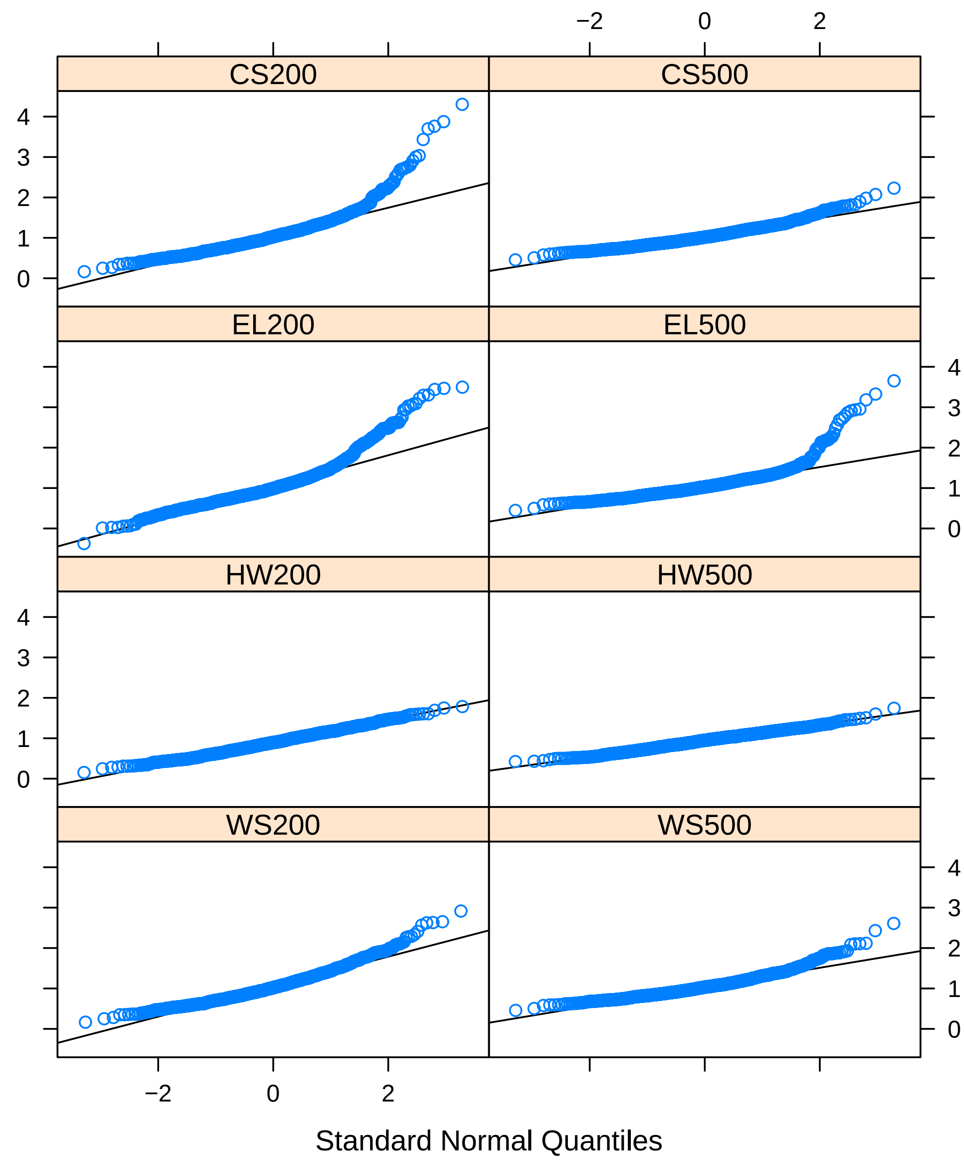

Figure 4.3: *QQ plots for Slope estimators in a single and error-prone covariate logistic regression model with successful root-finding only. Under the set ups in Table 4.3. $X \sim N(0,1)$. $U \sim N(0,1)$. CS: conditional score estimator; HW: Huang & Wang (2001)parametric-correction estimation procedure estimator; WS: weighted-correction estimating function estimator; EL: the maximum empirical likelihood estimator; 200 and 500 are sample sizes.*

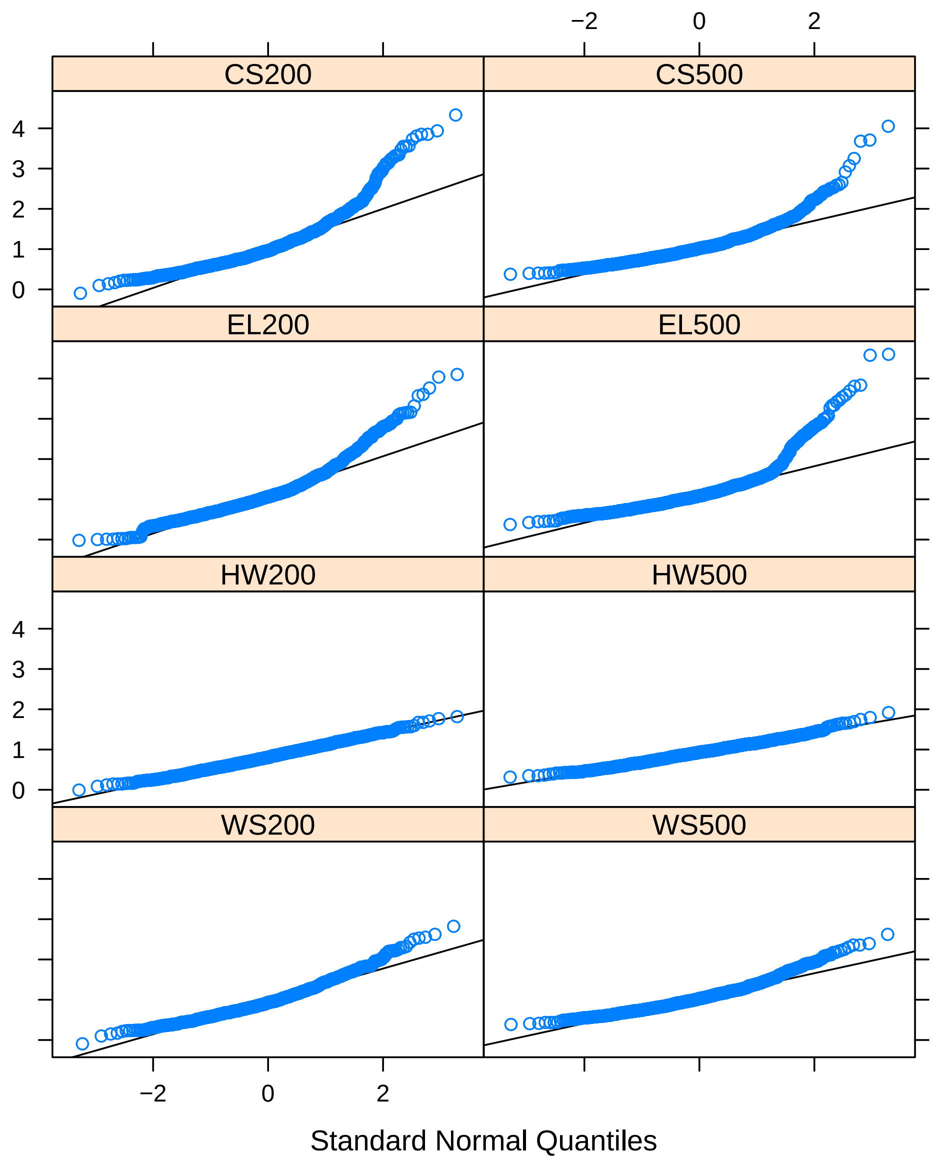

Figure 4.4: *QQ plots for Slope estimators in a single and error-prone covariate logistic regression model with successful root-finding only. Under the set ups in Table 4.3.* $X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$. $U \sim N(0,1)$. *CS: conditional score estimator; HW: Huang & Wang (2001)parametric-correction estimation procedure estimator; WS: weighted-correction estimating function estimator; EL: the maximum empirical likelihood estimator; 200 and 500 are sample sizes.*

Table 4.5: *The performance of Slope estimators in a single and error-prone covariate logistic regression model with very large measurement errors*

	NV	RC	CS	HW	WS	EL
			$X \sim N(0,1)$			
Bias	-669	-103.2	15.5	-148.6	9.6	9.6
SD	60.2	193.2	288.3	189.6	287.3	290.6
EC	0	87.2	94.4	78.0	96.8	95.2
F	0	0	2.7	0	10.6	0
			$X \sim \text{Unif}(-\sqrt{3}, \sqrt{3})$			
Bias	-647.1	-38.5	4.7	-139.1	-1.3	-21.3
SD	56.8	195.6	226.2	172.2	238.3	202.5
EC	0	91.3	96.1	84.0	96.2	95.4
F	0	0	1.2	0	8.5	0
			$X \sim \exp(1) - 1$			
Bias	-713.7	-227.6	-16.9	-188.8	-39.3	17.3
SD	57.2	179.2	319.6	214.2	300.2	341.4
EC	0	73.1	96.0	70.0	96.4	96.4
F	0	0	5.6	0	18.0	0
			$X \sim (\chi^2_{(1)} - 1)/\sqrt{2}$			
Bias	-740.6	-304.9	-4.6	-214.1	-44.4	33.8
SD	57.6	189.9	381.4	228.1	343.1	412.5
EC	0	65.2	93.3	67.4	97.2	96.2
F	0	0	7.1	0	20.8	0

Note: Sample size: 500. Measurement error $U \sim N(0, 1.3^2)$; NV: naive estimator; RC: regression calibration estimator; CS: conditional score estimator; HW: Huang & Wang (2001)parametric-correction estimation procedure estimator; WS: weighted-correction estimating function estimator; EL: the maximum empirical likelihood estimator; Bias: median bias (× 1000); SD: estimated standard deviations (× 1000) based on interquartile; EC: empirical coverage probability(%) based on 95% bootstrap percentile confidence interval; F: root-finding failure (%).

4.5 Real studies

The proposed maximum empirical likelihood estimator and the weighted-correction estimating function estimator are evaluated using two real examples. For comparisons, the analysis results from the naive approach, regression calibration approach, conditional score approach, and the HW's parametric-correction estimation procedure approach are also reported.

The ACTG study

The AIDS Clinical Trials group (ACTG) 175 study (Hammer et al. 1996) is a randomized, double-blind, placebo-controlled trial to compare treatment with either a single nucleoside or two nucleosides in adults infected with human immunodeficiency virus type 1 (HIV-1) whose CD4 cell counts were from 200 to 500 per cubic millimeter and had no history of an AIDS-defining illness. A total of 2467 HIV-1-infected patients were recruited from 43 AIDS Clinical Trials Units and 9 National Hemophilia Foundation sites in the United States and Puerto Rico. A particular research question is to assess the effect of the true baseline CD4 count on the symptomatic HIV infection defined as candidiasis, oral hairy leukoplakia, or herpes zoster reported within 30 days before randomization in antiretroviral-naive patients (Huang & Wang 2000, 2001). Since the true baseline CD4 count is unobservable, the screening baseline CD4 count is usually used as a substitute as the true baseline CD4 count. As a fact, the screening baseline CD4 count is subject to both instrumental error and biological diurnal fluctuation. Therefore, the true baseline CD4 count is measured with errors. In this study, 1067 patients without antiretroviral therapy were included and 1036 of them had duplicated screening baseline CD4 count measurements prior to the start of treatment and within 3 weeks of randomization. Huang & Wang (2001) adopted a single covariate logistic regression model by treating the symptomatic HIV infection as outcome and the true baseline log(CD4) count as the covariate to assess their relationship based on the 1036 patients. The average of the duplicated screening baseline log(CD4) counts were used as the observed covariate in the model. The measurement error is assumed to be zero mean normally distributed. The variances of the measurement error and the true baseline log(CD4) count are estimated to be 0.033 and 0.076 , respectively, using the formulas given by Carroll et al. (2006) in Section 4.4.2. The amount of the measurement error is substantial (43% variation compared to the true covariate).

Table 4.6 Shows the analysis results using the six approaches. The naive approach that treats the average of the two duplicated screening baseline log(CD4) counts as the true covariate. The 95% confidence intervals are bootstrap percentile intervals based on 999 bootstrap samples, as adopted in Huang & Wang (2001). As seen, the naive estimate has substantially smaller magnitude than others. All other approaches show similar results. The weighted-correction estimating function estimate and the maximum empirical likelihood estimate have slightly shorter lengths of confidence intervals than those of the conditional score estimate and the parametric-correction estimation procedure estimate. The regression approach works comparable to other consistent methods, which suggests that both the true covariate and the measurement error are approximately normal. A better approach to find the 95% confidence interval of the maximum empirical likelihood estimator is to perform bootstrap calibration on the empirical likelihood ratio. However, for illustration purpose, all the approaches adopt the same technique to obtain the interval estimates.

Table 4.6: *The summary of the coefficient estimators in ACTG 175 study: Comparison of different approaches*

	Intercept		ln(CD4)	
	Estimate	95% CI	Estimate	95% CI
NV	4.636	(1.516, 7.883)	-1.080	(-1.647, -0.531)
RC	6.004	(2.168, 9.981)	-1.313	(-2.003, -0.645)
CS	5.896	(2.134, 9.726)	-1.296	(-1.957, -0.644)
WS	5.767	(2.125, 9.659)	-1.274	(-1.943, -0.642)
HW	5.955	(2.231, 9.769)	-1.306	(-1.965, -0.652)
EL	5.939	(2.038, 9.679)	-1.304	(-1.941, -0.636)

Note: NV: naive estimator; RC: regression calibration estimator; CS: conditional score estimator; WS: weighted-correction estimating function estimator; HW: Huang & Wang (2001)parametric-correction estimation procedure estimator; EL: The proposed maximum empirical likelihood estimator; 95% CI: Bootstrap percentile confidence interval based on 999 bootstrap samples.

A blood pressure study

In this study, the relationships between 24-hour urinary sodium chloride and blood pressure were investigated in 397 middle-aged Chinese men living in Taipe by Pan et al. (1990). Hanfelt & Liang (1997) used this example to investigate their conditional quasi-likelihood. Seven overnight urinary sodium chloride measurements had a within-subject variability of 1.2 when standardized so that the mean urinary sodium chloride measurements had a variability of 1 across the 397 subjects (Hanfelt & Liang 1997). A logistic regression model with high systolic blood pressure as outcome and the 24-hour urinary sodium chloride measurement, plus age and body mass index (BMI) as covariates was adopted by Hanfelt & Liang (1997). The measurement error in the model is induced since the true 24-hour urinary sodium chloride measurement is not measurable. In the model, a single (the most recent) urinary sodium chloride measurement was used. Therefore, the variance of the measurement error can be assumed to be 1.2 as aforementioned. The estimated variance of the most recent urinary sodium chloride measurement is 1.82. Then the variance of the unobservable true covariate can be estimated by 0.62 under additive measurement error assumption. The amount of the measurement error is very large (about 200% variation compared to the true covariate). The age and BMI are assumed to be measured with a very small amount of error due to rounding off. Their measurement error variances are assumed to be 0.083.

