2012

Efficient estimation in missing data and survey sampling problems

Sixia Chen
Iowa State University

Efficient estimation in missing data and survey sampling problems

by

Sixia Chen

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Jae Kwang Kim, Co-major Professor

Wayne A. Fuller, Co-major Professor

Zhengyuan Zhu

Nordman Dan

Cindy L. Yu

Iowa State University

Ames, Iowa

2012

DEDICATION

I would like to dedicate this thesis to my parents and my wife without whose support I would not have been able to complete this work. I would also like to thank my friends and family for their loving guidance and financial assistance during the writing of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Jae Kwang Kim for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I am particularly grateful to Dr. Wayne A. Fuller for his helpful comments and suggestions. I would also like to thank my committee members for their efforts and contributions to this work.

ABSTRACT

The thesis consists of four research papers. The first paper deals with general theory for empirical likelihood under the standard setup. Instead of maximizing the empirical likelihood function, a functional-form approach is proposed to generalize the theory of empirical likelihood and to achieve computational efficiency. The second paper deals with an empirical likelihood approach for missing data. The proposed method uses a partial likelihood for the respondents and theories are developed for both a parametric response model and a nonparametric response model. Also, the proposed method is extended to two-phase sampling where the first-phase sample is obtained by complex survey sampling. The third paper deals with empirical likelihood in the survey sampling setup. In the proposed method, called the population empirical likelihood method, the empirical likelihood function is defined for the finite population and the sampling design is incorporated into one of the constraints in the optimization problem. The proposed method is quite useful when combining information from several independent surveys. The fourth paper proposes a novel application of the capture-recapture experiment to estimate the propensity score for nonignorable nonresponse. The proposed method can be used to reduce the selection bias associated with voluntary sampling.

CHAPTER 1. GENERAL INTRODUCTION

Hartley and Rao (1968) introduced the empirical likelihood (EL) approach under the name of "scale load". Owen (1988,1990) brought the EL method to standard statistical problems. For a comprehensive overview of EL method, see Owen (2001). Chen and Hall (1993) extended the EL method to inference for quantiles. Qin and Lawless (1994) extended the EL method to inference for parameters defined by some general estimating equations. DiCiccio et al. (1991) and Chen and Cui (2006) used bartlett correction techniques to improve the convergence rate of empirical likelihood ratio. The application of EL method in time series has been considered by Kitamura (1997), Nordman et al. (2007) and others. Recently, Hjort, McKeague and Van Keilogom (2009), Chen, Peng and Qin (2009) and Tang and Leng (2010) showed that the EL method continues to work when data dimensionality is growing. Newey and Smith (2004) proposed generalized empirical likelihood (GEL) which extended the scope of the traditional EL method. In chapter 2, we propose a different extension by using the functional-form empirical likelihood (FEL) method. The basic idea is to generalize the form of the EL weight or of the objective function. We prove the first order equivalence between our proposed estimator and the traditional EL estimator. The proposed estimator has certain advantages in terms of computation and choice of weights.

Missing data happens frequently in observational studies. If the missing mechanism is completely missing at random (CMAR) in the sense of Rubin (1976), we can safely remove the missing part of the data. However, if the response mechanism is missing at random (MAR) or not missing at random (NMAR), we may not ignore the missing data in order to produce efficient and consistent estimates. There are two main approaches for inference with missing data: Imputation and Propensity Score Weighting. Wang and Rao (2002) and Wang and Chen (2009) considered combining EL and imputation methods for inference with data missing at random. Alternatively, Qin, Leung and Shao (2002) proposed the EL method to deal with nonignorable missing data by using the propensity score method. Qin and Zhang (2007) applied the EL method in missing response problems. Chen, Leung and Qin (2008) proposed constructing two different empirical likelihood method with data MAR. Most recently, Qin et al. (2009) provided the complete EL method for missing covariate problem. The literature is somewhat sparse for modeling the response mechanism nonparametrically. Cheng (1994) discussed some asymptotic properties of the mean estimator based on the kernel regression method under ignorable missing data. Recently, Kim and Yu (2011) extended the approach of Cheng (1994) to handle nonignorable nonresponse. Xue (2009) discussed an empirical likelihood method for linear models using the weights computed from a nonparametric model where the kernel regression method is used to estimate the response model. Da Silva and Opsomer (2009) considered another type of nonparametric response probability estimator using local polynomial regression. Hirano et al (2003) and Cattaneo (2010) discussed semiparametric efficiency of the nonparametric response propensity estimators in the context of estimating average treatment effect in econometrics. In chapter 3, we propose a response EL method which can be used to handle both survey sampling and missing data problems. Specifically, we propose estimating the propensity score nonparametrically in the EL method. By doing this, the semi-parametric lower bound can be achieved automatically.

The use of the EL method for a finite population parameter was first considered by Chen and Qin (1993), but their method is only applicable under simple random sampling (SRS). Chen and Sitter (1999) proposed pseduo empirical likelihood (PEL) which can be used to deal with complex survey data. Wu and Rao (2006) constructed a likelihood ratio-based confidence interval for the population mean by using PEL. For the most recent development of PEL, see Rao and Wu (2009). The likelihood ratio property is the most attractive property of the EL method. The corresponding confidence region has several advantages compared to the normal approximation (NA) confidence region. These include better coverage rate, shape respecting, and transformation invariance. However, the PEL ratio converges to a scaled chi-squred distribution instead of the standard chi-squred distribution. The scale factor needs to be estimated and it often depends on the complex sampling design. In addition, the PEL estimator is not equivalent to the design optimal estimator. To avoid those drawbacks, we propose using the population empirical likelihood (POEL) estimator in chapter 4. The POEL likelihood ratio converges to the standard chi-squred distribution; the proposed estimator is equivalent to the design optimal estimator and the POEL method can combine several sources of auxiliary information.

A voluntary sample is a self-selected sample whose first order inclusion probabilities are unknown. The most popular method for the inference for a voluntary sample is propensity score weighting. Rosenbaum and Rubin (1983) and Rosenbaum (1987) proposed using propensity scores to estimate treatment effects in observational studies. Duncan and Stasny (2001) used the propensity score method to control coverage bias in telephone surveys. Lee (2006) applied the propensity score method to a volunteer panel web survey. Lee and Valliant (2009) and Valliant and Dever (2011) considered the propensity score method for a web-based voluntary sample. All of these studies assumed an ignorable selection mechanism. However, we often confront the case where the selection mechanism does depend on the study variable itself. In chapter 5, we propose a novel two-phase approach for estimators with a voluntary sample. The proposed method can be extended to handle a non-nested two-phase voluntary sample. The auxiliary information can be incorporated via the generalized method of moment (GMM).

We organize the thesis as followings. In chapter 2, we present the new functional form empirical likelihood (EL) method; We proposed a unified theory of using the EL method in missing data problems in chapter 3; In chapter 4, we propose using the population empirical likelihood (POEL) method for inference with survey data; In chapter 5, a novel approach is proposed for inference in the voluntary sample problem. Future works are presented in chapter 6. Technical details are presented in the appendixes.

CHAPTER 2. SEMI-PARAMETRIC INFERENCE WITH A FUNCTIONAL-FORM EMPIRICAL LIKELIHOOD

A paper submitted to the *Journal of the Korean Statistical Society*

Sixia Chen and Jae Kwang Kim

Abstract

A functional-form empirical likelihood method is proposed as an alternative for the empirical likelihood method. The proposed method has the same asymptotic properties as the empirical likelihood method but has more flexibility in choosing the weight construction. Also, some computational efficiency can be gained. Because it enjoys the likelihood-based interpretation, the profile likelihood ratio test has a chi-square limiting distribution. Some computational details are also discussed, and results from limited simulation studies are presented.

Key Words: Exponential tilting, Generalized method of moments, Nonparametric maximum likelihood method, Profile likelihood ratio test.

2.1 Introduction

The empirical likelihood method, proposed by Owen (1988, 1990), provides a useful tool for obtaining nonparametric confidence regions for statistical functionals. Even though the empirical likelihood method is a nonparametric approach in the sense that it does not require a parametric model for the underlying distribution of the sample observation, the empirical likelihood method enjoys some of the desirable properties of the likelihood-based method. Using a nonparametric likelihood function, the empirical likelihood method can easily incorporate known constraints on parameters and also incorporate prior information on parameters. For example, Chen and Qin (1993) and Qin (2000) discuss combining information using the empirical likelihood. A comprehensive overview of the empirical likelihood method is provided by Owen (2001).

We consider an extension of the empirical likelihood method by providing a class of nonparametric estimators that have the same asymptotic properties as the empirical likelihood method. In particular, instead of assuming a nonparametric likelihood, we consider a generalization of the empirical likelihood that uses a functional-form likelihood function in the likelihood maximization. The class of functional-form likelihood function contains the empirical likelihood function as a special case. The functional-form likelihood approach provides several useful alternatives to the classical empirical likelihood method in the sense that some of the computational difficulty of the empirical likelihood method can be avoided, and more clear insights can be obtained from the empirical likelihood method.

Let $z_1, \cdots, z_n$ be n independent realizations of a vector-valued random variable Z with a distribution function $F(z)$ that is completely unspecified. In the empirical likelihood approach, we consider a class of distribution functions, $\mathcal{F}_1 \subset \mathcal{F}$, that have support on $\{z_1, \cdots, z_n\}$. Thus, the elements in $\mathcal{F}_1$ can be written as

$$F_w(x) = \sum_{i=1}^{n} w_i I(z_i \leq x)$$

with $\sum_{i=1}^{n} w_i = 1$ and $w_i > 0$, where $I(z_i \leq x)$ takes the value one if $z_i \leq x$ and takes the value zero otherwise. The parameter w_i is the amount of point mass that unit z_i represents in the population. We are interested in making an inference about θ_0 that is defined as a unique solution to $E\{U(Z;\theta)\} = 0$, where $U(Z;\theta)$ is an r-dimensional vector of some function $U(Z;\theta)$ known up to θ and the dimension of θ equals $p \leq r$. Hansen (1982) and Imbens (1997) considered this over-identified situation in the context of a generalized method of moments in econometrics.