Table 4.7 shows the analysis results using the six approaches. As seen, for the coefficient of urinary sodium chloride, the naive estimate has substantially smaller magnitude than others. The regression calibration estimate has the largest magnitude and estimated standard error. The proposed empirical likelihood estimate appears to have a smaller magnitude than those of the other three consistent methods. The proposed empirical likelihood estimate and the parametric-correction estimation procedure estimate have obvious smaller estimated standard errors than the conditional score estimate and the weighted-correction estimating function estimate. This is not very surprise since the normal approximation does not perform well for the proposed estimator and the parametric-correction estimation procedure estimator in terms of interval estimations. Different from the ACTG example, those four consistent approaches show some discrepancies in estimation in the blood pressure study. Indeed, when the measurement error is relative small, each of the four consistent approach would perform similar. However, when the measurement error is very large, they may show some differences.

It is worth pointing out that using bootstrap method to obtain more accurate variance estimates does not apply to this study due to the numerical difficulties resulting from the small rate (7.8%) of subjects having high systolic blood pressure in the original data.

Table 4.7: *The summary of the coefficient estimators in the blood pressure study (Pan et al. 1990): Comparison of different approaches*

	Intercept		USC		Age		BMI	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
NV	-2.86	0.25	0.33	0.13	0.08	0.03	0.28	0.08
RC	-2.84	0.25	0.96	0.38	0.08	0.03	0.28	0.08
CS	-2.91	0.27	0.61	0.29	0.08	0.04	0.26	0.09
HW	-2.93	0.26	0.59	0.15	0.08	0.04	0.26	0.10
WS	-2.98	0.28	0.55	0.22	0.09	0.04	0.27	0.09
EL	-2.90	0.25	0.47	0.11	0.06	0.03	0.32	0.08

Note: USC: urinary sodium chloride. NV: naive estimator; RC: regression calibration estimator; CS: conditional score estimator; WS: weighted-correction estimating function estimator; HW: Huang & Wang (2001)parametric-correction estimation procedure estimator; EL: The proposed maximum empirical likelihood estimator; SE: estimated standard error.

4.6 Discussions

When the dimension of estimating functions is larger than the dimension of the parameters ($s > p$), empirical likelihood provides a way to combine estimating functions and yields consistent estimators. There are also other approaches available in the case of $s > p$. In generally, one could consider a class of linear combinations of $g(x, \theta)$

$$\psi(x, \theta) = A(\theta)g(x, \theta),$$

where $A(\theta)$ is a $p \times s$ real-functions matrix with rank of p for all the values of θ. The optimal choice of $A(\theta)$ yields the most efficient estimator or θ based on $g_1(x, \theta), \ldots, g_s(x, \theta)$. Surprisingly, maximizing the empirical likelihood automatically determine the optimal choice of $A(\theta)$. As a result, the MELE is fully efficient in the class of $p \times 1$ estimating functions based on any linear combinations of $g_1(x, \theta), \ldots, g_s(x, \theta)$.

Another popular method to combine estimating functions is the generalized method-of-moments (GMM) method (Hansen 1982). The GMM estimator is defined as the minimizer of the quadratical form

$$\bar{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(x_i, \theta) \right\}^{\mathrm{T}} W^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} g(x_i, \theta) \right\}, \tag{4.6}$$

where W is some positive definite matrix. Different choices of W yield different estimators and they are all consistent. The GMM estimator $\bar{\theta}$ is efficient in the class of estimators based on $g_1(x, \theta), \ldots, g_s(x, \theta)$ if the optimal W is chosen. More specifically,

$$W = \frac{1}{n} E\{g(\theta, X)g(\theta, X)^{\mathrm{T}}\}. \tag{4.7}$$

The optimal W can be consistently estimated by

$$\hat{W} = \left\{ \frac{1}{n^2} \sum_{i=1}^{n} g(\hat{\theta}^*, x_i) g(\hat{\theta}^*, x_i)^{\mathrm{T}} \right\}, \tag{4.8}$$

where $\hat{\theta}^*$ is any consistent estimator of θ. In practice, the GMM estimator is usually obtained using a popular two-steps routine. At the first step, a consistent estimator of θ is found by setting the weight matrix to be the identity matrix. At the second step, plug the consistent estimator of θ into (4.8) to obtain the estimated optimal W. Then, the efficient GMM estimator can be obtained by minimizing the quadratic form in (4.6) using the estimated optimal W. The HW's estimator (Huang & Wang 2001) was obtained this way.

Asymptotically, the GMM estimator is equivalent to MELE. However, Kitamura (2006) and Kunitomo & Matsushita (2003), among others, found that the MELE in general has a better finite-sample performance than the GMM estimator. Our simulations (not shown) also showed that the empirical likelihood is in general better than the GMM approach in finite samples.

4.7 Appendix

4.7.1 Proofs

Conditions for Lemma 4.2.1 and Theorem 4.2.2:

Assume that $\theta \in \Theta$ is sufficient for $-\theta \in \Theta$, and Θ is compact. Let $\beta = (\beta_z^{\mathrm{T}}, \beta_x^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{C} = (Z^{\mathrm{T}}, X^{\mathrm{T}})^{\mathrm{T}}$.

A1: $E\{F(\alpha_0+\beta_0^{\mathrm{T}}\mathbf{C})\binom{1}{\mathbf{C}}^{\otimes 2}\}$ and $E\{F(-\alpha_0-\beta_0^{\mathrm{T}}X)\binom{1}{\mathbf{C}}^{\otimes 2}\}$ are nonsingular, where $\mathbf{V}^{\otimes 2} \equiv \mathbf{V}\mathbf{V}^{\mathrm{T}}$ for a vector $\mathbf{V}$.

A2: $E(\mathbf{C}^{\mathrm{T}}\mathbf{C}) < \infty$ and $E\{\sup_{\theta\in\Theta}\mathbf{C}^{\mathrm{T}}\mathbf{C}\exp(2\beta^{\mathrm{T}}\mathbf{C})\} < \infty$.

A3: $E(U^{\mathrm{T}}U) < \infty$ and $E\{\sup_{\theta\in\Theta}\exp(2\beta^{T}U)\} < \infty$.

The proofs of Lemma 4.2.1 and Theorem 4.2.2 are similar to Huang & Wang (2001) and therefore omitted here.

Conditions for Theorem 4.3.1: Assume that the parameter space Θ is compact.

B1: θ_0 is uniquely determined by $E[g(O,\theta)] = 0$.

B2: In a neighborhood Θ of θ_0, there is a function $M(o)$ with $E(M(O)) < \infty$.

B3: $E[\partial g(O,\theta_0)/\partial\theta]$ has rank p.

B4: $E[g(O,\theta_0)g(O,\theta_0)^{\mathrm{T}}]$ is positive definite.

B5: $\partial g(o,\theta)/\partial\theta$ is continuous in Θ.

B6: $\partial^2 g(o,\theta)/\partial\theta\partial\theta^{\mathrm{T}}$ is continuous in θ in Θ.

B7: $||g(o,\theta)||^3 \leq M(o)$ in Θ.

B8: $||\partial g(o,\theta)/\partial\theta||^3 \leq M(o)$ in Θ.

B9: $||\partial^2 g(o,\theta)/\partial\theta\partial\theta^{\mathrm{T}}|| \leq M(o)$ in Θ.

The condition B1 is satisfied since $E[\Psi_{ws}(O,\theta)] = 0$ if and only if $\theta = \theta_0$. Under B2-B9, Qin & Lawless (1994) proved the asymptotical properties of maximum empirical likelihood estimators. Therefore, the results in Theorem 4.3.1 and Corollary 4.3.2 are rather standard.

4.7.2 Asymptotic relative efficiency

We give the details to generate the Table 4.1. Assume that $X \sim N(0, \sigma_x^2)$. When there are no measurement errors, the weighted-correction estimating function (4.2) reduces to the correction-amenable estimating function (4.1). For a single covariate logistic model without measurement error, (4.1) takes the form

$$\Psi(\theta) = \begin{pmatrix}\Psi_1\\ \Psi_2\end{pmatrix} = \left\{(Y-1)\exp(\frac{t}{2}) + Y\exp(-\frac{t}{2})\right\}\begin{pmatrix}1\\ X\end{pmatrix}, \tag{4.9}$$

where $t = \alpha + \beta X$. According to Lemma 4.2.1, the asymptotical variance of the estimator $\widetilde{\theta}$ is

$$AVAR(\widetilde{\theta}) = \frac{1}{n}\left[E\left(\frac{\partial\Psi}{\partial\theta}\right)^{\mathrm{T}}(E(\Psi\Psi^{\mathrm{T}}))^{-1}E\left(\frac{\partial\Psi}{\partial\theta}\right)\right]^{-1}. \tag{4.10}$$

In (4.10),

$$\begin{aligned}\frac{\partial\Psi}{\partial\theta} &= \begin{pmatrix}J_{11} & J_{12}\\ J_{21} & J_{22}\end{pmatrix}\\ &= \begin{pmatrix}\frac{\partial\Psi_1}{\partial\alpha} & \frac{\partial\Psi_1}{\partial\beta}\\ \frac{\partial\Psi_2}{\partial\alpha} & \frac{\partial\Psi_2}{\partial\beta}\end{pmatrix},\end{aligned}$$

where

$$\begin{aligned}J_{11} &= 0.5\sum_{i=1}^{n}\left\{(Y_i-1)\exp(\frac{t_i}{2}) - Y_i\exp(-\frac{t_i}{2})\right\}\\ J_{12} &= 0.5\sum_{i=1}^{n}\left\{(Y_i-1)X_i\exp(\frac{t_i}{2}) - Y_iX_i\exp(-\frac{t_i}{2})\right\}\\ J_{21} &= J_{12}\\ J_{22} &= 0.5\sum_{i=1}^{n}\left\{(Y_i-1)X_i^2\exp(\frac{t_i}{2}) - Y_iX_i^2\exp(-\frac{t_i}{2})\right\}.\end{aligned}$$

Since $Y|X \sim Bernoulli(P)$, where $P = \exp(t)/(1+\exp(t))$, $E(Y|X) = P$. Therefore,

$$\begin{aligned} E(J_{11}|X) &= -\sum_{i=1}^{n} \frac{\exp(\frac{t_i}{2})}{1+\exp(t_i)} \\ E(J_{12}|X) &= -\sum_{i=1}^{n} \frac{X_i \exp(\frac{t_i}{2})}{1+\exp(t_i)} \\ E(J_{22}|X) &= -\sum_{i=1}^{n} \frac{X_i^2 \exp(\frac{t_i}{2})}{1+\exp(t_i)} \end{aligned}$$

Then

$$E\left(\frac{\partial \Psi}{\partial \theta}\right) = E\left[E\left(\frac{\partial \Psi}{\partial \theta}|X\right)\right].$$

Similarly,

$$E(\Psi\Psi^{\mathrm{T}}) = \sum_{i=1}^{n} E\begin{pmatrix} \Psi_{1i}^2 & \Psi_{1i}\Psi_{2i} \\ \Psi_{1i}\Psi_{2i} & \Psi_{2i}^2 \end{pmatrix},$$

where

$$\Psi_{1i}^2 = (Y_i^2 - 2Y_i + 1)\exp(t_i) + Y_i^2 \exp(-t_i) + 2(Y_i^2 - Y_i),$$

$\Psi_{1i}\Psi_{2i} = \Psi_{1i}^2 X_i$ and $\Psi_{1i}\Psi_{2i} = \Psi_{1i}^2 X_i^2$. Since $E(Y^2|X) = Var(Y|X) + E^2(Y|X) = P(1-P) + P^2 = P$,

$$\begin{aligned} E(\Psi_{1i}^2|X_i) &= (P - 2P + 1)\exp(t_i) + P\exp(-t_i) \\ &= 1. \end{aligned}$$

Consequently,

$$\begin{aligned} E(\Psi_{1i}\Psi_{2i}|X_i) = X_i \\ E(\Psi_{2i}^2|X_i) = X_i^2. \end{aligned}$$

Then,

$$E(\Psi_{1i}^2) = 1$$
$$E(\Psi_{1i}\Psi_{2i}) = E\left[E(\Psi_{1i}\Psi_{2i}|X_i)\right] = E(X_i) = 0$$
$$E(\Psi_{2i}^2) = E\left[E(\Psi_{2i}^2|X_i)\right] = E(X_i^2) = \sigma_x^2.$$

Therefore,

$$E(\Psi\Psi^{\mathrm{T}}) = \begin{pmatrix} n & 0 \\ 0 & n\sigma_x^2 \end{pmatrix}.$$

For the parametric-correction estimation procedure (Huang & Wang 2001),

$$\Phi(\theta) = \begin{pmatrix} \Phi_-(\theta) \\ \Phi_+(\theta) \end{pmatrix}, \tag{4.11}$$

where

$$\begin{aligned} \Phi_-(\theta) &= \begin{pmatrix} \Phi_{1-} \\ \Phi_{2-} \end{pmatrix} = \{(Y-1) + Y\exp(-t)\}\begin{pmatrix} 1 \\ X \end{pmatrix} \\ \Phi_+(\theta) &= \begin{pmatrix} \Phi_{1+} \\ \Phi_{2+} \end{pmatrix} = \{Y + (Y-1)\exp(t)\}\begin{pmatrix} 1 \\ X \end{pmatrix}. \end{aligned}$$

the asymptotical variance of the estimator $\overline{\theta}$ is

$$AVAR(\overline{\theta}) = \frac{1}{n}\left[E\left(\frac{\partial\Phi}{\partial\theta}\right)^{\mathrm{T}}(E(\Phi\Phi^{\mathrm{T}}))^{-1}E\left(\frac{\partial\Phi}{\partial\theta}\right)\right]^{-1}. \tag{4.12}$$

Let $\frac{\partial \Phi}{\partial \theta} = M$. The elements of the 2 by 4 matrix M are of following:

$$
\begin{aligned}
M_{11} &= \frac{\partial \Phi_{1-}}{\partial \alpha} = -\sum_{i=1}^{n} Y_i \exp(t_i) \\
M_{12} &= \frac{\partial \Phi_{1-}}{\partial \beta} = -\sum_{i=1}^{n} Y_i X_i \exp(t_i) \\
M_{21} &= \frac{\partial \Phi_{2-}}{\partial \alpha} = -\sum_{i=1}^{n} Y_i X_i \exp(t_i) \\
M_{22} &= \frac{\partial \Phi_{2-}}{\partial \beta} = -\sum_{i=1}^{n} Y_i X_i^2 \exp(t_i) \\
M_{31} &= \frac{\partial \Phi_{1+}}{\partial \alpha} = \sum_{i=1}^{n} (Y_i - 1) \exp(t_i) \\
M_{32} &= \frac{\partial \Phi_{1+}}{\partial \beta} = \sum_{i=1}^{n} (Y_i - 1) X_i \exp(t_i) \\
M_{41} &= \frac{\partial \Phi_{2+}}{\partial \alpha} = \sum_{i=1}^{n} (Y_i - 1) X_i \exp(t_i) \\
M_{44} &= \frac{\partial \Phi_{2+}}{\partial \beta} = \sum_{i=1}^{n} (Y_i - 1) X_i^2 \exp(t_i)
\end{aligned}
$$

with the conditional expectation

$$
\begin{aligned}
E(M_{11}|X) &= -\sum_{i=1}^{n} \frac{1}{1 + \exp(t_i)} \\
E(M_{12}|X) &= -\sum_{i=1}^{n} \frac{X_i}{1 + \exp(t_i)} \\
E(M_{21}|X) &= -\sum_{i=1}^{n} \frac{X_i}{1 + \exp(t_i)} \\
E(M_{22}|X) &= -\sum_{i=1}^{n} \frac{X_i^2}{1 + \exp(t_i)}
\end{aligned}
$$

$$
\begin{aligned}
E(M_{31}|X) &= -\sum_{i=1}^{n} \frac{\exp(t_i)}{1+\exp(t_i)} \\
E(M_{32}|X) &= -\sum_{i=1}^{n} \frac{X_i\exp(t_i)}{1+\exp(t_i)} \\
E(M_{41}|X) &= -\sum_{i=1}^{n} \frac{X_i\exp(t_i)}{1+\exp(t_i)} \\
E(M_{42}|X) &= -\sum_{i=1}^{n} \frac{X_i^2\exp(t_i)}{1+\exp(t_i)}
\end{aligned}
$$

Straight algebra shows

$$
E(\Phi\Phi^{\mathrm{T}}|X) = \sum_{i=1}^{n} E\begin{pmatrix} \exp(-t_i) & X_i\exp(-t_i) & 1 & 0 \\ X_i\exp(-t_i) & X_i^2\exp(-t_i) & 0 & \sigma_x^2 \\ 1 & 0 & \exp(t_i) & X_i\exp(t_i) \\ 0 & \sigma_x^2 & X_i\exp(t_i) & X_i^2\exp(t_i) \end{pmatrix}
$$

Then

$$
E\left(\frac{\partial\Phi}{\partial\theta}\right) = E\left[E\left(\frac{\partial\Phi}{\partial\theta}|X\right)\right],
$$

and

$$
E(\Phi\Phi^{\mathrm{T}}) = E\left[E(\Phi\Phi^{\mathrm{T}}|X)\right].
$$

$E\left(\frac{\partial\Psi}{\partial\theta}\right)$, $E\left(\frac{\partial\Phi}{\partial\theta}\right)$, and $E(\Phi\Phi^{\mathrm{T}})$ can be evaluated numerically using QUADPACK routines. The asymptotic relative efficiency between $\widetilde{\theta}$ and $\overline{\theta}$ is then can be calculated by

$$
ARE = \frac{AVAR(\overline{\theta})}{AVAR(\widetilde{\theta})}.
$$

4.7.3 The modified Newton-Raphson procedures

In point estimation, the point estimator is usually defined as root of an unbiased estimating function. A suitable numerical algorithm is crucial to locate the point estimators successfully. In practice, it is usual that there is no analytic form for the roots of the unbiased estimating function. Therefore, finding the roots typically involves numerical iteration and is computationally intensive.