In this setup, Qin and Lawless (1994) considered the empirical likelihood estimator of θ_0 that can be obtained by maximizing

$$\sum_{i=1}^{n} \ln(w_i) \tag{2.1}$$

subject to

$$\sum_{i=1}^{n} w_i \{1, U(z_i;\theta)\} = (1, 0)\,. \tag{2.2}$$

Note that (2.2) is equal to the condition $E\{U(Z;\theta)\} = 0$ for $F \in \mathcal{F}_1$. Using the Lagrange multiplier method, the empirical likelihood estimator can be obtained by maximizing

$$l_e(\theta) = \sum_{i=1}^{n} \ln\{w_i(\theta)\}\,, \tag{2.3}$$

where $w_i(\theta)$ is of the form

$$w_i(\theta) = \frac{1}{n}\frac{1}{1+\hat{\lambda}_\theta^T U(z_i;\theta)} \tag{2.4}$$

and $\hat{\lambda}_\theta$ satisfies the second equation of (2.2). Qin and Lawless (1994) showed that the empirical likelihood estimator satisfies

$$2\left\{l_e(\hat{\theta}) - l_e(\theta_0)\right\} \to^d \chi_p^2 \tag{2.5}$$

where $\hat{\theta}$ is the empirical likelihood estimator. The result (2.5) is often called the Wilk's theorem for empirical likelihood and is quite useful in obtaining confidence regions for θ_0.

The weight (2.4) used to compute the empirical likelihood estimator can be expressed as

$$w_i(\theta, \hat{\lambda}_\theta) = \frac{m\left\{\hat{\lambda}_\theta^T U(z_i;\theta)\right\}}{\sum_{j=1}^{n} m\left\{\hat{\lambda}_\theta^T U(z_j;\theta)\right\}}, \tag{2.6}$$

where $m(x) = 1/(1-x)$ and $\hat{\lambda}_\theta = \hat{\lambda}(\theta; z_1, \cdots, z_n)$ satisfies

$$\sum_{i=1}^{n} w_i(\theta, \hat{\lambda}_\theta) U(z_i;\theta) = 0. \tag{2.7}$$

The Lagrange multiplier $\hat{\lambda}_\theta = \hat{\lambda}(\theta; z_1, \cdots, z_n)$ is completely determined by (2.7). We assume that, for given θ, the solution $\hat{\lambda}_\theta$ to (2.7) is unique. The unique solution exists for any given θ if 0 is inside the convex hull of the points $U(z_1;\theta), \cdots, U(z_n;\theta)$.

We consider an extension of the empirical likelihood estimator by allowing $m(x)$ in (2.6) to be some smooth function other than $m(x) = 1/(1-x)$. The proposed estimator can be called the functional-form empirical likelihood (FEL) estimator because it uses a known function $m(x)$ in computing the weights in the FEL estimator. For example, the exponential tilting (ET) estimator considered in Kitamura and Stutzer (1997) and Schennach (2007) is the same form (2.6) with $m(x) = \exp(x)$. Imbens, Spady, and Johnson (1998) advocated using the ET estimator over the empirical likelihood (EL) estimator based on Monte Carlo investigation and analytic comparison using higher order asymptotic expansion. In this paper, we discuss some asymptotic properties for the FEL estimator. In particular, asymptotic normality and a version of Wilk's theorem for the FEL estimator are established. We found that the asymptotic results in Qin and Lawless (1994) are special cases of the general results in this paper. The results in this paper can also be used to make inferences for other types of FEL estimators, including the ET estimator.

The main results are presented in Section 2. Some extensions are introduced in Section 3 to illustrate possible theoretical results of the proposed FEL estimator. In Section 4, the underlying algorithm is discussed. Results from a limited simulation study are presented in Section 5 and concluding remarks are made in Section 6.

2.2 Main Results

Based on the functional form of the FEL weights in (2.6), we can define a functional-form empirical log-likelihood function

$$l(\theta) = l(\theta, \hat{\lambda}_\theta) = \sum_{i=1}^{n} \ln \omega_i(\theta, \hat{\lambda}_\theta) = \sum_{i=1}^{n} \ln \left\{ \frac{m_i(\theta, \hat{\lambda}_\theta)}{\sum_{i=1}^{n} m_i(\theta, \hat{\lambda}_\theta)} \right\} \tag{2.8}$$

where $m_i(\theta, \hat{\lambda}_\theta) = m\{\hat{\lambda}_\theta^T U(z_i;\theta)\}$ for some function $m(\cdot)$ and $\hat{\lambda}_\theta$ satisfies (2.7). The log-likelihood function in (2.8) is a parametric form in the sense that the likelihood function is known except for some unknown parameter (θ, λ). The computation for optimization using (2.8) is generally simpler than the computation using the nonparametric likelihood (2.1) since the parameter space is reduced from n to $p + r$. The parameter λ is used to facilitate the computation for constrained optimization. Furthermore, the log-likelihood function (2.8) does not directly use any distributional assumptions. Thus, the nature of the maximum likelihood estimator using (2.8) is still nonparametric in the sense that it is valid without assuming any distributional assumptions. The only assumption we use is $E\{U(Z;\theta_0)\} = 0$.

Let $\hat{\theta}$ be the solution that maximizes $l(\theta, \hat{\lambda}_\theta)$ in (2.8). Let $\hat{Q}_1(\theta, \lambda) = \sum_{i=1}^{n} \omega_i(\theta, \lambda) U(z_i;\theta)$ and $\hat{Q}_2(\theta, \lambda) \equiv n^{-1} dl(\theta, \hat{\lambda}_\theta)/d\theta$. The solution $\hat{\theta}$ and its corresponding λ-value, denoted by

$\hat{\lambda} = \hat{\lambda}(\hat{\theta})$, satisfies $\hat{Q}_1(\hat{\theta}, \hat{\lambda}) = 0$ and $\hat{Q}_2(\hat{\theta}, \hat{\lambda}) = 0$. The solution $\hat{\theta}$ is called the FEL estimator of θ_0. For simplicity of notation, let $\gamma = (\theta, \lambda)$ and $\hat{\gamma} = (\hat{\theta}, \hat{\lambda})$. Also, let $\hat{Q}(\gamma) = (\hat{Q}_1(\gamma), \hat{Q}_2(\gamma))$.

To discuss the asymptotic properties of the FEL estimator, we assume the following conditions:

(C1) The solution θ_0 to $E\{U(Z;\theta)\} = 0$ is unique.

(C2) In the weight function (2.6), the function $m(x)$ is always positive and has continuous second-order derivatives at $x = 0$ with $m(0) = m'(0) = 1$.

(C3) The partial derivative $\dot{U}(\theta) = \partial U(\theta)/\partial\theta$ is a continuous function of θ in the compact set $\mathcal{A}$ and $\theta_0 \in \mathcal{A}$ almost surely.

(C4) The random functions $\hat{Q}(\gamma)$ converge uniformly in probability to $Q(\gamma) = E\left\{\hat{Q}(\gamma)\right\}$ in the compact set $\mathcal{B}$ and $\gamma_0 \in \mathcal{B}$, where $\gamma_0 = (\theta_0, 0)$.

The following theorem provides the consistency of the FEL estimator.

Theorem 2.2.1 *Assume that conditions (C1)-(C4) hold. Assume that the solution* $(\hat{\theta}, \hat{\lambda})$ *to* $\hat{Q}_1(\theta, \lambda) = 0$ *and* $\hat{Q}_2(\theta, \lambda) = 0$ *is uniquely determined. Then, the solution* $(\hat{\theta}, \hat{\lambda})$ *satisfies*

$$p \lim_{n\to\infty} (\hat{\theta}, \hat{\lambda}) = (\theta_0, 0) \tag{2.9}$$

where θ_0 *is a unique solution to* $E\{U(Z;\theta)\} = 0$.

In the special case of the empirical likelihood method, Qin and Lawless (1994) also proved (2.9). The proof of Theorem 2.2.1, which is different from that of Qin and Lawless (1994), is presented in Section A of Appendix A.

Theorem 2.2.2 *In addition to the conditions of Theorem 2.2.1, assume that*

(C5) $\partial^2 U(z, \theta)/(\partial\theta\partial\theta^T)$ *is continuous at* θ *in the compact set* $\mathcal{A}$ *almost surely.*

(C6) $||U(Z;\theta)||^3$, $||\partial U(Z;\theta)/\partial\theta||$, *and* $||\partial^2 U(Z, \theta)/(\partial\theta\partial\theta^T)||$ *are bounded by some integrable function* $G(Z)$.

(C7) The $r \times p$ matrix $E\{\partial U(Z;\theta_0)/\partial\theta\}$ has full column rank p. Also, $Var\{U(Z;\theta)\}$ is positive definite in the compact set $\mathcal{A}$.

Then, we have

$$\sqrt{n}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\lambda}-0\end{pmatrix} \to^d N(0,\mathbf{V}) \tag{2.10}$$

where

$$\mathbf{V}=\begin{pmatrix}V_1 & 0\\ 0 & V_2\end{pmatrix}$$

where

$$V_1=\{E(\frac{\partial U}{\partial\theta})^T(EUU^T)^{-1}E(\frac{\partial U}{\partial\theta})\}^{-1}$$

and

$$V_2=\left\{E(UU^T)\right\}^{-1}\left\{I-E(\frac{\partial U}{\partial\theta})V_1E(\frac{\partial U}{\partial\theta})^T[E(UU^T)]^{-1}\right\}.$$

The proof of Theorem 2.2.2 is presented in Section B of Appendix A. Using Theorem 2.2.2, we can construct a Wald-type confidence interval for θ_0. The asymptotic variance V_1 of $\sqrt{n}(\hat{\theta}-\theta_0)$ can be consistently estimated by

$$\left[\left\{\sum_{i=1}^n w_i\dot{U}\left(z_i;\hat{\theta}\right)\right\}^T\left\{\sum_{i=1}^n w_iU\left(z_i;\hat{\theta}\right)U\left(z_i;\hat{\theta}\right)^T\right\}^{-1}\left\{\sum_{i=1}^n w_i\dot{U}\left(z_i;\hat{\theta}\right)\right\}\right]^{-1},$$

where $w_i=w_i(\hat{\theta},\hat{\lambda})$ is the final FEL weight in (2.6) evaluated at $\hat{\theta}$ and $\hat{\lambda}$.

By Theorem 2.2.2, asymptotic variance of the FEL estimator can be derived. For example, if $z_i=(x_i,y_i)^T$ and $\mu_x=E(x)$ is known, the FEL estimator of $\theta=E(y)$ can be obtained using $\hat{\theta}=\sum_{i=1}^n\hat{m}_iy_i/\sum_{i=1}^n\hat{m}_i$ with $\hat{m}_i=m\{\hat{\lambda}(x_i-\mu_x)\}$ where $\hat{\lambda}$ satisfies $\sum_{i=1}^n\hat{m}_i(x_i-\mu_x)=0$. The asymptotic variance of $\hat{\theta}$ is equal to $n^{-1}V(y)\left\{1-\rho^2\right\}$ where ρ is the correlation coefficient of x and y in the population. Note that the asymptotic variance is equal to the asymptotic variance of the regression estimator

$$\hat{\theta}_{reg}=\bar{y}+S_{yx}S_{xx}^{-1}\left(\mu_x-\bar{x}\right) \tag{2.11}$$

and so the FEL estimator in this setup is asymptotically equivalent to the regression estimator (2.11). The regression estimator (2.11) is the maximum likelihood estimator under the bivariate normality assumption (Anderson, 1957). The asymptotic variance V_1 is equal to the semiparametric lower bound discussed in Chamberlain (1987) and so the FEL estimator achieves semiparametric efficiency.

Theorem 2.2.3 *The functional-form empirical likelihood ratio statistic for testing* $H_0 : \theta = \theta_0$ *is*

$$W(\theta_0) = l(\hat{\theta}) - l(\theta_0) \tag{2.12}$$

where $l(\theta)$ *is given by (2.8). Under the assumption of Theorem 2.2.1,we have that*

$$2W(\theta_0) \rightarrow^d \chi^2_p \tag{2.13}$$

as $n \rightarrow \infty$, *when* H_0 *is true.*

Theorem 2.2.3, which can be called the Wilk's theorem for FEL method, shows that the FEL log-likelihood in (2.8) can be used to construct a confidence interval based on the likelihood ratio statistics (2.12) as in the parametric likelihood method. In the following corollary, we show that the FEL method can be used to construct a profile of likelihood ratio confidence intervals. The proofs of Theorem 2.2.3 and Corollary 2.2.1 are presented in Sections C and D of Appendix A, respectively. Results similar to Corollary 2.2.1 are also presented in Qin and Lawless (1994) in the context of empirical likelihood method, but we presents a different proof of the corollary.

Corollary 2.2.1 *Let* $\theta^T = (\theta_1, \theta_2)^T$, *where* θ_1 *and* θ_2 *are* $q \times 1$ *and* $(p-q) \times 1$ *vectors, respectively. For* $H_0 : \theta_1 = \theta_1^0$, *the profile generalized empirical likelihood ratio test statistic is defined by*

$$W_2 = l(\hat{\theta}_1, \hat{\theta}_2) - l(\theta_1^0, \hat{\theta}_2^0) \tag{2.14}$$

where $\hat{\theta}_2^0$ *maximizes* $l(\theta_1^0, \theta_2)$ *with respect to* θ_2. *Then, under* H_0, *we have that*

$$2W_2 \rightarrow^d \chi^2_q$$

as $n \rightarrow \infty$.

Remark 2.2.1 *The FEL method could be called a generalized empirical likelihood method because it is essentially a generalization of the empirical likelihood method using functional-form weight function. The term "generalized empirical likelihood", however, was already used by Smith (1997) and Newey and Smith (2004) to denote another type of extension to empirical likelihood method in econometrics using a saddle point optimization problem. Our method is different from the GEL method because we do not have to specify the objective function for saddle point computation and we have only to directly specify the functional-form for the weights in FEL estimators.*

2.3 Extension

The log-likelihood function in (2.8) can be viewed as a negative divergence function between $1/n$ and w_i. Instead of using a divergence function based on the log-likelihood (2.8), one can also consider a more general class of divergence functions. Specifically, we consider a class of divergence functions based on power-divergence statistics, proposed by Cressie and Read (1984),

$$CR(\alpha) = \frac{2}{\alpha(\alpha+1)} \sum_{i=1}^{n} \left\{ \left(\frac{1/n}{\omega_i} \right)^{\alpha} - 1 \right\}. \tag{2.15}$$

Note that $CR(0) = -2\sum_{i=1}^{n} log(n\omega_i)$, which is the log-likelihood function in (2.6) and $CR(-1) = 2\sum_{i=1}^{n} n\omega_i log(n\omega_i)$, which is often called the Kullback-Leibler divergence measure.

The results in Section 2 show that the choice of weight function is not critical because the resulting estimators are all asymptotically equivalent. Surprisingly, we show in this section that the choice of the objective function is not critical either. The results presented here are an extension of Baggerly (1998) to the case when θ is defined through the solution to an estimating equation.

Theorem 2.3.1 *Let $\hat{Q}_1(\theta,\lambda) = \sum_{i=1}^{n} \omega_i U(z_i;\theta)$ and $\hat{Q}_2(\theta,\lambda) = n^{-1} dl_3(\theta,\lambda)/d\theta$ where ω_i is defined in (2.6) and*

$$l_3\left(\theta,\lambda\right) = -\frac{1}{\alpha(\alpha+1)} \sum_{i=1}^{n} \left[\{\omega_i(\theta,\lambda)n\}^{-\alpha} - 1 \right]. \tag{2.16}$$

Suppose that $(\hat{\theta}, \hat{\lambda})$ *is the solution of* $\hat{Q}_1(\theta, \lambda) = 0$ *and* $\hat{Q}_2(\theta, \lambda) = 0$. *Then under conditions stated in theorem 2.2.1 and theorem 2.2.2, we have*

$$\sqrt{n} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\lambda} \end{pmatrix} \rightarrow^d N(0, V) \tag{2.17}$$

where V *is defined in (2.10). Also, the generalized empirical likelihood ratio statistic for testing* $H_0 : \theta = \theta_0$ *satisfies*

$$2 \left\{ l_3(\hat{\theta}) - l_3(\theta_0) \right\} \rightarrow^d \chi^2_p \tag{2.18}$$

where $l_3(\theta)$ *is given by (2.16).*

Theorem 2.3.1 is a general result in the sense that, for the special case of $\alpha = 0$ in (2.15), it leads to Theorem 2.2.2 and Theorem 2.2.3. Also, for the special case of $\alpha = -1$, we have the following result. Its proof is very similar to that of Theorem 2.3.1 and is not presented here.

Corollary 2.3.1 *Let* $l_2(\theta) = -\sum_{i=1}^n n\omega_i \log(n\omega_i)$ *and assume that* $\hat{\theta}$ *maximizes* $l_2(\theta)$. *Then we have*

$$2 \left\{ l_2(\hat{\theta}) - l_2(\theta_0) \right\} \rightarrow^d \chi^2_p,$$

and θ_0 *is the true value of* θ.

2.4 Computational Aspects

The FEL estimator that maximizes the objective function (2.8) subject to the constraint (2.7) could be viewed as a standard optimization problem in the (θ, λ) space of dimension $p + r$. However, as shown in Section A of Appendix A, the probability limit $Q_2(\theta, \lambda)$ of $\hat{Q}_2(\theta, \lambda)$ satisfies $Q_2(\theta, 0) = 0$ for all θ. Thus, standard approaches to solving the systems of equations $\hat{Q}_1(\theta, \lambda) = 0$ and $\hat{Q}_2(\theta, \lambda) = 0$ can have erratic behavior in the neighborhood of $\lambda = 0$.

To avoid this numerical problem, we consider an approach using a penalty term used in the ridge regression method, as was also considered by Imbens, Spady, and Johnson (1998). The objective function with a penalty term can be expressed as

$$l^*(\theta, \lambda) = l(\theta, \lambda) - 0.5K \cdot \hat{Q}_1(\theta, \lambda)^T W \hat{Q}_1(\theta, \lambda), \tag{2.19}$$

where $l(\theta, \lambda)$ is the original objective function, such as (2.8) or (2.16), and K is a scalar penalty term that makes the optimization problem locally convex, and W is some $r \times r$ positive definite matrix. Note that $\hat{Q}_2^*(\theta, \lambda) = n^{-1} \partial l^*(\theta, \lambda) / \partial \theta$ can be written

$$\hat{Q}_2^*(\theta, \lambda) = Q_2(\theta, \lambda) - K \cdot n^{-1} \dot{Q}_{1\theta}(\theta, \lambda)^T W \hat{Q}_1(\theta, \lambda),$$

where $\dot{Q}_{1\theta}(\theta, \lambda) = \partial \hat{Q}_1(\theta, \lambda)/\partial \theta$. Thus, for sufficiently large $K = O(n)$, we have

$$Q_2^*(\theta, 0) \neq 0 \quad \text{for } \theta \neq \theta_0 \quad \text{and } Q_2^*(\theta_0, 0) = 0, \tag{2.20}$$

where $Q_2^*(\theta, \lambda)$ is the probability limit of $\hat{Q}_2^*(\theta, \lambda)$. Property (2.20) follows because

$$Q_2^*(\theta, \lambda) = Q_2(\theta, \lambda) + C(\theta, \lambda) Q_1(\theta, \lambda)$$

for some matrix $C(\theta, \lambda)$, and $Q_1(\theta, \lambda)$ satisfies (2.20). Once the solution $\left(\hat{\theta}^*, \hat{\lambda}^*\right)$ that maximizes $l^*(\theta, \lambda)$ in (2.19) is obtained, we solve

$$\hat{Q}_1\left(\hat{\theta}^*, \lambda\right) = \sum_{i=1}^n m\left\{\lambda^T U(z_i; \hat{\theta}^*)\right\} U(z_i; \hat{\theta}^*) = 0 \tag{2.21}$$

for λ to get the final solution. The Newton-type solution to (2.21) can be computed by

$$\hat{\lambda}_{(t+1)} = \hat{\lambda}_{(t)} - \left\{\sum_{i=1}^n \dot{m}\left(\hat{\lambda}_{(t)}^T U_i^*\right) U_i^* U_i^{*T}\right\}^{-1} \left\{\sum_{i=1}^n m\left(\hat{\lambda}_{(t)}^T U_i^*\right) U_i^*\right\},$$

where $U_i^* = U(z_i; \hat{\theta}^*)$, with an initial value $\hat{\lambda}_{(0)} = 0$.