The Newton-Raphson method is the most widely used method to solve nonlinear equations $G = \sum_{i=1}^{n} g(y_i, \theta) = 0$:

$$\theta^{(k+1)} = \theta^{(k)} - a^{(k)} \left\{ \dot{G}(\theta^{(k)}) \right\}^{-1} G(\theta^{(k)}), \tag{4.13}$$

where the step size $a^{(k)} \equiv 1$. The iterations stop upon achieving a pre-specified stopping rule. The Newton-Raphson method has a quadratic converge rate and can usually find the root quickly given that the start point is not too far from the root. Since the step size $a^{(k)} \equiv 1$ is greedy, the Newton-Raphson method may fail to converge at certain situations. Therefore, some cautions need to be paid when applying it to find roots. A common modification to (4.13) is to halve the step size at each iteration if necessary until the algorithm converges. This strategy appears work well in practice. The choice of stopping rule is also very important to secure a successful root finding. Two common stopping rules are

- $||\theta^{(k+1)} - \theta^{(k)}|| < \epsilon$
- $||G(\theta^{(k+1)})|| < \epsilon$,

where $|| \cdot ||$ denotes the norm and $\epsilon > 0$ is the pre-specified tolerance level. Either stopping rule works well for well-shaped estimating functions. However, when the estimating functions are ill-behaved, for example, the conditional score and the weighted-correction estimating function, one may encounter the situations that the Newton-Raphson ends up with some non-zero-crossings based on a single stopping rule. Sometimes those non-zero-crossings can be numerically quite large. Therefore, a stopping rule that guarantees to locate the zero-crossing whenever there is one needs to be established. We found that the desired approach is to combine these two stopping rules together and it appears to work excellent in our simulations. The modified Newton-Raphson procedure has following steps:

1. $k = 0$. $a^{(0)} = 1$. $\epsilon = 10^{-8}$.

2. Calculate

$$\Delta_1(\theta^{(k)}) = G(\theta^{(k)})$$

   and

$$\Delta_2(\theta^{(k)}) = \left\{\dot{G}(\theta^{(k)})\right\}^{-1} G(\theta^{(k)})$$

   If $\max(\|\Delta_1(k)\|, \quad \|\Delta_2(k)\|) < \epsilon$, stop the iterations. $\hat{\theta} = \theta^{(k)}$. Otherwise go to next step.

3. Let $a^{(k)} = 1$. If $\|\Delta_1(\theta^{(k)} - a^{(k)}\Delta_2(\theta^{(k)}))\| > \|\Delta_1(\theta^{(k)})\|, \quad a^{(k)} = a^{(k)}/2$ and repeat this step.

4. $\theta^{(k+1)} = \theta^{(k)} - a^{(k)}\Delta_2(\theta^{(k)})$. Go to step 2 with $k = k + 1$.

We conducted extensive simulations to evaluate this modified Newton-Raphson procedure for the conditional score and weighted-correction estimating function. The results show that this procedure works excellent and never pick up non-zero-crossings due to numerical issues.

Now we give the detailed formulas to perform the modified Newton-Raphson procedure for the conditional score and weighted-correction estimating function for a single errors-in covariate logistic regression model used in Table 4.2, 4.3 and 4.5. Let the measurement error $U \sim N(0, \sigma_u^2)$. Let the Hessian matrix for each estimating funciton be

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

Conditional score

The conditional score takes the form

$$\Psi_{cs}(\theta) = \{Y - F(\alpha + \beta_x(W + (Y - 1/2)\sigma_u^2\beta_x)\}\begin{pmatrix} 1 \\ W + (Y - 1/2)\sigma_u^2\beta_x \end{pmatrix}, \tag{4.14}$$

where $F(t) \equiv \{1 + e^{-t}\}^{-1}$. The elements in the Hessian matrix H are

$$\begin{aligned}
H_{11} &= -\sum_{i=1}^{n} \frac{F(\alpha + \beta_x(W_i + (Y_i - 1/2)\sigma_u^2\beta_x)}{1 + \exp(t_i)} \\
H_{12} &= -\sum_{i=1}^{n} \frac{F(\alpha + \beta_x(W_i + (Y_i - 1/2)\sigma_u^2\beta_x)}{1 + \exp(t_i)}\{W_i + (2Y_i - 1)\sigma_u^2\beta_x\} \\
H_{21} &= -\sum_{i=1}^{n} \frac{F(\alpha + \beta_x(W_i + (Y_i - 1/2)\sigma_u^2\beta_x)}{1 + \exp(t_i)}\{W_i + (Y_i - 1/2)\sigma_u^2\beta_x\} \\
H_{22} &= \sum_{i=1}^{n}[\frac{F(\alpha + \beta_x(W_i + (Y_i - 1/2)\sigma_u^2\beta_x))}{1 + \exp(t_i)}\{W_i + (2Y_i - 1)\sigma_u^2\beta_x\}\{W_i + (Y_i - 1/2)\sigma_u^2\beta_x\} \\
&\quad +(Y_i - F(\alpha + \beta_x(W_i + (Y_i - 1/2)\sigma_u^2\beta_x)))(Y_i - 1/2)\sigma_u^2].
\end{aligned}$$

weighted-correction estimating function

The weighted-correction estimating function takes the form

$$\begin{aligned}
\Psi_{ws}(\theta) &= (Y - 1)\exp((\alpha + \beta_x W)/2 - \sigma_u^2\beta_x^2/8)\begin{pmatrix} 1 \\ W - \sigma_u^2\beta_x/2 \end{pmatrix} \\
&\quad +Y\exp(-(\alpha + \beta_x W)/2 - \sigma_u^2\beta_x^2/8)\begin{pmatrix} 1 \\ W + \sigma_u^2\beta_x/2 \end{pmatrix}. \qquad (4.15)
\end{aligned}$$

The elements in the Hessian matrix H are

$$
\begin{aligned}
H_{11} &= 0.5 \sum_{i=1}^{n} [(Y_i - 1)A_i - Y_i B_i] \\
H_{12} &= 0.5 \sum_{i=1}^{n} [(Y_i - 1)A_i C_i - Y_i B_i D_i] \\
H_{21} &= H_{12} \\
H_{22} &= 0.5 \sum_{i=1}^{n} \left[(Y_i - 1)A_i (C_i^2 - \sigma_u^2) + Y_i B_i (\sigma_u^2 - D_i^2)\right],
\end{aligned}
$$

where

$$
\begin{aligned}
A_i &= \exp((\alpha + \beta_x W_i)/2 - \sigma_u^2 \beta_x^2/8) \\
B_i &= \exp(-(\alpha + \beta_x W_i)/2 - \sigma_u^2 \beta_x^2/8) \\
C_i &= W_i - \sigma_u^2 \beta_x^2/2 \\
D_i &= W_i + \sigma_u^2 \beta_x^2/2
\end{aligned}
$$

4.7.4 Computations of Empirical Likelihood

The main task of empirical likelihood approach is to find the MELE and confidence regions for some interested parameters. Especially, empirical likelihood provides a way to find the consistent estimator when there are more estimating functions than the parameters of interest. Owen (2001) gives detailed discussions on empirical likelihood computations. Kitamura (2006) discussed valuable computational issues on empirical likelihood. Chen et al. (2002), Wu (2004, 2005), Chen et al. (2008) proposed some useful bisection and modified Newton-Raphson algorithms to compute empirical likelihood ratios. Wood et al. (1996) discussed bootstrapping empirical likelihood ratios to obtain correct critical values. Indeed, there are many different methods to compute the empirical likelihood and each method has its own merits on different situations. In this section, we will present the algorithms to find the MELE $\tilde{\theta}$ in (2.24) and its empirical likelihood confidence regions.

The objective function to be maximized is the empirical loglikelihood ratio for θ

$$l_E(\theta) = -\sum_{i=1}^{n} \log\{1 + \lambda^{\mathrm{T}} g(x_i, \theta)\}. \tag{4.16}$$

The maximization of $l_E(\theta)$ is best executed via a nested optimization routine (Owen 2001). For a fixed θ, the maximization of $l_E(\theta)$ is equivalent to the minimization of $l_E(\theta)$ over λ, an example of convex duality. Therefore, the nested optimization contains two loops:

- Inner loop minimization : $l_E(\theta)$ is minimized over λ for a fixed θ
- Outer loop maximization: $l_E(\theta)$ is maximized over θ,

where the inner loop is nested in the outer loop.

Since it is required that p_i lies between 0 and 1, $1 + \lambda^{\mathrm{T}} g(x_i, \theta) \geq 1/n$ is needed for each i. In addition, the success of the optimizations relies on the crucial condition that 0 is inside of the convex hull of $g(x_i, \theta)$'s. As long as those conditions are met, the inner loop minimization is a well-behaved convex function optimization problem and is usually done by Newton-Raphson procedure. To see why $l_E(\theta)$ is convex in λ

$$\frac{\partial l_E(\theta)}{\partial \lambda} = -\sum_{i=1}^{n} \frac{g(x_i,\theta)}{1+\lambda^{\mathrm{T}} g(x_i,\theta)}, \tag{4.17}$$

$$\frac{\partial^2 l_E(\theta)}{\partial \lambda \lambda^{\mathrm{T}}} = \sum_{i=1}^{n} \frac{g(x_i,\theta) g(x_i,\theta)^{\mathrm{T}}}{(1+\lambda^{\mathrm{T}} g(x_i,\theta))^2}. \tag{4.18}$$

As seen, the Hessian for λ is always positive definite. As a result, Newton-Raphson procedure should work well here.

The outer loop maximization is much more complicated than the inner loop minimization in that $l_E(\theta)$ is in general not concave in θ. Owen (2001) gave detailed formulas for the outer loop optimization using Newton-Raphson procedure with a warning that it can be unstable. Indeed, it is also the author's experience that Newton-Raphson procedure is not stable and can sometimes behaves unpredictable in the outer loop. The major problem is that the Hessian in the outer loop can be nearly singular, especially when the estimating functions are not well-behaved. A more stable optimization method is the simplex method by Nelder & Mead (1965). This method does not evaluate the second derivative of estimating functions and it is very reliable with a little sacrifice on efficiency compared to Newton-Raphson method. We adopted this optimization routine to perform outer loop maximization. The software **R** has a built-in function for the simplex method.

Even though the inner loop optimization is a well-behaved convex problem, some cautions are still needed and the Newton-Raphson procedure should be tuned to guarantee a successful minimization process. Let

$$\dot{l}(\lambda) = \frac{\partial l_E(\theta)}{\partial \lambda}, \quad \ddot{l}(\lambda) = \frac{\partial^2 l_E(\theta)}{\partial \lambda \lambda^{\mathrm{T}}}.$$

The modified Newton-Raphson procedure for a fixed θ in the inner loop contains following steps

1. $k = 0$. $\lambda^{(0)} = 0$, $a^{(0)} = 1$. $\epsilon = 10^{-8}$.

2. Calculate
$$\Delta(\lambda^{(k)}) = \left\{\ddot{l}(\lambda^{(k)})\right\}^{-1} \dot{l}(\lambda^{(k)})$$
If $(\|\Delta(\lambda^{(k)})\|) < \epsilon$, stop the iterations. $\hat{\lambda} = \lambda^{(k)}$. Otherwise go to next step.

3. Let $a^{(k)} = 1$. If $1+\lambda^{\mathrm{T}} g(x_i, \theta) \leq 1/n$ for some i or $l_E(\theta, \lambda^{(k)} - a^{(k)}\Delta(\lambda^{(k)})) < l_E(\theta, \lambda^{(k)})$, $a^{(k)} = a^{(k)}/2$ and repeat this step.

4. $\lambda^{(k+1)} = \lambda^{(k)} - a^{(k)}\Delta(\lambda^{(k)})$. Go to step 2 with $k = k + 1$.

The inner loop optimization usually causes no computational issues since the objective function is convex in λ. However, the inner loop optimization could fail to converge if 0 is not in the convex hull of the points $\{g_i(\theta), i = 1, \ldots, n\}$. This scenario could happen when the updated $\tilde{\theta}$ is far enough from the truth during iterations. Hence, a bad start point may cause algorithm fail to converge. Adopting the strategy suggested by Stefanski & Carroll (1987), we recommend starting from the naive estimator to find the maximum empirical likelihood estimator. Indeed, in our finite-sample simulation studies with different amounts of measurement errors, the nested optimization never failed if starting from the naive estimators.

The confidence regions for $\tilde{\theta}$ can be constructed using either normal or χ^2 approximation applying the asymptotic results in the previous section. The χ^2 calibration is usually preferred since it reflects the data-shaped confidence regions, a property that is not shared by the symmetric normal calibration that can cause undesired coverage probabilities. However, the χ^2 approximation is an asymptotic result and its validity can be distorted with small sample size. Indeed, the χ^2 confidence region is usually too conservative in finite samples. As a result, its coverage probability in finite samples can be consistently below the nominal level. A second-order correction, Bartlett correction, can reduces the coverage error of the χ^2 calibration from $O(n^{-1})$ to $O(n^{-2})$ to improve the coverage accuracy (Hall & Scala 1990, DiCiccio et al. 1991). However, such analytical correction procedures can be very difficult to perform in practice. In addition, the Bartlett correction is accurate only if n is large. in finite samples, the Bartlett correction may not be able to make much improvement. Owen (2001), among others, suggests using bootstrap empirical likelihood ratio method to obtain the confidence regions. This method also achieves an $O(n^{-2})$ coverage error asymptotically with a much better finite-sample performance than the Bartlett correction procedure.