To demonstrate the computation, we use a sample of size $n = 50$ generated from a bivariate normal distribution

$$(X, Y) \sim^{iid} N\left[\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right]. \tag{2.22}$$

In the computation, we set $W = I$ and let K vary from 10 to 1000. We assume that $\mu_x = 1$ is known and we are interested in estimating μ_y. We used the exponential tilting weight of the form

$$\omega_i = \frac{\exp(\lambda_1 x_i + \lambda_2 y_i)}{\sum_{j=1}^n \exp(\lambda_1 x_j + \lambda_2 y_j)}$$

From the realized sample, the estimates of $(\mu_y, \lambda_1, \lambda_2)$ that maximize the penalized likelihood (2.19) are computed for each K using

$$\hat{Q}_1(\theta, \lambda) = \left(\sum_{i=1}^n \omega_i (x_i - 1), \sum_{i=1}^n \omega_i (y_i - \theta)\right).$$

< Figure 2.1 around here. >

Figure 2.1 presents the plot of the solution $(\hat{\mu}_y, \hat{\lambda}_1, \hat{\lambda}_2)$ against the value of the penalty parameter K. The estimates of μ_y and λ_1 converge as K gets larger, but the estimate of λ_2 does not converge even for large K. Because the computation in Figure 1 is based on a single realization of the sample, the resulting $\hat{\mu}_y$ is not necessarily equal to $\mu_y = 1$. The estimate for μ_y can be used for final computation but $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$ need to be updated using (2.21).

2.5 Simulation Study

To check the finite sample performance of the FEL estimators, we performed two limited simulation studies. In the first simulation study, we generated two sets of bivariate data (x_i, y_i) from two different sampling distributions: the bivariate normal distribution (2.22) and a bivariate non-normal distribution defined by

$$\begin{aligned} x_i &\sim \chi^2(1) \\ y_i &= \sqrt{M}(x_i - 1) + e_i, \end{aligned} \tag{2.23}$$

where $M = 0.5$, $e_i \sim \exp(1)$, and e_i is independent of x_i for $i = 1, 2, \ldots, n$. Note that, in both distributions, $E(X) = E(Y)$, $V(X) = V(Y)$, and $Corr(X, Y) = 0.5$. For each distribution, we generated $B = 2,000$ independent Monte Carlo samples of size n, where we used the three different sample sizes: $n = 20, 50$, and 100.

For each sample generated above, we computed three FEL estimators of $\mu_y = E(Y)$ under the following scenarios:

(Scenario 1) We have no extra information.

(Scenario 2) We use $\mu_x = 1$ as the constraint.

(Scenario 3) We use $\mu_x = \mu_y$ as the constraint.

(Scenario 4) We use $\mu_x = \mu_y$ and $\sigma_x = \sigma_y$ as the constraints.

In Scenario 1, we used the sample mean to estimate θ. In Scenarios 2-4, the FEL methods are used to incorporate the additional information. In Scenario 3, for example, the additional information can be incorporated by using the FEL weights

$$\omega_i = \frac{m\left\{\lambda_1(x_i - y_i) + \lambda_2(y_i - \theta)\right\}}{\sum_{j=1}^{n} m\left\{\lambda_1(x_j - y_j) + \lambda_2(y_i - \theta)\right\}}$$

where λ_1 and λ_2 are computed by (2.21) with

$$U(x_i, y_i; \theta) = (x_i - y_i, y_i - \theta)$$

and θ is determined by maximizing the given objective function.

For the choice of $m(\cdot)$ function in ω_i, we considered three different FEL estimators as below:

1. Empirical likelihood estimator (EL) using $m(x) = 1/(1-x)$ with the objective function (2.8).

2. Exponential tilting estimator (ET1) using $m(x) = \exp(x)$ with the objective function $l(\theta) = -\sum_{i=1}^{n} n\omega_i log(n\omega_i)$.

3. Exponential tilting estimator (ET2) with the objective function (2.8).

Monte Carlo mean and Monte Carlo variance of the FEL estimators are computed for each scenario based on the Monte Carlo sample of size $B = 2,000$. All of the FEL estimators are essentially unbiased, and the Monte Carlo means are not presented here. Table 2.1 presents the Monte Carlo estimates of the relative efficiency of the FEL estimators. The efficiency is computed by the ratio of the variance of the sample mean (under Scenario 1) to the variance of the corresponding FEL estimator. Under the normal distribution, the theoretical values of the standardized variance of the FEL estimators are all approximately equal to $1/(1-\rho^2) = 1.333$ for the three scenarios, which is consistent with the simulation results in Table 2.1. The simulation results in Table 2.1 show that all of the FEL estimators show similar efficiency for large sample size ($n = 100$) but the ET estimators are slightly more efficient than the EL estimator for small sample size ($n = 20, 50$).

In the second simulation study, we compared the statistical power of test statistics derived from the FEL methods. In this simulation study, we first generated 6 different samples from

$$(X, Y) \sim^{iid} N\left[\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

with 6 different values of ρ, varying from 0 to 0.5. In addition to the normal model, we also generated samples from the non-normal model (2.23) where M is chosen to make $\rho = (0, 0.1, 0.2, 0.3, 0.4, 0.5)$.

In the second study we considered the same three FEl estimators. We used $\theta = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ and $U(x, y; \theta)$ is a 5-dimensional vector of unbiased estimating function for θ. For each FEL method, the profile likelihood test is constructed by computing the full maximum likelihood estimator $\hat{\theta}$ and the profile maximum likelihood estimator $\hat{\theta}^0$ that is computed under the null hypothesis $H_0 : \rho = 0$. The profile likelihood test with level α rejects the null hypothesis $H_0 : \rho = 0$ if

$$2\left\{l\left(\hat{\theta}_1, \hat{\theta}_2\right) - l\left(0, \hat{\theta}_2^0\right)\right\} \geq \chi_1^2(1-\alpha)$$

where $\theta_1 = \rho$, $\theta_2 = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$ and $\chi_1^2(1-\alpha)$ is the $1-\alpha$ quantile of the chi-square distribution with 1 degrees of freedom. In additional to the FEL method, we also computed the normal-based Pearson test for comparison.

The Monte Carlo power of the level $\alpha = 0.05$ test statistic was computed by the relative frequency of rejecting the null hypothesis $H_0 : \rho = 0$. Table 2.2 presents the Monte Carlo power of the test statistics obtained from three FEL methods for each sample. For $\rho = 0$, the power is the size of the test and it converges to $\alpha = 0.05$ as n gets larger. In the normal sample, the power of the test based on ET method is higher than that for EL method when $n = 100$. The ET1 method shows smaller type-1 error than the ET2 method when the sample size is small. In the non-normal sample, the EL method seems to have better statistical powers than the ET methods. Overall, the three FEL methods show similar performances in most cases, which is consistent with our theory.

2.6 Conclusions

Empirical likelihood method is useful in incorporating the known constraints of parameters and also in combining information from different sources. The functional-form empirical likelihood method proposed in this paper provides a unified approach of handling such constraints without using distributional assumptions on the sample observation. FEL methods allow us to set a more flexible objective function as well as a flexible weight function. Thus, computational efficiency can be achieved by finding a simple weight function in the FEL method. For example, in the simulation study, the computing time for the ET method is much shorter than the computing time for the EL method.

The FEL method can be used to provide a likelihood ratio test with a chi-square limiting distribution. Also, a profile likelihood ratio test can be derived using the orthogonality of the log-likelihood functions. To improve the coverage properties of the FEL in the small sample sizes, some cutting-edge techniques such as bootstrap calibration (Hall and Horowitz, 1996) or the Bartlett correction (Chen and Cui, 2006) can be used. Further investigation in this direction, including the Higher order expansion as in Liu and Chen (2010), is not discussed here and will be a topic of future research.

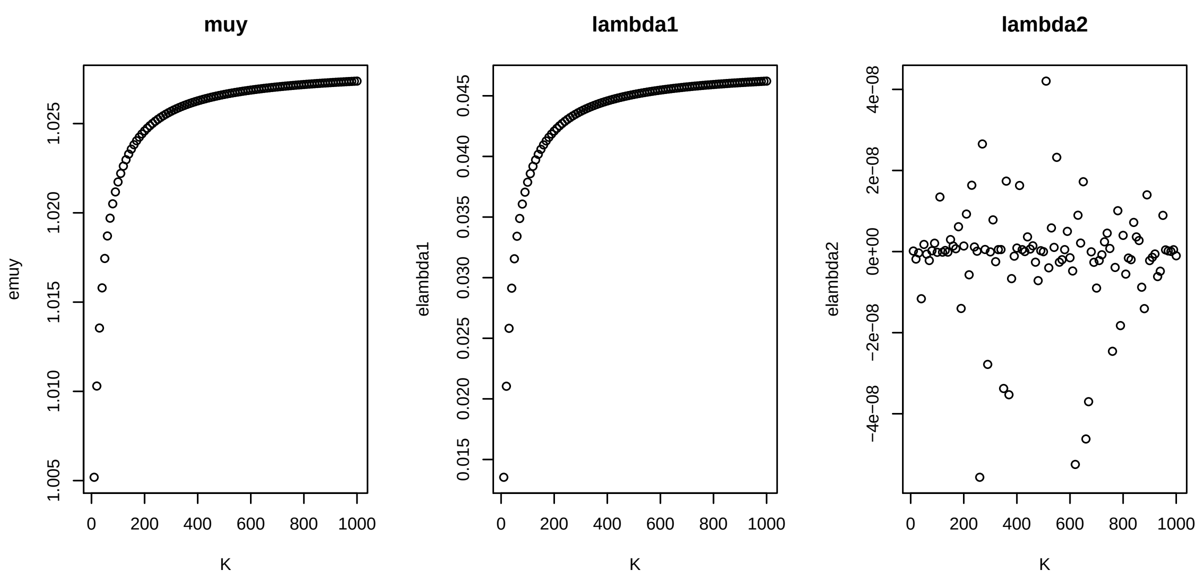

Figure 2.1 Parameter estimations versus penalty parameter.

Table 2.1 Monte Carlo relative efficiency of the point estimators.

Model	Situation	Sample size(n)	EL	ET1	ET2
Normal	S1	$n = 20$	1.0000	1.0000	1.0000
		$n = 50$	1.0000	1.0000	1.0000
		$n = 100$	1.0000	1.0000	1.0000
	S2	$n = 20$	1.2192	1.2496	1.2375
		$n = 50$	1.2729	1.2782	1.2772
		$n = 100$	1.3184	1.3146	1.3153
	S3	$n = 20$	1.2765	1.3377	1.3267
		$n = 50$	1.3183	1.3364	1.3337
		$n = 100$	1.3295	1.3339	1.3337
	S4	$n = 20$	1.1478	1.2244	1.2262
		$n = 50$	1.2558	1.2721	1.2754
		$n = 100$	1.3022	1.3042	1.3084
Non-normal	S1	$n = 20$	0.9960	1.0000	1.0000
		$n = 50$	0.9988	1.0000	1.0000
		$n = 100$	1.0000	1.0000	1.0000
	S2	$n = 20$	1.5547	1.7117	1.6455
		$n = 50$	1.5597	1.9005	1.8908
		$n = 100$	1.5233	1.9040	1.8990
	S3	$n = 20$	1.0676	1.1901	1.1632
		$n = 50$	1.0875	1.1592	1.1388
		$n = 100$	1.1518	1.2014	1.1839
	S4	$n = 20$	1.2721	1.3700	1.3067
		$n = 50$	1.2966	1.3691	1.3579
		$n = 100$	1.4289	1.5062	1.5188

Table 2.2 Power comparisons for testing $H_0 : \rho = 0$

Model	Method	Sample size	ρ					
			0	0.1	0.2	0.3	0.4	0.5
Normal	Pearson	$n = 20$	0.044	0.060	0.141	0.244	0.426	0.647
		$n = 50$	0.043	0.107	0.279	0.565	0.837	0.963
		$n = 100$	0.052	0.172	0.521	0.848	0.988	1.000
	EL	$n = 20$	0.100	0.126	0.235	0.358	0.545	0.738
		$n = 50$	0.062	0.141	0.323	0.610	0.863	0.970
		$n = 100$	0.059	0.192	0.529	0.856	0.985	1.000
	ET1	$n = 20$	0.096	0.121	0.229	0.344	0.520	0.721
		$n = 50$	0.061	0.138	0.323	0.608	0.859	0.971
		$n = 100$	0.060	0.189	0.534	0.859	0.986	1.000
	ET2	$n = 20$	0.117	0.140	0.254	0.379	0.570	0.756
		$n = 50$	0.064	0.147	0.335	0.619	0.866	0.971
		$n = 100$	0.059	0.196	0.536	0.861	0.987	1.000
Non -normal	Pearson	$n = 20$	0.048	0.075	0.165	0.285	0.403	0.610
		$n = 50$	0.041	0.113	0.277	0.536	0.769	0.917
		$n = 100$	0.039	0.174	0.492	0.848	0.965	0.996
	EL	$n = 20$	0.149	0.152	0.264	0.408	0.545	0.710
		$n = 50$	0.099	0.151	0.380	0.651	0.843	0.946
		$n = 100$	0.075	0.212	0.619	0.913	0.980	0.996
	ET1	$n = 20$	0.127	0.124	0.227	0.353	0.496	0.678
		$n = 50$	0.089	0.123	0.337	0.614	0.823	0.937
		$n = 100$	0.079	0.193	0.585	0.900	0.979	0.998
	ET2	$n = 20$	0.161	0.165	0.282	0.430	0.558	0.733
		$n = 50$	0.115	0.169	0.397	0.658	0.851	0.950
		$n = 100$	0.092	0.234	0.633	0.918	0.983	0.998

CHAPTER 3. A UNIFIED THEORY ON EMPIRICAL LIKELIHOOD METHODS WITH MISSING DATA AND SURVEY SAMPLING

A paper submitted to the *Australian and New Zealand Journal of Statistics(revision invited)*

Sixia Chen and Jae Kwang Kim

Abstract

Efficient estimation with missing data is an important practical problem with many application areas. Survey sampling can be treated as a missing data problem where the sample is treated as a realization of a known response mechanism. Parameter estimation under nonresponse is considered when the parameter is defined as a solution to an estimating equation. Using a response probability model, a complete-response empirical likelihood method can be constructed and the nonparametric maximum likelihood estimator can be obtained by solving the weighted estimating equation where the weights are computed by maximizing the complete-response empirical likelihood subject to the constraints that incorporate the auxiliary information obtained from the full sample. Often the constraints are constructed from the working outcome regression model for the conditional distribution of the estimating function given the observation. The proposed method achieves the semi-parametric lower bound when we correctly specify the conditional expectation of the estimating function, regardless of whether the response probability is known or estimated. When the response probability is estimated nonparametrically, the resulting empirical likelihood method automatically achieves the semi-parametric lower bound without specifying the conditional distribution of the estimating function. The proposed method is also applicable to two-phase sampling. Asymptotic theories

are derived and simulation studies are also presented.

Key Words: Missing at random; Nonparametric estimation; Response mechanism; Propensity score.

3.1 Introduction

The empirical likelihood (EL) method, proposed by Owen (1988, 1990), has become a very powerful tool for nonparametric inference in statistics. It uses a likelihood-based approach without having to make a parametric distributional assumption about the data observation. Thus, the EL method often leads to efficient estimation and enables likelihood-ratio type inference. Qin and Lawless (1994) considered the situation when the parameter of interest is the solution to a system of estimating equations. Owen (2001) provides a comprehensive overview of the EL method.

Under existence of missing data or survey data, however, the EL method is not directly applicable and some adjustment needs to be made. Qin (1993) addressed this problem using a biased sampling argument of Vardi (1985). Wang and Rao (2002) used regression-type imputation approaches to empirical likelihood inference. Wang and Chen (2009) used a nonparametric regression imputation approach to handle missing data in the empirical likelihood inference. The imputation approach uses some assumptions about the missing data given the observed data and usually assumes that the response mechanism is ignorable in the sense of Rubin (1976). Under an ignorable missing mechanism, the explicit modeling of the response model is avoided. In the case of survey sampling, Chen and Sitter (1999) considered the pseudo empirical likelihood estimator that uses the sampling weight in the empirical log-likelihood function. Kim (2009) considered an alternative empirical likelihood function based on the biased sampling likelihood of Vardi (1985) and Qin (1993). Wu and Rao (2006) discussed interval estimation using the pseudo empirical likelihood. Note that the survey sampling can be treated as a special case of missing data problem, where the sample is obtained by a planned missing mechanism and the first-order sample inclusion probability corresponds to the response probability in the usual missing data problem. The main difference is that the sample inclusion probabilities are known in survey sampling, as the missing mechanism is planned by the sampling design.

In this paper, we consider an alternative approach to handling missing data using a model for response probability. Use of parametric response probability model in the empirical likelihood inference has been considered in Qin and Zhang (2007) and in Chen et al. (2008). Qin et al. (2009) and Tan (2011) considered using EL to model the complete likelihood, where the nonparametric likelihood function is computed for the whole sample including the units with missing data. The use of complete likelihood attains the full efficiency and also provides a nice theory of the limiting chi-square distribution in the likelihood ratio test statistics. However, in some practical case, the unit-level information for the complete likelihood is not always available and the complete likelihood cannot be computed. For example, in survey sampling, the individual values of auxiliary variable in the non-sampled part are not usually available. In this case, the approach of using the complete likelihood for the finite population may not be applicable.

If the response mechanism is nonparametrically modeled, the literature is somewhat sparse. Cheng (1994) discussed some asymptotic properties of the mean estimator using the kernel regression method to estimate the conditional outcome regression model under an ignorable missing case. Recently, Kim and Yu (2011) extended the approach of Cheng (1994) to handle nonignorable nonresponse. Xue (2009) discussed an empirical likelihood method for linear models using the weights computed from a nonparametric model where the kernel regression method is used to estimate the response model. Da Silva and Opsomer (2009) considered another type of nonparametric response probability estimation using local polynomial regression. Hirano et al (2003) and Cattaneo (2010) discussed semiparametric efficiency of the nonparametric response propensity estimators in the context of estimating average treatment effect in econometrics.

In this paper, we propose a unified approach of the EL method with missing data that avoids using the complete likelihood. Under the setup of estimating function in Qin and Lawless (1994), the proposed method can handle the situation regardless of whether the response probabilities are known or estimated, parametrically or even nonparametrically. When the response probabilities are known, the proposed method can be applied to survey weighting problems when the first-order inclusion probabilities are known. Incorporating the population level auxiliary information into the weights in the sample is an important problem in survey sampling and is often called calibration weighting. Calibration weighting is considered in Deville and Särndal (1992), Fuller (2002), and Kim and Park (2010), among others. The proposed method can be directly applicable to the calibration weighting problem.

When the response probabilities are estimated from a parametric model, the proposed method under ignorable response mechanism is similar to the method of Qin and Zhang (2007). The proposed method is directly applicable to the problem of the propensity score weighting method. The propensity score weighting method can be found, for example, in Durrant and Skinner (2006), Kim and Kim (2007), and Chang and Kott (2008). We show that employing EL method using a suitable choice of control variable leads to efficient estimation in the sense that it achieves the lower bound of the asymptotic variance. Optimal choice of the control variable requires correct specification of the conditional distribution of the missing data given the observation. Under the nonparametric propensity score method, which will be discussed in Section 5, the lower bound of the asymptotic variance can be achieved without correctly specifying the conditional distribution.

In Section 2, we first review the existing methods of empirical likelihood under missing data and discuss a unified approach of the EL method. Asymptotic properties of the proposed estimator under known response probabilities are discussed in Section 3. The proposed EL estimator is discussed under estimated response probability in Section 4. Use of the nonparametric response model for the EL approach is discussed in Section 5. The proposed method is extended to two-phase sampling in Section 6. Results from two simulation studies are reported in Section 7.

3.2 Basic setup

Consider a multivariate random variable (X, Y) with distribution function $F(x, y)$ which is completely unspecified except that $E\{U(X, Y; \theta_0)\} = 0$ for some θ_0. We are interested in estimating the parameter θ_0 from a random sample of the distribution. To avoid unnecessary details, we assume that the solution to $E\{U(X, Y; \theta)\} = 0$ is unique. For simplicity, we assume that the dimension of U is equal to the dimension of θ.

If $(x_i, y_i),\ i = 1, 2, \ldots, n$, are n independent realizations of the random variable (X, Y), a consistent estimator of θ_0 can be obtained by solving

$$\sum_{i=1}^{n} U(x_i, y_i; \theta) = 0. \tag{3.1}$$

In this paper, we consider the problem of estimating θ_0 when x is always observed and y is subject to missingness. Let $r_i = 1$ if y_i is observed and $r_i = 0$ otherwise. We consider an approach based on the empirical likelihood (EL) method. To explain the idea, first note that the joint density of the observed data can be written as

$$p^{n_r}(1-p)^{n-n_r} \times \prod_{r_i=1} f(x_i, y_i | r_i = 1) \prod_{r_i=0} f(x_i | r_i = 0), \tag{3.2}$$

where n_r is the response sample size, $p = Pr(r = 1)$, $f(x, y|r)$ is the conditional density of (X, Y) given r, and $f(x_i|r_i = 0) = \int f(x_i, y_i|r_i = 0)dy_i$ is the marginal density of X among $r = 0$.