Now we give the algorithm to obtain the 95% bootstrap empirical likelihood ratio confidence interval of the scalar β in Table 4.5. Let $\theta = (\alpha, \beta)^{\mathrm{T}}$.

1. Find the MELE $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})^{\mathrm{T}}$ from the original data using the algorithms aforementioned.

2. Draw bootstrap samples $(X_1^b, \ldots, X_n^b)$ with replacement and with equal probability from the original data, $b = 1, 2, \ldots, B$.

3. Calculate

$$R^b = -2\{\max_{\alpha} l_E(X_1^b, \ldots, X_n^b, \alpha, \tilde{\beta}) - \max_{\alpha,\beta} l_E(X_1^b, \ldots, X_n^b, \alpha, \beta)\}, \quad b = 1, 2, \ldots, B$$

   and find the 95% quantile of R^b, denotes as C_B.

4. Find β such that

$$\left\{\beta : -2\{\max_{\alpha} l_E(X_1, \ldots, X_n, \alpha, \tilde{\beta}) - \max_{\alpha,\beta} l_E(X_1, \ldots, X_n, \alpha, \beta)\} \le C_B\right\},$$

   which gives the 95% bootstrap calibration confidence interval of $\tilde{\beta}$. The standard empirical likelihood confidence interval is obtained by replacing C_B by $\chi_1^2 = 3.841$.

Chapter 5

Building Objective Functions

5.1 Introduction

Combining the conditional score with the weighted-correction estimating function provides a way to resolve the multiple-roots problem of the conditional score and yield locally efficient estimators. Even though this approach is theoretically solid, its finite-sample properties might be distorted by combining a typically less efficient estimating function with the asymptotically locally efficient conditional score. We observed that the combined procedure estimator may have noticeable skewness (Figure 4.3 and 4.4). In addition, our simulation did not show an obvious efficiency gain by combining two estimating functions (Table 4.3 and Table 4.5). In fact, we usually found that the combined procedure estimator has larger estimated standard deviations than those of the conditional score estimator.

In theory, combining estimating functions would not result in first-order efficiency loss. Typically, possible efficiency gains may be expected by combining estimating functions. However, this desirable efficiency property is based on first-order asymptotical theory. In finite samples, combining estimating functions may reduce the second-order efficiency that can offset the first-order efficiency gains (Qu et al. 2008). In our case, the conditional score is already locally efficient. By including a typically less efficient weighted-correction estimating function, we may encounter efficiency loss, as shown in the simulation studies (Table 4.3 and Table 4.5).

Another approach to resolve the multiple-roots problem of the conditional score is to build an objective function to distinguish among multiple roots of the conditional score. Once the consistent estimator of the conditional score is identified, the local efficiency is secured. This approach could avoid the potential efficiency loss of the combining approach in Chapter 4.

In the presence of multiple solutions, Stefanski & Carroll (1987) suggested choosing the root closest to the naive estimator. This approach usually works well unless measurement errors are too large. However, this approach is rather heuristic, and lacks of solid theoretical justifications compared with maximum likelihood estimation. To resolve this issue, Hanfelt & Liang (1997) constructed an objective function, called the conditional quasi-likelihood, by performing a line integral on the conditional score. The correct solution to the conditional score is the one maximizing the conditional quasi-likelihood in large samples. This approach appears to work well in practice. Nevertheless, the conditional quasi-likelihood is in general not unique in finite smaples.

In this chapter, we aim to develop objective functions to distinguish among multiple roots of the conditional score.

5.2 Conditional quasi-likelihood

Building an objective function to distinguish among multiple roots for estimating functions is a very attractive approach since it shares the merit of maximum likelihood estimation. However, since most of the unbiased estimating functions are not conservative, they can not be written as derivatives of objective functions. As a result, building objective functions is nontrivial in general.

Inspired by the unconditional quasi-likelihood approach (Wedderburn 1974, McCullagh 1983), Hanfelt & Liang (1997) proposed to integrate the conditional score (Stefanski & Carroll 1987) to form the conditional quasi-likelihood. For an arbitrary path $\theta(s), a \le s \le b$, in $\theta-$space from two values θ and η, the conditional quasi-likelihood ratio of θ to η is give by the line integral

$$Q_{\theta(s)}(\theta, \eta) = \int_{\theta(a)}^{\theta(b)} \Psi_{cs}(\theta(s))\, \mathrm{d}\theta(s). \tag{5.1}$$

For the logistic model, the conditional score (Stefanski & Carroll 1987) takes the form

$$\Psi_{cs}(\theta) = \{Y - F(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}}(W + (Y - 1/2)\Sigma_{uu}\beta_x)\}\begin{pmatrix} 1 \\ Z \\ t(\Delta) \end{pmatrix}. \tag{5.2}$$

Since the conditional score is not conservative:

$$\frac{\partial \Psi_j(\theta)}{\partial \theta_i^{\mathrm{T}}} \neq \frac{\partial \Psi_i(\theta)}{\partial \theta_j^{\mathrm{T}}} \qquad \text{for} \quad i \neq j,$$

it can not be written as the derivative of an objective function in general. To overcome this issue, Hanfelt & Liang (1997) carefully chose

$$t(\Delta) = \Delta + (E(Y|\Delta) - 1))\Sigma_{uu}\beta_x. \tag{5.3}$$

The specific choice of $t(\Delta)$ (5.3) makes $Q_{\theta(s)}$ to be independent of path $\theta(s)$ as $n \to \infty$.

Therefore, one could apply $Q_{\theta(s)}$ to distinguish among multiple roots in large samples. Hanfelt & Liang (1997) showed that in large samples the consistent root of Ψ_{cs} is corresponding to a local maximizer of $Q_{\theta(s)}$. On the other hand, local minimizers of $Q_{\theta(s)}$ are corresponding to inconsistent roots.

In finite samples, $Q_{\theta(s)}$ is typically path-dependent and not unique. However, simulations results suggest that $Q_{\theta(s)}$ works very well under large measurement errors (Hanfelt & Liang 1997).

The conditional quasi-likelihood is a valuable method to resolve the multiple-roots problem of the conditional score. Even though its finite-sample performance is very good, it is typically path-dependent in finite samples and therefore, remains an incomplete solution to the multiple-roots problem.

5.3 The objective functions

In this section, we construct two objective functions that are based on the weighted-correction estimating function $\Psi_{ws}(\theta)$ (4.2). We will show that these two objective functions are well defined in a compact parameter space and can be used to distinguish among multiple roots of the conditional score in large samples.

In section 4.2, we developed $\Psi_{ws}(\theta)$ by performing correction on the correction-amenable estimating function $\Psi(\theta)$ (4.1). However, $\Psi_{ws}(\theta)$ can also be developed in terms of objective function. It is easy to show that $\Psi(\theta)$ can be written as the derivative of the following objective function

$$K(Y,X,Z,\theta) = 2\left\{(Y-1)\exp(\frac{\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}X}{2}) - Y\exp(-\frac{\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}X}{2})\right\}. \tag{5.4}$$

Similar to the score function (2.3), $K(\theta)$ is strictly concave since its Hessian matrix is always negative definite for all θ:

$$H_K(Y,X,Z,\theta) = 0.5\left\{(Y-1)\exp(\frac{\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}X}{2}) - Y\exp(-\frac{\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}X}{2})\right\}\begin{pmatrix}Z\\X\end{pmatrix}\begin{pmatrix}Z\\X\end{pmatrix}^{\mathrm{T}}. \tag{5.5}$$

In the presence of measurement error, the corrected score based on the logistic regression loglikelihood function does not exist (Stefanski 1989). However, one could perform correction on $K(Y,X,Z,\theta)$. Under additive measurement error model, The resulting objective function $Q(Y,W,Z,\theta)$ is given by

$$Q(Y,W,Z,\theta) = 2\left\{\frac{(Y-1)\exp((\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}W)/2)}{E(\exp(\beta_x^{\mathrm{T}}U/2))} - \frac{Y\exp(-(\alpha+\beta_z^{\mathrm{T}}Z+\beta_x^{\mathrm{T}}W)/2)}{E(\exp(-\beta_x^{\mathrm{T}}U/2))}\right\}, \tag{5.6}$$

which satisfies

$$E\{Q(Y,W,Z,\theta)|(Y,X,Z)\} = K(Y,X,Z,\theta).$$

That is, $Q(Y,W,Z,\theta)$ and $K(Y,X,Z,\theta)$ achieve the same limit, and $Q(Y,W,Z,\theta)$ is concave at the limit. The weighted-correction estimating function $\Psi_{ws}(\theta)$ can then be obtained by take the derivative of $Q(Y, W, Z, \theta)$ with respect to θ. It can be shown that $Q(Y, W, Z, \theta)$ is non-positive. As usual, the measurement error is often assumed to be normally distributed: $U \sim N(0, \Sigma_{uu})$. Then $Q(Y, W, Z, \theta)$ becomes to

$$Q(Y, W, Z, \theta) = 2\left\{\frac{(Y-1)\exp((\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} W)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)} - \frac{Y\exp(-(\alpha + \beta_z^{\mathrm{T}} Z + \beta_x^{\mathrm{T}} W)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)}\right\}. \tag{5.7}$$

We define the corrected quasi-likelihood as:

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, w_i, z_i, \theta). \tag{5.8}$$

For convenience, we use the notation $Q(\theta)$ to denote $Q(Y, W, Z, \theta)$ (5.7). The existence of an objective function $Q_n(\theta)$ for $\Psi_{ws}(\theta)$ makes it possible to define the estimator of $\Psi_{ws}(\theta)$ as the maximizer of $Q_n(\theta)$. A potential issue here is that $Q_n(\theta)$ has a correction factor $\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)$. As a result, $Q_n(\theta)$ will be maximized towards zero as $|\beta_x| \to \infty$. However, if the parameter space is compact, the estimator of $\Psi_{ws}(\theta)$ can be defined as the maximizer of $Q_n(\theta)$. Similar scenarios were also found for the corrected score (Nakamura 1990, Stefanski 1989) that is derived as the derivative of corrected loglikelihood functions (2.8). Unlike the loglikelihood functions in the absence of measurement errors, the corrected loglikelihood functions are, in general, unbounded as $|\beta_x| \to \infty$ (Nakamura 1990, Stefanski 1989). Therefore, the corrected score estimator may not be regarded as the maximizer of the corrected loglikelihood functions either unless the parameter space is compact.

In a compact parameter space, the maximizer of $Q_n(\theta)$ is consistent. The maximizer can be found by solving the weighted-correction estimating function. By the fact that both the weighted-correction estimating function and conditional score yield consistent roots of the same parameter, one can use $Q_n(\theta)$ as an objective function to distinguish among multiple roots of the conditional score in large samples.

Theorem 5.3.1 *Assume that there exists a consistent root for the conditional score. Then, among all the roots of the conditional score as n $\to \infty$, the maximizer of $Q_n(\theta)$ is consistent if following regularity conditions are met:*

1. *The parameter space $\theta \in \Theta$ is compact*

2. *X and Z are bounded*

In addition to solve the unbiased estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \Psi_{ws}(y_i, w_i, z_i, \theta) = 0, \tag{5.9}$$

one can also use empirical likelihood method to obtain the weighted-correction estimating function estimator $\hat{\theta}_{ws}$. Based on $\Psi_{ws}(\theta)$ only, the profile empirical likelihood function for θ is

$$L(\theta) = \sup\left\{\prod_{i=1}^{n} p_i \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\Psi_{ws}(y_i, w_i, z_i, \theta) = 0\right\}.$$

Using Lagrange multipliers (Section 2.3.2), one can show that the empirical loglikelihood ratio for θ is

$$l_{ws}(\theta) = -\sum_{i=1}^{n} \log\{1 + \lambda^{\mathrm{T}}\Psi_{ws}(y_i, w_i, z_i, \theta)\},$$

where λ is the Lagrange multiplier that solves

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\Psi_{ws}(y_i, w_i, z_i, \theta)}{1 + \lambda^{\mathrm{T}}\Psi_{ws}(y_i, w_i, z_i, \theta)} = 0$$

for fixed θ. Note that λ is a continuous differentiable function of θ (Qin & Lawless 1994). Maximizing the objective function $l_{ws}(\theta)$ yields the maximum empirical likelihood estimator for θ. Since the parameters and estimating functions have the same dimension, the resulting maximum empirical likelihood estimator is the same as the one obtained by solving the equation (5.9). For any roots of $\Psi_{ws}(\theta)$, $l_{ws}(\theta)$ will attain its maximum value of 0. As n $\to \infty$, $l_{ws}(\theta)$ will be only maximized to 0 by its consistent roots in a compact parameter space. Define the corrected empirical likelihood

$$D_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\{1 + \lambda^{\mathrm{T}}\Psi_{ws}(y_i, w_i, z_i, \theta)\}. \tag{5.10}$$

Since $\Psi_{ws}(y_i, w_i, z_i, \theta)$ has a unique zero crossing at θ_0 at the limit, $D_n(\theta)$ has the desired concavity property at the limit with the maximum value of zero at θ_0. Again, by the fact that both the weighted-correction estimating function and conditional score yield consistent roots of the same true parameter, one may use $D_n(\theta)$ as an objective function to distinguish among multiple roots of the conditional score in large samples.

We conjecture that under the same conditions in Theorem 5.3.1, among all the roots of the conditional score as $n \to \infty$, the maximizer of $D_n(\theta)$ is consistent (See Section 5.5 for a discussion on this conjecture).

In finite samples, both approaches have their own advantages and disadvantages. The weighted-correction estimating function $\Psi_{ws}(\theta)$ could have multiple solutions. $Q_n(\theta)$ and $D_n(\theta)$ behave different at those multiple solutions. As a matter of fact, the correct root of $\Psi_{ws}(\theta)$ is corresponding to a local maximizer of $Q_n(\theta)$, whereas local minimizers of $Q_n(\theta)$ are corresponding to improper roots. Indeed, for a correct root of $\Psi_{ws}(\theta)$, its observed information matrix must be positive-definite. On the other hand, $D_n(\theta)$ will be maximized to zero at all of roots of $\Psi_{ws}(\theta)$. That is, $Q_n(\theta)$ provides more information about the roots of $\Psi_{ws}(\theta)$ than $D_n(\theta)$ in the sense of positive or negative definiteness of the observed information matrix at roots .

However, due to the existence of the correction factor $\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)$, $Q_n(\theta)$ might be maximized towards zero for arbitrarily large $|\beta_x|$. In addition, when the measurement error is large enough, $Q_n(\theta)$ might also be maximized towards zero because of the correction factor. On the other hand, $D_n(\theta)$ does not necessarily become larger as $|\beta_x|$ becomes larger. Therefore, $D_n(\theta)$ may have a better tail behavior than $Q_n(\theta)$ in finite samples, which could be crucial for the ability to distinguish among conditional score roots in practice.

To better understand the finite-sample behaviors of $Q_n(\theta)$ and $D_n(\theta)$, let's consider a single and error-prone covariate logistic model with sample size of 200. True values of (α_0, β_0) considered are $(0, 1)$. Both the true covariate and the additive measurement error follow a standard normal distribution. Figure 5.1 plots out the weighted-correction estimating function $\Psi_{ws}(\theta)$, the corrected quasi-likelihood $Q_n(\theta)$, and the corrected empirical likelihood $D(\theta)$ with respect to β. In this example, $\Psi_{ws}(\theta)$ has three roots. only The middle one ($\hat{\beta} = 0.94$) appears to be the correct root since it is the only root that has positive-definite observed information matrix. As seen from the graph of $Q_n(\theta)$, this root ($\hat{\beta} = 0.94$) is corresponding to a local maximizer of $Q_n(\theta)$, whereas the other two roots are corresponding to local minimizers of $Q_n(\theta)$. On the other hand, $D_n(\theta)$ is maximized to zero at all of those three roots. Examining the tails of $Q_n(\theta)$ and $D_n(\theta)$, we can find that the tails of $Q_n(\theta)$ go up as $|\beta|$ increases, whereas the tails of $D_n(\theta)$ go down eventually as $|\beta|$ increases.

In the next section, the abilities of $Q_n(\theta)$ and $D_n(\theta)$ to distinguish among roots of the conditional score will be investigated through simulations.

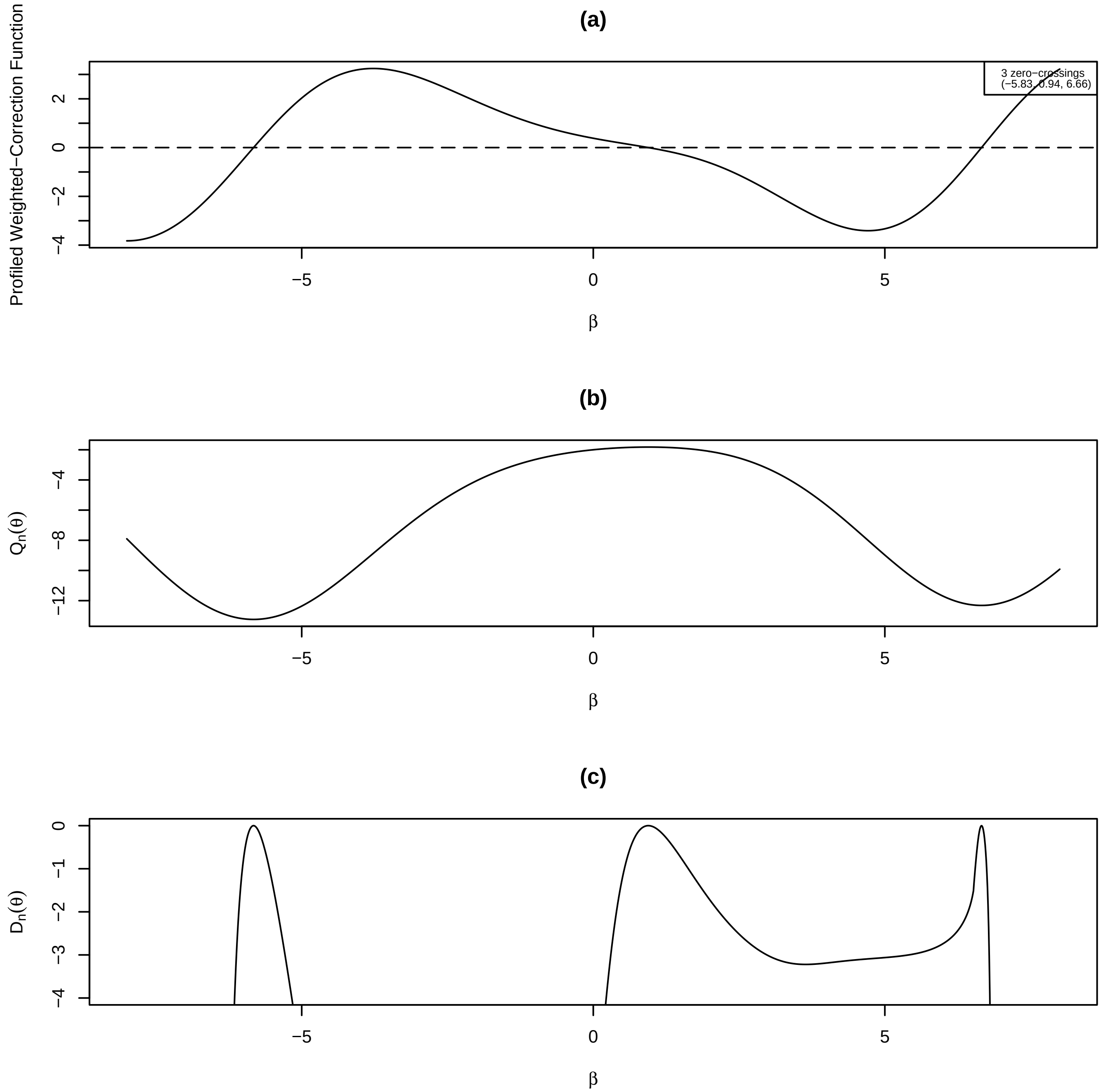

Figure 5.1: *An illustration of $Q_n(\theta)$ and $D_n(\theta)$. The weighted-correction estimating function has three roots. Plots in this graph are (a) profiled weighted-correction estimating function, (b) $Q_n(\theta)$, and (c) $D_n(\theta)$. For this sample data: N=200. True values $(\alpha_0, \beta_0) = (0, 1)$. Both the true covariate and the additive measurement error follow a standard normal distribution.*

5.4 Distinguish among multiple roots

5.4.1 Simulations

We conduct three simulation studies to investigate the abilities of $Q_n(\theta)$ and $D_n(\theta)$ to distinguish among multiple roots of the conditional score in finite samples. To apply objective functions to the conditional score, one wants to find all the possible roots of the conditional score. We adopted the random start points approach (Robbins 1968, Thode et al. 1987, Finch et al. 1989). This approach was also adopted by Hanfelt & Liang (1997) in their study of the conditional quasi-likelihood. For each simulated data set, we search all possible roots using 8000 random start points from $\{-6 \leq \theta \leq 6\}$. Then the probability of a new root remains unfound is less than 0.05.

In Chapter 3, we have discussed that the conditional score may fail to generate a good root when the measurement error is large. Typically, the conditional score yields a single but bad root (Figure 3.4) at that situation. In this simulation study, we focus on the cases where the conditional score yield at least a good root. Indeed, this is a prerequisite for applying an objective function to distinguish among roots.

In our simulation studies, we evaluate the values of $Q_n(\theta)$ and $D_n(\theta)$ at each conditional score root. The root being selected by each objective function is the one maximizing the corresponding objective function. The performance of each objective function will be evaluated by its successful rate (%) of identifying the correct conditional score root, which is defined as the one closest to the truth in Euclidean distance.

In the first simulation study, we consider a single measured with error covariate logistic regression. The true covariate $X \sim N(0, 1)$. A large measurement error is considered: $\varepsilon \sim N(0, 1)$. The true values $\theta_0 = (0, 1)$. We simulated 1000 data sets with a sample size of sizes of 200. 17 samples fail to generate a good conditional score root. That is, the conditional score yields a single but bad root (Figure 3.4). In other 973 samples, the conditional score generated 3 roots for each data set. Similar to the conditional score, the weighted-correction estimating function yields either one or three roots. When only one root is generated, that single root is a bad root and far from the truth. We also found out whenever the conditional score has a single but bad root, the weighted-correction estimating function yields a single but bad root for sure. This indicated that the conditional score is more computationally stable than the weighted-correction estimating function. For the weighted-correction estimating function, 84 samples fail to generate a good root. Table 5.1 shows the multiple conditional score roots of a simulated data set. Each roots are evaluated for their values of $Q_n(\theta)$ and $D_n(\theta)$. As seen, both $Q_n(\theta)$ and $D_n(\theta)$ successfully identify the root ($\hat{\theta}_1$) closest to θ_0 in Euclidean distance. $Q_n(\theta)$ and $D_n(\theta)$ had a successful rate of 89.9% and 98.3%, respectively, to identify the correct rootin 973 samples. The heuristic procedure by Stefanski & Carroll (1987) identifies all of the correct roots in 973 samples.

Table 5.1: *Distinguish among multiple roots of the conditional score: one sample from simulation study 1*

Multiple Roots	d	$Q_n(\theta)$	$D_n(\theta)$
$\hat{\theta}_1$ = (0.11,1.08)	0.14	-1.83	-0.07
$\hat{\theta}_2$ = (-3.27,6.74)	6.61	-12.32	-0.89
$\hat{\theta}_3$ = (0.64,-6.12)	7.15	-13.14	-3.14
Naive Estimator			
$\hat{\theta}_N$ = (-0.31,0.71)	0.42		

Note: d: Euclidean distance to θ_0. $n = 200$. $\theta_0 = (0, 1)$. $x_i \sim N(0, 1)$ and the measurement errors $u_i \sim N(0, 1)$. The naive estimator $\hat{\theta}_N$ ignores the measurement error.

We now draw graphs to illustrate how $Q_n(\theta)$ and $D_n(\theta)$ distinguish among conditional score roots in simulation study 1. For illustration purpose, we profile the intercept out and only draw the plots with respect to slope β only.

When the weighted-correction estimating function yields a single root, $Q_n(\theta)$ would fail to identify the correct root of the conditional score for sure in simulation study 1 (Figure 5.2 and 5.3) since $Q_n(\theta)$ would just keep increasing at the tail. On the other hand, $D_n(\theta)$, even though fails in some cases (Figure 5.2), can pick up the correct root most of the time (Figure 5.3). The reason is that $D_n(\theta)$ typically have a local maximizer around the correct conditional score root even though the local maximizer can not be zero at that case. That local maximizer may still provide some information to help to identify the correct conditional score root.

When the weighted-correction estimating function yields three roots. Both $Q_n(\theta)$ (Figure 5.4) and $D_n(\theta)$ (Figure 5.4 and 5.5) have high successful rates to identify the correct root. However, $Q_n(\theta)$ is more likely to fail than $D_n(\theta)$ does (Figure 5.5). The cases that both $Q_n(\theta)$ and $D_n(\theta)$ fail to pick up the correct root, or $Q_n(\theta)$ succeeds but $D_n(\theta)$ fails to identify the correct root are extremely rare when the weighted-correction estimating function yields three roots.

The second simulation study has a sample size of 300. We simulated 500 data sets . Four covariates were considered with two of them measured with errors. The true parameter values are $\theta_0 = (0, 1, -0.4, 0.4, -1)$. Four covariates follows a multivariate normal distribution: $X_i \sim N_4(0, \tau I)$ for $i = 1, 2, 3, 4.$ and $\tau = 1$. x_1 and x_2 are measured with zero mean normal errors with variances of $\tau/2$ and $\tau/4$, respectively. The measurement errors in this model are substantial. The conditional score generated 3-11 roots with 5 roots in average. Table (5.2) shows a successful example of $Q_n(\theta)$ and $D_n(\theta)$ to identify the correct root. In the 500 simulations, both $Q_n(\theta)$ and $D_n(\theta)$ identify the correct roots 99.4% of time. The heuristic procedure by Stefanski & Carroll (1987) identifies all of the correct roots in 500 samples.

The third simulation study was used by Hanfelt & Liang (1997) to investigate the performances of the conditional quasi-likelihood to distinguish among conditional score multiple roots. The sample size they considered is 200. 500 data sets were simulated. The true $\theta_0 = (-1.4, 1.4, -0.3, 0.3, 0.6)$. The true covariates $X_i \sim N_4(0, \tau I)$ and the measurement errors $U_i \sim N_4(0, \Sigma_{uu})$, where $\tau = 0.1$ and $\Sigma_{uu} = \mathrm{diag}\{\tau/3, \tau, \tau, \tau\}$. Hanfelt & Liang (1997) found that 490 samples yielded roots in the interior of the parameter space and an average of over seven roots were generated by the conditional score. The conditional quasi-likelihood (Hanfelt & Liang 1997) successfully identified the correct root in all 490

Table 5.2: *Distinguish among multiple roots of the conditional score: one sample from simulation study 2*

Multiple Roots	d	$Q_n(\theta)$	$D_n(\theta)$
$\hat{\theta}_1$ = (-0.04,0.96,-0.29,0.46,-0.98)	0.14	-1.67	-0.12
$\hat{\theta}_2$ = (1.11, -8.8, -5.75, -1.81,-1.14)	11.44	-24.85	-4.25
$\hat{\theta}_3$ = (0.1,6.49,-9.92,4.69,-3.45)	12.05	-17.42	-3.85
$\hat{\theta}_4$ = (-3.45,9.76,-5.58,3.08,-6.31)	12.28	-32.90	-4.13
$\hat{\theta}_5$ = (-5.44,10.03,-0.38,0.04,-7.44)	12.36	-22.84	-1.84
Naive Estimator			
$\hat{\theta}_N$ = (-0.02, 0.58, -0.21, 0.42, -0.93)	0.47		

Note: d: Euclidean distance to θ_0. $n = 300$. $\theta_0 = (0, 1, -0.4, 0.4, -1)$. $X_i \sim N_4(0, \tau I)$ and the measurement errors $U_i \sim N_4(0, \Sigma_{uu})$, where $\tau = 1$ and $\Sigma_{uu} = \text{diag}\{\tau/2, \tau/4, 0, 0\}$. The naive estimator $\hat{\theta}_N$ ignores the measurement error.

samples. We adopted the same set ups as in Hanfelt & Liang (1997) to compare three approaches. Table(5.3) shows the multiple conditional score roots of a simulated data set. In our simulated data, 10 samples fail to generate a good root and an average of over 9 roots were founded in other 490 samples for the conditional score. On the other hand, the weighted-correction estimating function fails to generate a good root in 276 samples. $Q_n(\theta)$ and $D_n(\theta)$ had a successful rate of 90.0% and 99.0%, respectively, to identify the correct root. An interested finding is that $Q_n(\theta)$ can still identify some correct conditional score roots even if the weighted-correction estimating function fails to generate a good root. Recall that in the simulation study 1 where the model has a single covariate, $Q_n(\theta)$ would fail for sure. Indeed, the finite-sample behavior of $Q_n(\theta)$ is complicated under multiple covariates and may deserve more investigations in future study. The heuristic procedure by Stefanski & Carroll (1987) identifies all of the correct roots in 490 samples.