In the empirical likelihood approach, the distribution is assumed to have the support on the sample observation. Let $F_1(x, y) = Pr(X \leq x, Y \leq y|r = 1)$ and $F_0(x, y) = Pr(X \leq x, Y \leq y|r = 0)$. Under the empirical likelihood approach, we can express

$$F_1(x, y) = \sum_{r_i=1} \omega_i I(x_i \leq x, y_i \leq y), \tag{3.3}$$

where $\sum_{r_i=1} \omega_i = 1$, ω_i is the point mass assigned to (x_i, y_i) in the nonparametric distribution of $F_1(x, y)$, and $I(B)$ is an indicator function for event B. To express $F_0(x, y)$ using ω_i, note that we can write

$$f(x_i, y_i|r_i = 0) = f(x_i, y_i|r_i = 1) \times \frac{Odd(x_i, y_i)}{E\{Odd(x_i, y_i)|r_i = 1\}},$$

where

$$Odd(x, y) = \frac{Pr(r = 0 \mid x, y)}{Pr(r = 1 \mid x, y)}.$$

Thus, we can express $F_0(x, y) = Pr(X \leq x, Y \leq y|r = 0)$ by

$$F_0(x, y) = \frac{\sum_{r_i=1} \omega_i O_i I(x_i \leq x, y_i \leq y)}{\sum_{r_i=1} \omega_i O_i}, \tag{3.4}$$

where $O_i = Odd(x_i, y_i)$. Note that $F_0(x,y)$ is completely determined by two factors: ω_i and O_i. The factor ω_i is determined by the distribution $F_1(x,y)$ and the factor O_i is determined by the response mechanism. If $Odd(x,y)$ is a known function of (x,y), then we have only to determine ω_i.

From (3.4), the joint distribution of (x,y) can be written as

$$\begin{aligned} F_w(x,y) &= p \times \sum_{r_i=1} \omega_i I(x_i \le x, y_i \le y) + (1-p) \times \left\{ \frac{\sum_{r_i=1} \omega_i O_i I(x_i \le x, y_i \le y)}{\sum_{r_i=1} \omega_i O_i} \right\} \\ &= p \times \left\{ \sum_{r_i=1} \omega_i I(x_i \le x, y_i \le y) + (1/p - 1) \frac{\sum_{r_i=1} \omega_i O_i I(x_i \le x, y_i \le y)}{\sum_{r_i=1} \omega_i O_i} \right\}. \end{aligned}$$

Note that (3.3) implies

$$\begin{aligned} \sum_{r_i=1} \omega_i (O_i + 1) &= E\left\{ \frac{1}{\pi(X,Y)} | r = 1 \right\} \\ &= \int \frac{1}{\pi(x,y)} f(x,y|r=1) dxdy \\ &= \int \frac{1}{\pi(x,y)} \frac{\pi(x,y) f(x,y)}{p} dxdy = 1/p. \end{aligned}$$

Thus, we have $\sum_{r_i=1} \omega_i O_i = 1/p - 1$ and

$$F_w(x,y) = \frac{\sum_{r_i=1} \omega_i (1+O_i) I(x_i \le x, y_i \le y)}{\sum_{r_i=1} \omega_i (O_i + 1)}.$$

We propose maximizing the partial likelihood $\prod_{r_i=1} f(x_i, y_i | r_i = 1)$ in (3.2) in constructing the empirical likelihood. Thus, the proposed empirical likelihood approach can be formulated as maximizing

$$l_e(\theta) = \sum_{r_i=1} \log(\omega_i), \tag{3.5}$$

subject to

$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i (1+O_i) U(x_i, y_i; \theta) = 0. \tag{3.6}$$

Note that, in constraint (3.6), the observed values of x_i with $r_i = 0$ are not used. To incorporate the partial information, we can impose

$$\frac{\sum_{r_i=1} \omega_i (1+O_i) h(x_i;\theta)}{\sum_{r_i=1} \omega_i (1+O_i)} = n^{-1} \sum_{i=1}^{n} h(x_i;\theta). \tag{3.7}$$

as an additional constraint for some $h(x_i;\theta)$. The choice of $h(x;\theta)$ will be discussed later.

There are several other approaches using the empirical likelihood with missing data. Qin et al. (2002) considered using empirical likelihood for nonignorable nonresponse. Wang and Rao (2002) proposed empirical likelihood-based inference under imputation for missing response data. Qin and Zhang (2007) proposed an empirical likelihood method for estimating the mean response under ignorable missing data where the response probability $\pi_i = Pr(r_i = 1|X_i)$ is parametrically modeled by $\pi_i = \pi_i(\phi_0)$ for some ϕ_0. Specifically, they proposed maximizing

$$l = \sum_{r_i=1} \log \left\{ \pi_i(\hat{\phi}) p_i / \hat{\nu} \right\},$$

subject to

$$\sum_{r_i=1} p_i = 1, \quad \sum_{r_i=1} p_i \pi_i(\hat{\phi}) = \hat{\nu}, \quad \sum_{r_i=1} p_i h(x_i) = n^{-1} \sum_{i=1}^{n} h(x_i), \tag{3.8}$$

where $\hat{\phi}$ is the maximum likelihood estimator of ϕ_0 in the response probability, $h(x_i)$ is an arbitrary variable and $\hat{\nu} = n^{-1} \sum_{i=1}^{n} \pi_i(\hat{\phi})$. Once the estimated probability $\hat{p}_i$ is computed by the above maximization procedure, the population mean can be estimated by $\hat{\theta} = \sum_{r_i=1} \hat{p}_i y_i$. Chen et al. (2008) built two empirical likelihoods for response and non-response variables separately and formulated two estimating equations based on these two empirical likelihoods. In the context of the current setup, their proposed method can be described as maximizing $l = \sum_{r_i=1} \log(p_i) + \sum_{r_j=0} \log(q_j)$, subject to $\sum_{r_i=1} p_i = 1, p_i \geq 0, \sum_{r_j=0} q_j = 1, q_j \geq 0$, and

$$\sum_{r_i=1} p_i \frac{h(x_i;\theta) - \mu}{\pi_i(\hat{\phi})} = 0, \quad \sum_{r_j=0} q_j \frac{h(x_j;\theta) - \mu}{1 - \pi_j(\hat{\phi})} = 0, \tag{3.9}$$

where $\hat{\phi}$ is the maximum likelihood estimator. Qin et al. (2009) considered maximizing the complete likelihood $l_c = \sum_{i=1}^{n} \log(\omega_i)$ subject to

$$\sum_{i=1}^{n} \omega_i = 1, \quad \sum_{i=1}^{n} \omega_i \frac{r_i}{\pi_i} U_i(\theta) = 0, \tag{3.10}$$

and

$$\sum_{i=1}^{n} \omega_i (\frac{r_i}{\pi_i} - 1) h_i(\theta) = 0, \quad \sum_{i=1}^{n} \omega_i \left\{ r_i - \pi_i(\phi) \right\} \frac{\partial \pi_i(\phi)/\partial \phi}{\pi_i(\phi)\{1 - \pi_i(\phi)\}} = 0. \tag{3.11}$$

The computation requires that the individual values of x_i for $r_i = 0$ be available, which is not always possible, as discussed in Section 1. For example, in survey sampling problem, we only observe (x_i, y_i) for $r_i = 1$ and the aggregate information $\bar{x}_n = n^{-1} \sum_{i=1}^{n} x_i$ is available. In this case, the method of Qin et al. (2009) is not applicable.

In Section 3, some asymptotic properties of the proposed EL estimator described in (3.5)-(3.7) are developed for the case when $\pi_i = Pr(r_i = 1 \mid x_i, y_i)$ is a known function of (x_i, y_i). In particular, we show that the optimal choice of $h(x_i;\theta)$ that minimizes the asymptotic variance of the resulting EL estimator of θ is

$$h^*(x_i;\theta) = \tilde{U}(x_i;\theta) \equiv E\left\{U(x_i, y_i;\theta) \mid x_i\right\}.$$

In Section 4, we consider the case where $\pi_i = Pr(r_i = 1 \mid x_i, y_i)$ is a parametric model of the form $Pr(r = 1 \mid x, y) = \pi(x;\phi_0)$ for some ϕ_0. By plugging estimator $\hat{\phi}$ of ϕ_0 into the empirical likelihood procedure, we can find the empirical likelihood estimator. The asymptotical properties of this estimator are discussed in Section 4. If a parametric form of π is unknown, we can use a nonparametric model for π. Asymptotical properties of the EL estimator using a nonparametric estimator of π are discussed in Section 5.

3.3 Estimation with known response probability

In this section, we assume that the true response probability $\pi = Pr(r = 1|X, Y)$ is known, which is often the case with survey sampling where π_i denotes the first-order inclusion probability and the response indicator, r, represents the sampling indicator. The regularity conditions of this section can be found in the section A of Appendix B. Our proposed estimator introduced in Section 2, (3.5)-(3.7), can be described as maximizing

$$l = \sum_{r_i=1} \log(\omega_i), \tag{3.12}$$

subject to

$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i \pi_i^{-1} \left\{ h_i(\theta) - n^{-1} \sum_{i=1}^{n} h_i(\theta) \right\} = 0, \quad \sum_{r_i=1} \omega_i \pi_i^{-1} U_i(\theta) = 0. \tag{3.13}$$

For $\theta = E(Y)$, a popular choice of $h(\theta)$ is $h(\theta) = x$. In this case, the EL estimator of θ is obtained by $\hat{\theta}_{h1} = \sum_{r_i=1} w_i^* \pi_i^{-1} y_i / \sum_{r_i=1} w_i^* \pi_i^{-1}$ where $w_i^* = n_r^{-1}\{1 + \hat{\lambda}\pi_i^{-1}(x_i - \bar{x}_n)\}^{-1}$, $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$ and $\hat{\lambda}$ is constructed to satisfy $\sum_{r_i=1} w_i^* \pi_i^{-1}(x_i - \bar{x}_n) = 0$.

The following theorem presents some asymptotic properties of the EL estimator $\hat{\theta}_{h1}$. The proof is presented in Section B of Appendix B.