According to these 3 simulation studies, $D_n(\theta)$ appears to perform better than $Q_n(\theta)$ to distinguish among roots of the conditional score in finite samples (Table 5.4). Indeed, when the measurement error is large enough, the weighted-correction estimating function is more likely to fail to yield a good root. This fact could reduce the ability of $Q_n(\theta)$, which is heavily depends on the behave of the weighted-correction estimating function, to distinguish among conditional score roots. For example, in simulation study 1 and 3. We also

Table 5.3: *Distinguish among multiple roots of the conditional score: one sample from simulation study 3*

Multiple Roots	d	$Q_n(\theta)$	$D_n(\theta)$
$\hat{\theta}_1$ = (-1.22,1.50,-0.37,0.13,0.80)	0.34	-1.66	-0.06
$\hat{\theta}_2$ = (-1.51,1.82,-9.27,4.65,6.92)	11.81	-2.40	-1.88
$\hat{\theta}_3$ = (-1.04,-2.52,-5.81,-8.40,8.48)	13.55	-2.52	-0.57
$\hat{\theta}_4$ = (-0.96,-1.1,-11.22,-4.13,8.82)	14.59	-2.65	-2.07
$\hat{\theta}_5$ = (-2.37,1.75,12.54,3.78,-5.55)	14.69	-2.96	-1.27
$\hat{\theta}_6$ = (-0.91,-2.13,-9.38,-11.79,-2.85)	15.91	-2.84	-8.61
$\hat{\theta}_7$ = (-0.47, -3.09, -11.2, -11.37,0.89)	16.62	-2.88	-4.15
$\hat{\theta}_8$ = (-0.91,-5.17,-2.01,5.52,-14.77)	17.60	-4.87	-3.95
$\hat{\theta}_9$ = (-1.45,-3.17,-9.87,-1.04,-14.32)	18.35	-5.07	-2.02
$\hat{\theta}_{10}$ = (-1.21,-2.8,-13.4,-1.89,-11.91)	18.72	-4.90	-4.89
$\hat{\theta}_{11}$ = (-0.95,5.30,12.29,-9.59,-8.51)	18.83	-6.30	-4.95
$\hat{\theta}_{12}$ = (-2.58,-5.34,6.45,8.23,-13.71)	18.97	-5.14	-2.13
Naive Estimator			
$\hat{\theta}_N$ = (-1.19, 1.19, -0.18, 0.09, 0.35)	0.46		

Note: d: Euclidean distance to θ_0. $n = 200$. $\theta_0 = (-1.4, 1.4, -0.3, 0.3, 0.6)$. $X_i \sim N_4(0, \tau I)$ and the measurement errors $U_i \sim N_4(0, \Sigma_{uu})$, where $\tau = 0.1$ and Σ_{uu} =diag$\{\tau/3, \tau, \tau, \tau\}$. The naive estimator $\hat{\theta}_N$ ignores the measurement error.

conducted more simulations on different sample sizes and amounts of measurement errors. The results show that $D_n(\theta)$ performs consistently better than $Q_n(\theta)$, and its successful rates are unanimously high.

In summary, $D_n(\theta)$ has a satisfactory performance on distinguishing among multiple roots of the conditional score in finite samples. We suggested using the corrected empirical likelihood $D_n(\theta)$ to distinguish among multiple roots of the conditional score for logistic regression with errors-in-covariates in practice.

Table 5.4: *The successful rates (%) of identifying correct conditional score roots using $Q_n(\theta)$ and $D_n(\theta)$ in the three simulation studies*

Simulation	N	P	R	$Q_n(\theta)$	$D_n(\theta)$	S&C
1	200	1	3	89.9	98.3	100
2	300	4	5	99.4	99.4	100
3	200	4	9	90.9	99.0	100

Note: S&C: The heuristic procedure by Stefanski & Carroll (1987). P: number of covariates. R: average number of roots

5.4.2 A High Blood Pressure study

We apply the proposed objective function $D_n(\theta)$, along with the $Q_n(\theta)$ to a real study. In this study, the relationships between 24-hour urinary sodium chloride and blood pressure were investigated in 397 middle-aged Chinese men living in Taipe by Pan et al. (1990). Hanfelt & Liang (1997) used this example to investigate their conditional quasi-likelihood. Seven overnight urinary sodium chloride measurements had a within-subject variability of 1.2 when standardized so that the mean urinary sodium chloride measurements had a variability of 1 across the 397 subjects (Hanfelt & Liang 1997). A logistic regression model with high systolic blood pressure as outcome and the 24-hour urinary sodium chloride measurement, plus age and body mass index (BMI) as covariates was adopted. The age and BMI were standardized by minus 53.2 years and 23.5, respectively. The measurement error in the model is induced since the true 24-hour urinary sodium chloride measurement is not measurable. In the model, a single (the most recent) urinary sodium chloride measurement was used. Therefore, the variance of the measurement error can be assumed to be 1.2 as aforementioned. The estimated variance of the most recent urinary sodium chloride measurement is 1.82. Then the variance of the unobservable true covariate can be estimated by 0.62 under additive measurement error assumption. The amount of the measurement error is very large (194% variation compared to the true covariate). The age and BMI are assumed to be measured with a very small amount of error due to rounding off. Their measurement error variances are assumed to be 0.083. Three roots were found by the conditional score approach. Table 5.5 shows the multiple roots and the corresponding values of $Q_n(\theta)$ and $D_n(\theta)$. Both objective functions identify $\hat{\theta}_1$ to be the correct root, which is also the closest root to the naive estimator. Hanfelt & Liang (1997) also identified $\hat{\theta}_1$ as the correct root. Therefore, the estimated effects of 24-hour urinary sodium chloride, the standardized age and BMI are 0.61, 0.08, and 0.26, respectively.

Table 5.5: *Distinguish among multiple roots of the conditional score: A High Blood Pressure study*

Multiple Roots	$d^{\dagger}$	$Q_n(\theta)$	$D_n(\theta)$
$\hat{\theta}_1$ = (-2.91,0.61,0.08,0.26)	0.29	-0.92	-0.15
$\hat{\theta}_2$ = (1.24,-4.47,-0.01,0.05)	6.32	-4.39	-2.76
$\hat{\theta}_3$ = (-11.26,6.61,-0.30,-0.92)	10.85	-19.04	-2.40

Note: $d^{\dagger}$: Euclidean distance to the naive estimator $\hat{\theta}_N = (-2.86, 0.33, 0.08, 0.28)$. The measurement errors $U \sim N_3(0, \Sigma_{uu})$, where $\Sigma_{uu} = \mathrm{diag}\{1.2, 0.083, 0.083\}$.

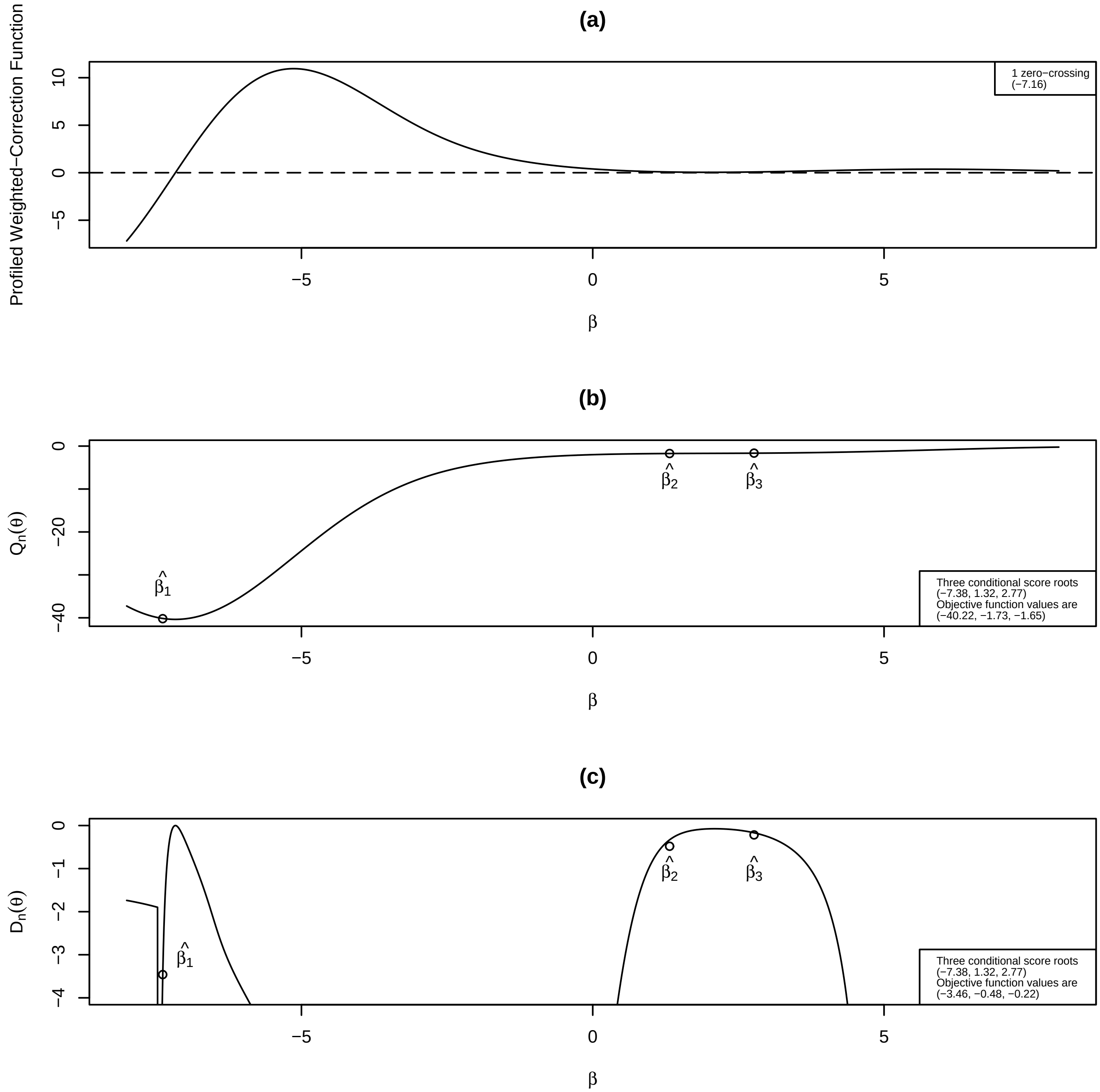

Figure 5.2: *Distinguish among multiple conditional score roots using $Q_n(\theta)$ and $D_n(\theta)$. The weighted-correction estimating function yields a single root. Both $Q_n(\theta)$ and $D_n(\theta)$ fail to identify the correct root. Plots in this graph are (a) profiled weighted-correction estimating function, (b) $Q_n(\theta)$, and (c) $D_n(\theta)$. For this sample data: N=200. True values $(\alpha_0, \beta_0) = (0, 1)$. Both the true covariate and the additive measurement error follow a standard normal distribution.*

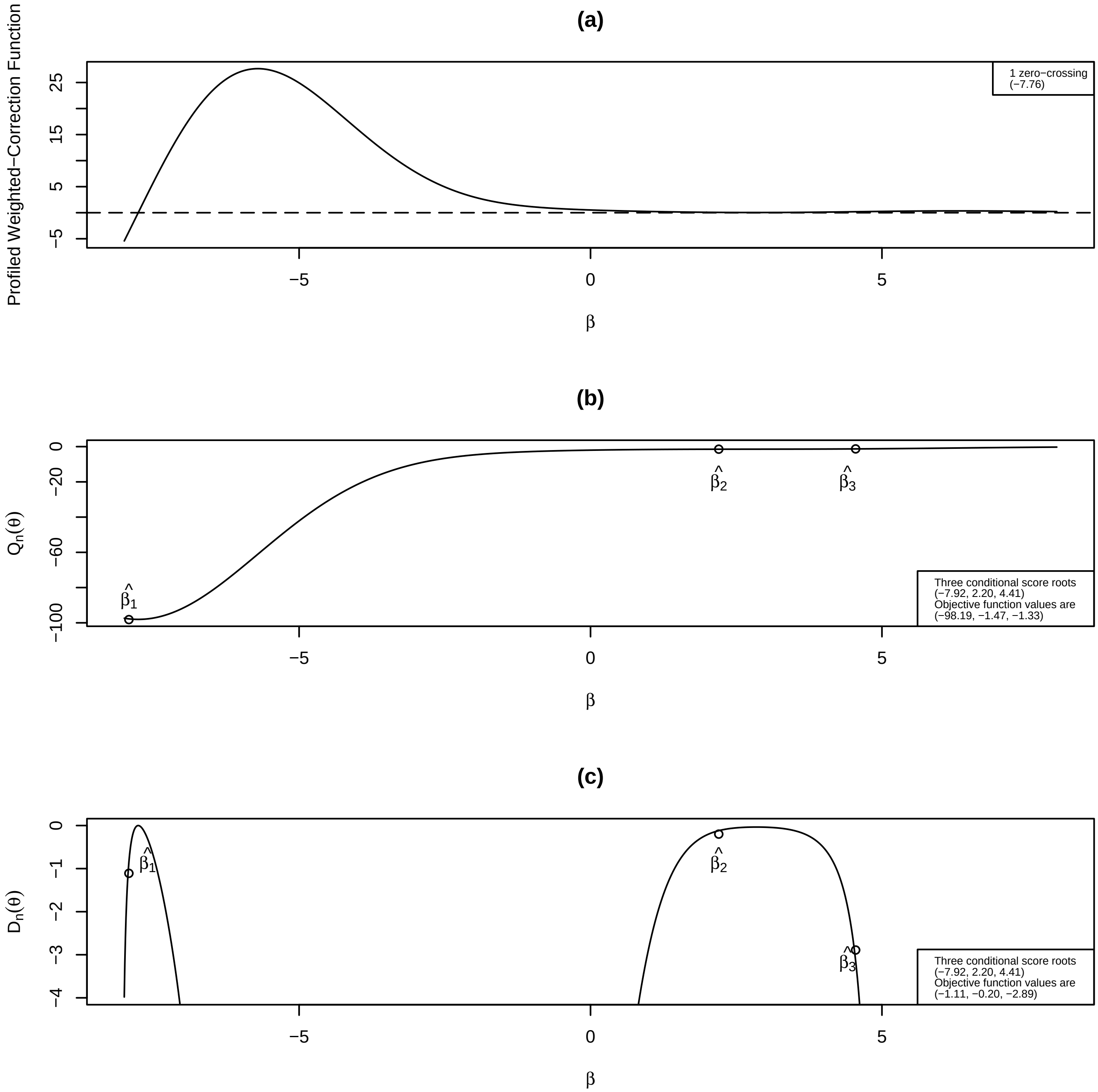

Figure 5.3: *Distinguish among multiple conditional score roots using $Q_n(\theta)$ and $D_n(\theta)$. The weighted-correction estimating function yields a single root. $Q_n(\theta)$ fails, whereas $D_n(\theta)$ succeeds to identify the correct root. Plots in this graph are (a) profiled weighted-correction estimating function, (b) $Q_n(\theta)$, and (c) $D_n(\theta)$. For this sample data: N=200. True values* $(\alpha_0, \beta_0) = (0, 1)$. *Both the true covariate and the additive measurement error follow a standard normal distribution.*

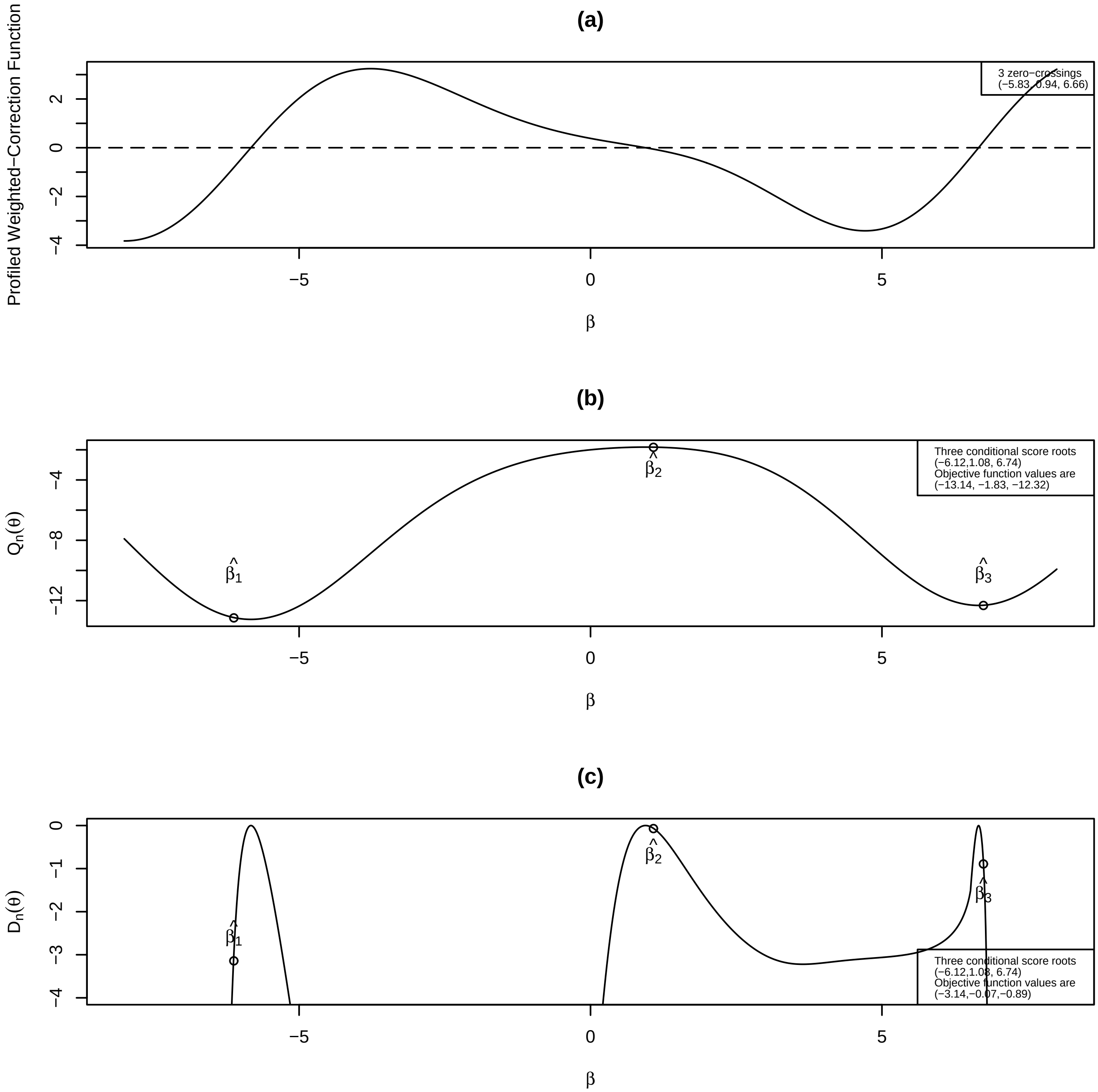

Figure 5.4: *Distinguish among multiple conditional score roots using $Q_n(\theta)$ and $D_n(\theta)$. The weighted-correction estimating function yields three roots. Both $Q_n(\theta)$ and $D_n(\theta)$ succeed to identify the correct root. Plots in this graph are (a) profiled weighted-correction estimating function, (b) $Q_n(\theta)$, and (c) $D_n(\theta)$. For this sample data: N=200. True values* $(\alpha_0, \beta_0) = (0, 1)$. *Both the true covariate and the additive measurement error follow a standard normal distribution.*

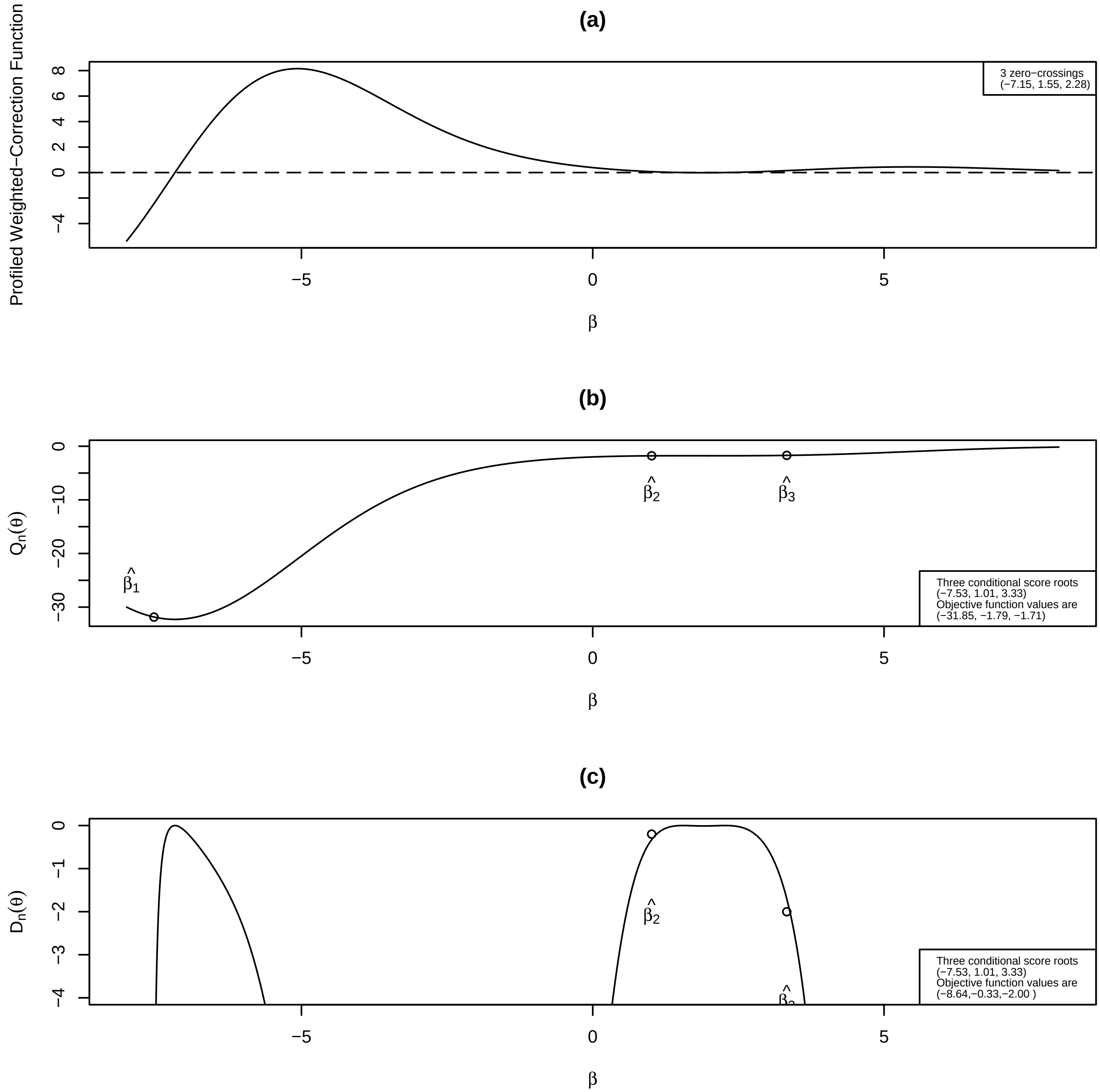

Figure 5.5: *Distinguish among multiple conditional score roots using $Q_n(\theta)$ and $D_n(\theta)$. The weighted-correction estimating function yields three roots. $Q_n(\theta)$ fails, whereas $D_n(\theta)$ succeeds to identify the correct root. Plots in this graph are (a) profiled weighted-correction estimating function, (b) $Q_n(\theta)$, and (c) $D_n(\theta)$. For this sample data: N=200. True values* $(\alpha_0, \beta_0) = (0, 1)$. *Both the true covariate and the additive measurement error follow a standard normal distribution.*

5.5 Discussions

In this chapter, we constructed two objective functions: the corrected quasi-likelihood $Q_n(\theta)$ and the corrected empirical likelihood $D_n(\theta)$, to the conditional score. We stated that both objective functions have desirable large-sample properties: achieve consistent inference of θ as $n \to \infty$. However, we only gave a rigorous proof to $Q_n(\theta)$ (Section 5.6). We could possibly establish a theorem to $D_n(\theta)$, similar to Theorem 5.3.1 for $Q_n(\theta)$. Unfortunately, $D_n(\theta)$ is a much more complicated function in structure than $Q_n(\theta)$ since $D_n(\theta)$ involves a Lagrange multiplier λ that is a continuous differentiable function of θ.

A crucial step to show that, among all the roots of the conditional score as $n \to \infty$, the maximizer of $D_n(\theta)$ is consistent is to show that $D_n(\theta)$ converge uniformly in probability over the compact parameter space of θ and λ. Important sufficient conditions for $D_n(\theta)$ to converge uniformly Hong et al. (2003) are (a) $\Psi_{ws}(y_i, w_i, z_i, \theta)$ is uniformly continuous in θ; and (b) $\Psi_{ws}(y_i, w_i, z_i, \theta)$ is uniformly bounded. Once the uniform convergence in probability is true for $D_n(\theta)$, a rigorous proof of our conjecture on $D_n(\theta)$ can be done by applying similar contradiction technique used in the proof of $Q_n(\theta)$ (Section 5.6). The fact that $\Psi_{ws}(y_i, w_i, z_i, \theta)$ is continuous in θ and the parameter space is compact implies that (a) is true, by the Heine-Cantor theorem. However, (b) appears to be violated since the underlying normal measurement error is unbounded. Note that (b) is a very strong statement and typically may not be met. However, in practice, the observed surrogates are typically finite regardless the underlying distribution of measurement error. That is, $\Psi_{ws}(y_i, w_i, z_i, \theta)$ is typically finite, even though in theory we could not make that claim. Therefore, even though the proof might be intractable, we believe that $D_n(\theta)$ is an operational valid method in practice.

In finite samples, $Q_n(\theta)$ could be ill-behaved for arbitrarily large $|\beta_x|$. In addition, when the underlying weighted-correction estimating function fails to generate a good root, $Q_n(\theta)$ may not be able to provide adequate information to distinguish roots. On the other hand, $D_n(\theta)$, does not necessarily favor large $|\beta_x|$. Moreover, $D_n(\theta)$ can still generate information about the correct conditional score root even if the underlying weighted-correction estimating function fails to generate a good root. Therefore, $D_n(\theta)$ is expected to be superior to $Q_n(\theta)$ in finite samples, which is confirmed by the simulations. This is also the reason that we suggest using $D_n(\theta)$ in practice.

Our simulations showed that $D_n(\theta)$ has a satisfactory performance on identifying correct conditional score roots in finite samples with large measurement errors. The simulation results also indicated that the correct root is typically the one closest to the naive estimator. Even though $D_n(\theta)$ still remains an incomplete solution to the conditional score according to the simulations, it is a promising method to distinguish among conditional score roots and may have the potential to be improved.

The implementation of $D_n(\theta)$ relies on a crucial assumption that a correct root exists, which is also a requirement for the heuristic approach (Stefanski & Carroll 1987) and the conditional quasi-likelihood (Hanfelt & Liang 1997). However, such correct root might not exist. Indeed, the requirement of existence of correct roots is a drawback of objective function approaches. As suggested in Hanfelt & Liang (1997), a good practical approach is to reduce the magnitude of the measurement error by using replicated surrogates. By doing so, one not only increases the likelihood of generating a correct root but also reduces the possibility of observing multiple roots.

5.6 Proofs

We give a proof of Theorem 5.3.1.

First, the expectation notation E in this proof denotes the expectation with respect to θ_0.

The basic assumption is that $E[Q(\theta)] < \infty$ for all $\theta \in \Theta$. It is known that $E[Q(\theta)]$ has a unique maximum at the true value θ_0:

$$E[Q(\theta)] < E[Q(\theta_0)] \quad \text{for} \quad \theta \neq \theta_0. \tag{5.11}$$

Let $\hat{\theta}_{1n} \in \Theta$ and $\hat{\theta}_{1n} \xrightarrow{p} \theta_0$ be a consistent root of the conditional score and $\hat{\theta}_{2n} \in \Theta$ be the maximizer of $Q_n(\theta)$ among all the roots of the conditional score. We need to show that $\hat{\theta}_{2n} \xrightarrow{p} \theta_0$.

We have, by definition,

$$Q_n(\hat{\theta}_{1n}) \leq Q_n(\hat{\theta}_{2n}). \tag{5.12}$$

A crucial part of this proof is that $Q_n(\theta)$ converges uniformly in θ:

$$\sup_{\theta \in \Theta} |Q_n(\theta) - E[Q(\theta)]| \xrightarrow{a.s.} 0. \tag{5.13}$$

Three conditions sufficient for (5.13) to hold are given in Theorem 16(a) on page 108 (Ferguson 1996). In our case, those conditions are (a) Θ is compact; (b) $Q(y, w, z, \theta)$ is continuous in θ for all y, w, and z; (c) There exists a function $M(y, w, z) \geq |Q(y, w, z, \theta)|$ such that $E[M(Y, W, Z))] < \infty$ for all $\theta \in \Theta$, y, w, and z. In those conditions, (a) is assumed and (b) is also true. We now verify the condition (c).

$$\begin{aligned} Q(y,w,z,\theta) &= 2\left\{\frac{(y-1)\exp((\alpha+\beta_z^{\mathrm{T}}z+\beta_x^{\mathrm{T}}w)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)} - \frac{y\exp(-(\alpha+\beta_z^{\mathrm{T}}z+\beta_x^{\mathrm{T}}w)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)}\right\} \\ &= 2\left\{\frac{(y-1)\exp((\alpha+\beta_z^{\mathrm{T}}z+\beta_x^{\mathrm{T}}x)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)}\exp(\beta_x^{\mathrm{T}}u/2) - \frac{y\exp(-(\alpha+\beta_z^{\mathrm{T}}z+\beta_x^{\mathrm{T}}x)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)}\exp(-\beta_x^{\mathrm{T}}u/2)\right\} \end{aligned}$$

Since Θ is compact, $\theta = (\alpha, \beta_z, \beta_x)$ is bounded. Also by assumption, X and Z are bounded. Moreover, $\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8) \geq 1$. Therefore, both

$$\frac{(y-1)\exp((\alpha+\beta_z^{\mathrm{T}}z+\beta_x^{\mathrm{T}}x)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)} \quad \text{and} \quad \frac{y\exp(-(\alpha+\beta_z^{\mathrm{T}}z+\beta_x^{\mathrm{T}}x)/2)}{\exp(\beta_x^{\mathrm{T}}\Sigma_{uu}\beta_x/8)}$$

are bounded. We now show that $\exp(\beta_x^{\mathrm{T}}u/2)$ can be bounded by a function $M(u)$ with finite mean.

Let $a = (a_1, ..., a_p)^{\mathrm{T}}$ and $b = (b_1, ..., b_p)^{\mathrm{T}}$ such that $a_k \leq \beta_{x_k} \leq b_k < \infty$, where $k = 1, ..., p$. It is understood that $p < \infty$. The measurement error $U \sim N(0, \Sigma_{uu})$, where Σ_{uu} is finite.

Let

$$M(u) = \prod_{k=1}^{p} \left[\exp(a_k u_k/2) + \exp(b_k u_k/2)\right].$$

Then $\exp(\beta_x^{\mathrm{T}}u/2) \leq M(u)$.

$$E[M(U)] = E\prod_{k=1}^{p} \left[\exp(a_k u_k/2) + \exp(b_k u_k/2)\right],$$

which can be written as the sum of a finite number of terms with each term taking the form

$E[\exp(v^{\mathrm{T}}u)/2]$. Here, v is a vector of length p and v_k is either a_k or b_k, where $k = 1, ..., p$. Hence, $v_k < \infty$. Since each term of $E[M(U)]$: $E[\exp(v^{\mathrm{T}}u)/2] = \exp(v^{\mathrm{T}}\Sigma_{uu}v/8) < \infty$, $E[M(U)] < \infty$.