Theorem 3.3.1 *Let $\hat{\theta}_{h1}$ be the solution to the maximization above. Then, under the regularity conditions (C1)-(C5) in section A of Appendix B, we have*

$$\hat{\theta}_{h1} - \theta_0 = \tau \frac{1}{n} \sum_{i=1}^{n} \{ \frac{r_i}{\pi_i} U_i(\theta_0) - B(\frac{r_i}{\pi_i} - 1)\tilde{h}_i(\theta_0) \} + o_p(n^{-1/2}), \tag{3.14}$$

where $\tilde{h}_i(\theta_0) = h_i(\theta_0) - \mu_h$, $\mu_h = E(h)$, $B = E(U\tilde{h}'/\pi) \left\{ E(\tilde{h}\tilde{h}'/\pi) \right\}^{-1}$ and $\tau = -\{E(\partial U/\partial\theta)\}^{-1}$ evaluated at $\theta = \theta_0$. Hence, we have

$$\sqrt{n} \left(\hat{\theta}_{h1} - \theta_0 \right) \rightarrow^d N(0, V_{h1}), \tag{3.15}$$

where $\rightarrow^d$ denotes convergence in distribution, $V_{h1} = \tau \Omega_{h1} \tau'$ and

$$\Omega_{h1} = V \left\{ \frac{r}{\pi} \left(U - B\tilde{h} \right) + B\tilde{h} \right\} = E \left\{ (\frac{1}{\pi} - 1)(U - B\tilde{h})^{\otimes 2} \right\} + V(U), \tag{3.16}$$

and $A^{\otimes 2} = AA'$.

Because $\Omega_{h1} = E\{(\pi^{-1} - 1)(U - B\tilde{h})^{\otimes 2}\} + V(U)$, we always have $V(\hat{\theta}_{h1}) \geq V(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the solution to (3.1). According to the above theorem, we can get the consistent variance estimator by using $\hat{V}_{h1} = \hat{\tau}\hat{\Omega}_{h1}\hat{\tau}'$, where $\hat{\tau} = -\left\{ n^{-1} \sum_{i=1}^{n} r_i \pi_i^{-1} (\partial U_i(\hat{\theta}_{h1})/\partial\theta) \right\}^{-1}$ and $\hat{\Omega}_{h1} = (n-1)^{-1}(\eta_i - \bar{\eta})^2$, where $\eta_i = r_i \pi_i^{-1} \left\{ U_i(\hat{\theta}_{h1}) - \hat{B}\tilde{h}_i(\hat{\theta}_{h1}) \right\} + \hat{B}\tilde{h}_i(\hat{\theta}_{h1})$, $\hat{B} = \hat{E}(U\tilde{h}'/\pi) \left\{ \hat{E}(\tilde{h}\tilde{h}'/\pi) \right\}^{-1}$,

$$\hat{E}(U\tilde{h}'/\pi) = n^{-1} \sum_{i=1}^{n} r_i \pi_i^{-2} U_i(\hat{\theta}_{h1}) \tilde{h}_i'(\hat{\theta}_{h1}), \quad \hat{E}(\tilde{h}\tilde{h}'/\pi) = n^{-1} \sum_{i=1}^{n} r_i \pi_i^{-2} \tilde{h}_i(\hat{\theta}_{h1}) \tilde{h}_i'(\hat{\theta}_{h1}),$$

with $\tilde{h}_i(\hat{\theta}_{h1}) = h_i(\hat{\theta}_{h1}) - \hat{\mu}_h$ and $\hat{\mu}_h = n^{-1} \sum_{i=1}^{n} h_i(\hat{\theta}_{h1})$.

For the special case of $\theta = E(Y)$ and $h = x$, after some algebra, we have

$$\hat{\theta}_{h1} = \hat{\bar{y}}_d - \hat{B}_1 \hat{B}_2^{-1} (\hat{\bar{x}}_d - \bar{x}_n) + o_p(n^{-1/2}),$$

where $(\hat{\bar{y}}_d, \hat{\bar{x}}_d) = (\sum_{r_i=1} \pi_i^{-1})^{-1} (\sum_{r_i=1} \pi_i^{-1} y_i, \sum_{r_i=1} \pi_i^{-1} x_i)$, $\hat{B}_1 = n^{-1} \sum_{r_i=1} \pi_i^{-2} (x_i - \hat{\bar{x}}_d)(y_i - \hat{\bar{y}}_d)$ and $\hat{B}_2 = n^{-1} \sum_{r_i=1} \pi_i^{-2} (x_i - \hat{\bar{x}}_d)^2$, which is close to the optimal estimator within the linear class. The resulting estimator is asymptotically equivalent to the optimal EL estimator considered in Kim (2009).

Remark 3.3.1 *The EL estimator of Chen et al. (2008) satisfies*

$$\sqrt{n}(\hat{\theta}_c - \theta_0) \to^d N(0, V_c),$$

where $V_c = \tau \Omega_{hc} \tau'$,

$$\Omega_{hc} = V\left\{\frac{r}{\pi}U - B^* \frac{(r-\pi)\tilde{h}}{\pi(1-\pi)}\right\},$$

where $B^* = E(U\tilde{h}'/\pi)\left[E\left\{\tilde{h}\tilde{h}'/(\pi(1-\pi))\right\}\right]^{-1}$. *Thus, the estimator of Chen et al. (2008) achieves the minimum variance when* $h/(1-\pi) \propto E(U \mid x)$ *while the asymptotic variance of the proposed EL estimator is minimized when* $h \propto E(U \mid x)$. *The Qin-Zhang-Leung (QZL) estimator* $\hat{\theta}_{QZL}$ *defined in (3.10) and (3.11) satisfies*

$$\sqrt{n}(\hat{\theta}_{hq} - \theta_0) \to^d N(0, V_q),$$

where $V_q = \tau \Omega_{hq} \tau'$,

$$\Omega_{hq} = V\left\{\frac{r}{\pi}U - \frac{r-\pi}{\pi}B_q h\right\},$$

and $B_q = E\{(\pi^{-1}-1)Uh'\}[E\{(\pi^{-1}-1)hh'\}]^{-1}$. *Note that the choice of* $B = B_q$ *minimizes the variance of* $(r/\pi)U - ((r-\pi)/\pi)Bh$ *and the QZL estimator is optimal in the sense that it minimizes the variance among its class. This is because QZL estimator uses the complete likelihood* $\sum_{i=1}^n \log(\omega_i)$ *while our proposed estimator uses only partial likelihood. If* $h \propto E(U|X)$, *then all the estimators, excluding the estimator of Chen et al. (2008), achieve the same asymptotic variance. A numerical comparison is also made through a simulation study in Section 6.*

In the following corollary, we find an optimal constraint that minimizes the asymptotic variance in (3.15). The proof is presented in section C of Appendix B.

Corollary 3.3.1 *Under the setup of Theorem 3.3.1, the asymptotic variance of* $\hat{\theta}_{h1}$ *is minimized when* $h \propto h^* = E(U|X)$. *The asymptotic variance satisfies*

$$V_{h1} \geq \tau\left\{E\left(\frac{UU'}{\pi}\right) - E\left(\frac{1-\pi}{\pi}h^*U'\right)\right\}\tau'. \tag{3.17}$$

The lower bound in (3.17) is the same as the semi-parametric lower bound for the asymptotic variance discussed in Robins et al. (1994) and Chen et al. (2008).

Remark 3.3.2 *To compute the solution to the constrained optimization problem of maximizing (3.12) subject to (3.13), the following two-step algorithm can be used. In the first step, the optimal weight that maximizes (3.12) subject to* $\sum_{r_i=1}\omega_i = 1$ *and* $\sum_{r_i=1}\omega_i\pi_i^{-1}\{\hat{h}_i - n^{-1}\sum_{j=1}^{n}\hat{h}_j\} = 0$ *are computed, where* $\hat{h}_i = h_i(\hat{\theta}_0)$ *and* $\hat{\theta}_0$ *is the solution to* $\sum_{r_i=1}\pi_i^{-1}U_i(\theta) = 0$*. In the second step, we can get the resulting EL estimator* $\hat{\theta}_{h1}$ *by solving*

$$\sum_{i=1}^{N}\hat{\omega}_i\frac{I_i}{\pi_i}U_i(\theta) = 0.$$

Such two-step algorithm was discussed in Chaudhuri, Handcock and Rendall (2008) when the control function h_i *does not depend on* θ*. Using* $\hat{h}_i = h(x_i;\hat{\theta})$*, where* $\hat{\theta}$ *is any* $\sqrt{n}$*-consistent estimator of* θ*, in the two-step optimization is asymptotically equivalent to the original solution.*

3.4 Estimation with unknown response probability

We now consider the case when the response probability is known up to some parameter and has the known form

$$Pr(r = 1|X, Y) = \pi(X;\phi_0),$$

for some ϕ_0. Thus, we assume that the response mechanism is ignorable. We also assume that there exists $\hat{\phi}$ such that

$$\hat{\phi} - \phi_0 = \frac{1}{n}\sum_{i=1}^{n} b(x_i, r_i;\phi_0) + o_p(n^{-1/2}), \tag{3.18}$$

for some function b with $E\{b(X_i, r_i;\phi_0)\} = 0$ and $Var\{b(X_i, r_i;\phi_0)\} = V_b$, where V_b is positive definite.

If the true response probability $\pi_i = \pi_i(\phi_0)$ is estimated by $\hat{\pi}_i = \pi_i(\hat{\phi})$, then the proposed EL estimator can be described as maximizing (3.12) subject to

$$\sum_{r_i=1}\omega_i = 1, \quad \sum_{r_i=1}\omega_i\hat{\pi}_i^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^{n}h_i(\theta)\right\} = 0, \tag{3.19}$$

and

$$\sum_{r_i=1}\omega_i\hat{\pi}_i^{-1}U(\theta;x_i,y_i) = 0. \tag{3.20}$$

The following theorem presents some asymptotic properties of the proposed EL estimator.

Theorem 3.4.1 *Let $\hat{\phi}$ be a $\sqrt{n}$-consistent estimator of ϕ_0, satisfying (3.18). Let $\hat{\theta}_{h2}$ be obtained by maximizing (3.12) subject to the constraints (3.19) and (3.20). Under the same regularity conditions as Theorem 3.3.1 and (3.18), we have*

$$\hat{\theta}_{h2} - \theta_0 = \tau \frac{1}{n} \sum_{i=1}^{n} \{ \frac{r_i}{\pi_i} U_i(\theta_0) - B(\frac{r_i}{\pi_i} - 1)\tilde{h}_i(\theta_0) - Cb_i(\phi_0) \} + o_p(n^{-1/2}), \tag{3.21}$$

where B is defined in (3.14), $\tau = -\{E(\partial U/\partial \theta)\}$, $C = E\left\{\pi^{-1}(U - B\tilde{h})(\partial \pi/\partial \phi)^{'}\right\}$ and $b_i(\phi_0) = b(x_i, r_i; \phi_0)$ defined in (3.18). Hence, we have

$$\sqrt{n}(\hat{\theta}_{h2} - \theta_0) \to^d N(0, V_{h2}),$$

where $V_{h2} = \tau \Omega_{h2} \tau^{'}$, and $\Omega_{h2} = V\left\{r\pi^{-1}(U - B\tilde{h}) + B\tilde{h} - Cb\right\}$.