Similarly, once can show that $\exp(-\beta_x^{\mathrm{T}}u/2)$ can be bounded by a function of u with finite mean.

Hence, condition (c) is met and (5.13) holds.

By Mann-Wald theorem,

$$E[Q(\hat{\theta}_{1n})] \xrightarrow{p} E[Q(\theta_0)], \tag{5.14}$$

where $E[Q(\hat{\theta}_{1n})] = E[Q(\theta)] \mid_{\theta=\hat{\theta}_{1n}}$.

We have

$$\begin{aligned} |Q_n(\hat{\theta}_{1n}) - E[Q(\theta_0)]| &\leq |Q_n(\hat{\theta}_{1n}) - E[Q(\hat{\theta}_{1n})]| + |E[Q(\hat{\theta}_{1n})] - E[Q(\theta_0)]| \\ &\leq \sup_{\theta\in\Theta} |Q_n(\theta) - E[Q(\theta)]| + |E[Q(\hat{\theta}_{1n})] - E[Q(\theta_0)]|. \end{aligned}$$

So that, by (5.13) and (5.14),

$$Q_n(\hat{\theta}_{1n}) \xrightarrow{p} E[Q(\theta_0)]. \tag{5.15}$$

The uniform convergence (5.13) implies

$$Q_n(\theta_0) \xrightarrow{p} E[Q(\theta_0)]. \tag{5.16}$$

Therefore,

$$Q_n(\theta_0) - Q_n(\hat{\theta}_{1n}) = o_p(1). \tag{5.17}$$

By (5.12)

$$\begin{aligned} Q_n(\hat{\theta}_{2n}) &\geq Q_n(\hat{\theta}_{1n}) \\ &= Q_n(\theta_0) - o_p(1). \end{aligned} \tag{5.18}$$

For every $\varepsilon > 0$, (5.11) implies that

$$\sup_{\theta:\|\theta-\theta_0\|>\varepsilon} E[Q(\theta)] < E[Q(\theta_0)] \tag{5.19}$$

Therefore, by (5.13), (5.18)and (5.19),

$$\hat{\theta}_{2n} \xrightarrow{p} \theta_0,$$

according to Theorem 5.7 on page 45 (van der Vaart 1998) and subsequent comments on page 46, taking $M_n(\theta) = Q_n(\theta)$ and $M(\theta) = E[Q(\theta)]$.

Chapter 6

Summary and future work

6.1 Summary

This dissertation research focused on resolving the multiple-roots problem of the conditional score (Stefanski & Carroll 1987) for logistic regression with errors-in-covariates. Even though the multiple-roots problem has been known in the literature for a long time, limited research has been done to exam its seriousness in this particular problem of logistic regression with errors-in-covariates. Our finite-sample root behaviors study showed that this issue of multiple roots could be serious, especially when the measurement error is large. We also found that the conditional score may have a single but bad root (Figure 3.4) in finite samples when the measurement is large enough. This research mainly focused on the cases where the conditional score has multiple solutions.

In this dissertation research, we have proposed two methods to resolve the multiple-roots problem of the conditional score. Each approach has its own advantages and limitations.

The first approach is to combine the conditional score with an estimating function that does not yield inconsistent roots. We developed a weighted-correction estimating function for logistic regression with errors-in-covariates. This new estimating function, even though is typically less efficient than the asymptotically locally efficient conditional score, does not yield inconsistent roots. We proposed to combine the conditional score with the weighted-correction estimating function using empirical likelihood. By doing so, the inconsistent roots of the conditional score are eliminated in large samples since those two estimating functions only share consistent roots in large samples. The estimator of the combined estimation procedure is the proposed maximum empirical likelihood estimator. This proposed estimator is guaranteed to be asymptotically locally efficient. Simulation studies showed that the proposed combined estimation procedure works well in finite samples with large measurement errors. The results also showed that it outperforms existing consistent methods in many situations.

In summary, the first approach provides a new estimation procedure to resolve the multiple-roots problem of the conditional score for logistic regression with errors-in-covariates. The limitations of this approach are the following. First, the weighted-correction estimating function and the conditional score could have wrong roots that are close to each other in finite samples. As a result, the empirical likelihood based on these two functions could be maximized around their wrong roots in finite samples. Therefore, in finite samples, the correct empirical likelihood maximizers are usually local maximizers, not global maximizers. Based on our experience, starting from the naive estimators, we typically found good local maximizers around the truth. Second, by combining the asymptotically locally efficient conditional score with the typically less efficient weighted-correction estimating function, we may reduce the second-order efficiency that can offset the first-order efficiency gains. Indeed, our simulations indicate that the proposed estimator is usually less efficient than the conditional score estimator. Third, the proposed estimator has unsatisfied performance on approximately normality when sample size is small. i.e., it may have outliers and large skewness (Figure 4.3 and 4.4).

The second approach is to build an objective function to distinguish multiple roots of the conditional score. Stefanski & Carroll (1987) suggested choosing the root closest to the naive estimator in the presence of multiple solutions. However, this heuristic approach lacks theoretical support. Hanfelt & Liang (1997) developed the conditional quasi-likelihood to choose from roots of the conditional score. However, their objective function is based on a path-dependent line integral and therefore, is not unique in general. In this research, we have developed two objective functions to the conditional score. One is the corrected quasi-likelihood, which is based on the path-independent integral of the weighted-correction estimating function. The other one is the corrected empirical likelihood, which is based on the empirical likelihood ratio of the weighted-correction estimating function. The simulations show that the corrected empirical likelihood performs very well and works better than the corrected quasi-likelihood in finite samples and therefore, is regraded as our recommended objective function to the conditional score.

In summary, the second approach resolves the multiple-roots problem of the conditional score by building an objective function: the corrected empirical likelihood. The corrected empirical likelihood is a promising objective function to the conditional score. Its finite sample performance is satisfied according to the simulation results. Compared to the first approach, this approach does not suffer from the possible efficiency loss once it identifies the correct root. The limitations of this approach are the following. First, it requires the parameter space to be compact, which may limit its applications. Second, it requires that a good conditional score root exist to apply the corrected empirical likelihood. However, such good root might not be attainable if the measurement error is large enough (Figure 3.4). Third, it is time consuming to find all the possible roots of the conditional score. In addition, it may be impossible to check whether or not all the roots have been found. Last, according to the simulations, the corrected empirical likelihood did not identify the correct roots 100% of the time. In other words, the proposed corrected empirical likelihood does not outperform the heuristic procedure by Stefanski & Carroll (1987).

As summarized above, the two proposed approaches have their own advantages and disadvantages. We suggest using the first approach: The combined estimation procedure, in practice. The reasons are: First, when the measurement is large enough, the conditional score may not be able to generate a good root (Figure 3.4). However, the first approach still apples and is typically able to yield good solutions. On the other hand, the second approach is not applicable at this case. Second, the corrected empirical likelihood is not guaranteed to identify the correct roots according to our simulations.

6.2 Future work

The proposed combined estimation procedure yields locally efficient estimators only in large samples. Unfortunately, the weighted-correction estimating function may have wrong roots that are close to the wrong roots of the conditional score in finite samples. In future studies, we would like to explore the reason why this is the case, which might lead us to understand the relationship between those two estimating functions. Such investigations should help us to refine the combined estimation procedure to achieve better finite-sample properties.

We would also like to perform more simulation studies to the corrected quasi-likelihood and the corrected empirical likelihood in other situations. Such studies may help us to understand more about their finite-sample properties. By doing that, we may be able to find a way to combine the corrected quasi-likelihood and the corrected empirical likelihood to develop a new objective function that has better finite-sample performances.

Bibliography

B.Armstrong (1985), 'Measurement error in generalized linear models', *Communications in Statistics, Series B* **14**, 529–544.

Biernacki, C. (2005), 'Testing for a global maximum of the likelihood', *Journal of Computational & Graphical Statistics* **14**, 657–674.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models*, second edn, Chapman & Hall/CRC.

Chen, J., Sitter, R. R. & Wu, C. (2002), 'Using empirical likelihood method to obtain range restricted weights in regression estimators for surveys', *Biometrika* **89**, 230–237.

Chen, J., Variyath, A. M. & Abraham, B. (2008), 'Adjusted empirical likelihood and its properties', *Journal of Computational & Graphical Statistics* **17**, 426–443.

Cook, J. & Stefanski, L. A. (1995), 'A simulation extrapolation method for parametric measurement error models', *Journal of the American Statistical Association* **89**, 1314–1328.

Crowder, M. (1986), 'On consistency and inconsistency of estimating equations', *Econom. Theory* **2**, 305–330.

DiCiccio, T. J., Hall, P. & Romano, J. P. (1991), 'Empirical likelihood is bartlett-correctable', *The Annals of Statistics* **19**, 1053–1061.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall/CRC.

Finch, S. J., Mendell, N. R. & Jr., H. C. T. (1989), 'Probabilistic measures of adequacy of a numerical search for a global maximum', *Journal of the American Statistical Association* **84**, 1020–1023.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley & Sons.

Gan, L. & Jiang, J. (1999), 'A test for global maximum', *Journal of the American Statistical Association* **94**, 847–854.

Hall, P. & Scala, B. L. (1990), 'Methodology and algorithms of empirical likelihood', *International Statistical Review* **58**, 109–127.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S. & Merigan, T. C. (1996), 'Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for the mrfit. multiple risk factor intervention trial', *The New England Journal of Medicine* **335**, 1081–1090.

Hanfelt, J. J. & Liang, K.-Y. (1995), 'Approximate likelihood ratios for general estimating functions', *Biometrika* **82**, 461–477.

Hanfelt, J. J. & Liang, K.-Y. (1997), 'Approximate likelihoods for generalized linear errors-in-variables models', *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **59**, 627–637.

Hansen, L. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**, 1029–1054.

Heyde, C. C. & Morton, R. (1998), 'Multiple roots in general estimating equations', *Biometrika* **85**, 954–959.

Hong, H., Preston, B. & Shum, M. (2003), ‘Generalized empirical likelihood-based model selection criteria for moment condition models’, *Econometric Theory* **19**, 923–943.

Huang, Y. & Wang, C. (2000), ‘Cox regression with accurate covariates unascertainable : A nonparametric-correction approach’, *Journal of the American Statistical Association* **95**, 1209–1219.

Huang, Y. & Wang, C. (2001), ‘Consistent functional methods for logistic regression with errors in covariates’, *Journal of the American Statistical Association* **96**, 1469–1482.

Huzurbazar, V. S. (1948), ‘The likelihood equation, consistency and the maxima of the likelihood function’, *Ann. Eugenics* **14**, 185–200.

Kitamura, Y. (2006), ‘Empirical likelihood methods in econometrics: Theory and practice’, *Cowles Foundation Discussion Paper* (1569), 66.

Kunitomo, N. & Matsushita, Y. (2003), ‘On finite sample distributions of the empirical likelihood estimator and the gmm estimator’, *CIRJE Discussion Paper, University of Tokyo* (CIRJE-F-200), 26.

Lehmann, E. L. (1983), *Theory of Point Estimation*, Wiley, New York.

Li, B. (1993), ‘A deviance function for the quasi-likelihood method’, *Biometrika* **80**, 741–753.

Li, B. & McCullagh, P. (1994), ‘Potential functions and conservative estimating functions’, *The Annals of Statistics* **22**, 340–356.

Lindsay, B. (1982), ‘Conditional score functions: Some optimality results’, *Biometrika* **69**, 503–512.

Markatou, M., Basu, A. & Lindsay, B. G. (1998), ‘Weighted likelihood equations with bootstrap root search’, *Journal of the American Statistical Association* **93**, 740–750.

McCullagh, P. (1983), 'Quasi-likelihood functions', *The Annals of Statistics* **11**, 59–67.

McCullagh, P. (1991), Quasi-likelihood and estimating functions, *in* 'In Statistical Theory and Modelling: In Honour of Sir David Cox(D. V. Hinkley, N. Reid and E. J. Snell, eds.)', Chapman & Hall, London, pp. 265–286.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, second edn, Chapman & Hall/CRC.

Nakamura, T. (1990), 'Corrected score function for errors-in-variables models: Methodology and application to generalized linear models', *Biometrika* **77**, 127–137.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *The Computer Journal* **7**, 308–313.

Owen, A. B. (1988), 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika* **75**, 237–249.

Owen, A. B. (1990), 'Empirical likelihood ratio confidence regions', *The Annals of Statistics* **18**, 90–120.

Owen, A. B. (1991), 'Empirical likelihood for linear models', *The Annals of Statistics* **19**, 1725–1747.

Owen, A. B. (2001), *Empirical Likelihood*, Chapman & Hall/CRC.

Pan, W.-H., Tseng, W.-P., F.-J.You, Tai, Y. & Chou, J. (1990), 'Positive relationship between urinary sodium chloride and blood pressure in chinese health examinees and its association with calcium excretion', *J. Hypertens.* **8**, 873–878.

Prentice, R. (1982), 'Covariate measurement errors and parameter estimation in a failuer time regression models', *Biometrika* **69**, 331–342.

Qin, J. & Lawless, J. (1994), 'Empirical likelihood and general estimating equations', *The Annals of Statistics* **22**, 300–325.

Qu, A., Lee, J. & Lindsay, B. G. (2008), 'Model diagnostic tests for selecting informative correlation structure in correlated data', *Biometrika* **95**, 891–905.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, second edn, Wiley, New York.

Robbins, H. E. (1968), 'Estimating the total probability of the unobserved outcomes of an experiment', *Ann. Math. Statist.* **39**, 256–257.

Rosner, B., Spiegelman, D. & Willett, W. (1989), 'Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error', *Statistics in Medicine* **8**, 1051–1070.

Rosner, B., Willett, W. & Spiegelman, D. (1990), 'Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error', *American Journal of Epidemiology* **132**, 734–745.

Small, C. G. & Wang, J. (2003), *Numerical Methods for Nonlinear Estimating Equations*, Oxford University Press.

Small, C. G., Wang, J. & Yang, Z. (2000), 'Eliminating multiple root problems in estimation', *Statistical Science* **15**, 313–341.

Stefanski, L. A. (1989), 'Unbiased estimation of a nonlinear function a normal mean with application to measurement error models', *Communications in Statistics, Series A - Theory and Methods* **18**, 4335–4358.

Stefanski, L. A. & Carroll, R. J. (1987), 'Conditional scores and optimal scores for generalized linear measurement-error models', *Biometrika* **74**, 703–716.

Thode, H. C., Finch, S. J. & Mendell, N. R. (1987), 'Finding the mle in a two component normal mixture', *Proc. Statist. Comput. Sect. Am. Statist. Ass.* pp. 472–475.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.

Wedderburn, R. W. M. (1974), 'Quasi-likelihood functions, generalized linear models, and gausscnewton method', *Biometrika* **61**, 439–447.

Wilks, S. S. (1938), 'The large-sample distribution of the likelihood ratio for testing composite hypotheses', *Ann. Math. Statist.* **9**, 60–62.

Wood, A., Do, K.-A. & Broom, B. (1996), 'Sequential linearization of empirical likelihood constraints wwith application to u-statistics', *Journal of Computational & Graphical Statistics* **5**, 365–385.

Wu, C. (2004), 'Some algorithmic aspects of the empirical likelihood method in survey sampling', *Statistica Sinica* **14**, 1057–1067.

Wu, C. (2005), 'Algorithms and r codes for the pseudo empirical likelihood method in survey sampling', *Survey Methodology* **31**, 239–243.