A consistent variance estimator of V_{h2} can be constructed by

$\hat{V}_{h2} = \hat{\tau}\hat{\Omega}_{h2}\hat{\tau}^{'}$, where $\hat{\tau} = -\left\{n^{-1}\sum_{i=1}^{n} r_i \hat{\pi}_i^{-1}(\partial U_i(\hat{\theta}_{h2})/\partial\theta)\right\}^{-1}$ and $\hat{\Omega}_{h2} = (n-1)^{-1}\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2$, where $\eta_i = r_i\hat{\pi}_i^{-1}\left\{U_i(\hat{\theta}_{h2}) - \hat{B}\tilde{h}_i(\hat{\theta}_{h2})\right\} + \hat{B}\tilde{h}_i(\hat{\theta}_{h2}) - \hat{C}b_i(\hat{\phi})$,

$$\hat{C} = n^{-1}\sum_{i=1}^{n} r_i\hat{\pi}_i^{-2}\left\{U_i(\hat{\theta}_{h2}) - \hat{B}\tilde{h}_i(\hat{\theta}_{h2})\right\}(\partial\hat{\pi}_i/\partial\phi)^T, \quad \hat{B} = \hat{E}(Uh^{'}/\pi)\left\{\hat{E}(\tilde{h}\tilde{h}^{'}/\pi)\right\}^{-1},$$

where $\hat{E}(U\tilde{h}^{'}/\pi) = n^{-1}\sum_{i=1}^{n} r_i\hat{\pi}_i^{-2}U_i(\hat{\theta}_{h2})\tilde{h}_i^{'}(\hat{\theta}_{h2})$, $\hat{E}(\tilde{h}\tilde{h}^{'}/\pi) = n^{-1}\sum_{i=1}^{n} r_i\hat{\pi}_i^{-2}\tilde{h}_i(\hat{\theta}_{h2})\tilde{h}_i^{'}(\hat{\theta}_{h2})$, with $\tilde{h}_i(\hat{\theta}_{h2}) = h_i(\hat{\theta}_{h2}) - \hat{\mu}_h$ and $\hat{\mu}_h = n^{-1}\sum_{i=1}^{n} h_i(\hat{\theta}_{h2})$.

Comparing (3.21) with (3.14), we have an extra term, $-Cb_i(\phi_0)$, in the linearization. This is because we have additional randomness due to estimating parameter ϕ_0.

Remark 3.4.1 *If we use $h = ah^* = aE(U|X)$ in the constraint (3.19) for some constant $a \neq 0$, we have $B = E(Uh'/\pi)E^{-1}(hh'/\pi) = a^{-1}I$ and*

$$C = E\left\{\pi^{-1}(U - Bh)(\partial\pi/\partial\phi)'\right\} = E\left[E\left\{\pi^{-1}(U - h^*)(\partial\pi/\partial\phi)'|X\right\}\right] = 0.$$

Thus, the asymptotic variance is equal to

$$V_{h2} = \tau\left\{E\left(\frac{UU'}{\pi}\right) - E\left(\frac{1-\pi}{\pi}h^*U'\right)\right\}\tau',$$

which is equal to the lower bound in (3.17) when the propensity score is known. Under the optimal choice of h, the lower bound for the asymptotic variance is achieved regardless of whether the propensity score is known or estimated.

According to Remark 3.4.1, the choice of $\hat{\phi}$ does not make any difference in the asymptotic variance of $\hat{\theta}_{h2}$, as long as $h \propto E(U \mid X)$ is used in (3.19). If $h \propto E(U \mid X)$ does not hold, then the choice of $\hat{\phi}$ makes a difference. While the MLE of ϕ_0 is a popular choice, it does not necessarily lead to the optimal estimator. To see this, let $\hat{\phi}_q$ be a consistent estimator of ϕ_0 that can be obtained by solving the following equation:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{r_i}{\pi_i(\phi)}-1\right\}q_i(\phi)=0, \tag{3.22}$$

where $q_i(\phi)$ is an arbitrary function to make the solution of (3.22) unique. Note that equation (3.22) can be called the calibration equation in the sense that the estimator for the mean of q_i using the propensity score is equal to the sample mean of q_i. The MLE of ϕ_0 also belongs to the class because it satisfies (3.22) with $q_i = \pi_i \text{logit}(\pi_i)$. Under some regularity conditions, we have $\hat{\phi}_q \to^p \phi_0$, regardless of the choice of q_i. However, the efficiency can be different for a different choice of q_i. The following theorem discusses the optimal choice of q in the calibration equation (3.22).

Theorem 3.4.2 *Let $\hat{\phi}_q$ be the estimator which solves (3.22) and satisfies $\hat{\phi}_q \to^p \phi_0$. Under the same regularity conditions as Theorem 3.4.1, we have*

$$\hat{\theta}_{h2}-\theta_0=\tau\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\pi_i}U_i(\theta_0)-(\frac{r_i}{\pi_i}-1)\left\{B\tilde{h}_i(\theta_0)+CS^{-1}q(\phi_0)\right\}\right]+o_p(n^{-1/2}), \tag{3.23}$$

where B is defined in (3.14), $\tau = -\{E(\partial U/\partial\theta)\}$, $S = E\left\{\pi^{-1}q(\partial\pi/\partial\phi)'\right\}$, $C = E\left\{\pi^{-1}(U-B\tilde{h})(\partial\pi/\partial\phi)'\right\}$. Hence, we have

$$\sqrt{n}(\hat{\theta}_{h2}-\theta_0)\to^d N(0,V_{h2}),$$

where $V_{h2} = \tau\Omega_{h2}\tau'$, and $\Omega_{h2} = V\left\{r\pi^{-1}U-(r\pi^{-1}-1)(B\tilde{h}+CS^{-1}q)\right\}$. In addition, we have

$$V_{h2} \geq \tau V\left\{r\pi^{-1}U-(r\pi^{-1}-1)h^*\right\}\tau'$$

with equality if $\alpha' q = h^* - B\tilde{h}$ for some α.

In Theorem 3.4.2, the meaning of $h^* - B\tilde{h}$ is the residual for the regression of $h^* = E(U|X)$ on $\tilde{h}$. If $h \propto h^*$, then the residual is equal to zero and the lower bound is achieved, as discussed in Remark 3.4.1. If $h \propto h^*$ does not hold, we cannot achieve the lower bound and the efficiency can be improved by how well q explains the conditional expectation $E(U \mid X)$. In the extreme case of $h \equiv 0$, the choice of $q \propto h^*$ achieves the lower bound while the choice of $q_i \propto \pi_i h^*$, which corresponds to the maximum likelihood estimation of ϕ_0, does not achieve the lower bound and thus leads to less efficient estimation.

3.5 Nonparametric estimation of the response mechanism

In this section, we consider nonparametric estimation of the response probability. For simplicity, we assume the response mechanism is ignorable, $\pi(x) = Pr(r = 1|x)$, and consider estimation of π nonparametrically. To this end, we consider kernel estimation of the response model as below:

$$\hat{\pi}_H(x) = \frac{\sum_{i=1}^n K_H(x - X_i) r_i}{\sum_{j=1}^n K_H(x - X_j)}, \tag{3.24}$$

where $K_H(s)$ is the kernel function which satisfies certain regularity conditions and H is the bandwidth. In addition, we define $K_H(s,t) = K((s-t)/H)$. Let $f(x)$ be the probability density function of X. In addition to regularity conditions (C1)-(C5) in the section A of Appendix B, we also assume the following regularity conditions:

(C6) $f(x)$ and $\pi(x)$ have bounded partial derivatives with respect to x up to an order q with $q \geq 2$, $2q > d_x$ almost surely, where d_x is the dimension of x.

(C7) The kernel function $K(s)$ is a probability density function such that

1. It is bounded and has compact support.
2. $\int K(s_1, \ldots, s_{d_x}) ds_1 \ldots ds_{d_x} = 1$,
3. $\int s_i^l K(s_1, \ldots, s_{d_x}) ds_1 \ldots ds_{d_x} = 0$ for any $i = 1, \ldots, d_x$ and $1 \leq l < q$.
4. $\int s_i^q K(s_1, \ldots, s_{d_x}) ds_1 \ldots ds_{d_x} \neq 0$.

(C8) $nH^{2d_x} \to \infty, \quad \sqrt{n} H^q \to 0$, as $n \to \infty$.

(C9) $1 > \pi(x) > d > 0$ almost surely.

Conditions (C6)-(C8) are common conditions used for nonparametric problems. In condition (C8), we used $\sqrt{n} H^q \to 0$ to control the bias due to kernel smoothing, and $nH^{2d_x} \to \infty$

is used to produce consistent estimation of the conditional distribution as well as control the convergence rate of response probability estimation. Condition (C9) is used to avoid extreme propensity scores.

Under those regularity conditions, the proposed empirical likelihood method can be constructed similarly by maximizing (3.12) subject to

$$\sum_{r_i=1}\omega_i=1,\quad \sum_{r_i=1}\omega_i\hat{\pi}_{i,H}^{-1}\{h_i(\theta)-n^{-1}\sum_{i=1}^{n}h_i(\theta)\}=0,\quad \sum_{r_i=1}\omega_i\hat{\pi}_{i,H}^{-1}U_i(\theta)=0, \tag{3.25}$$

where $\hat{\pi}_{i,H}=\hat{\pi}_H(x_i)$.

The following theorem presents some asymptotic properties of the proposed EL estimator of θ_0 using nonparametric response probability (3.24).

Theorem 3.5.1 *Let $\hat{\theta}_{h3}$ be the empirical likelihood estimator that is obtained by maximizing (3.12) subject to (3.25). Under the regularity conditions (C1)-(C9), we have*

$$\hat{\theta}_{h3}-\theta_0 \;=\; -\tau\{\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i}U_i(\theta_0)-\frac{1}{n}\sum_{i=1}^{n}(\frac{r_i}{\pi_i}-1)h_i^*(\theta_0)\}+o_p(n^{-1/2}), \tag{3.26}$$

where $h^*(\theta_0)=E\{U(\theta_0)|X\}$ and $\tau=-\{E(\partial U/\partial\theta)\}^{-1}$ evaluated at $\theta=\theta_0$. Furthermore, we have

$$\sqrt{n}(\hat{\theta}_{h3}-\theta_0)\to^d N(0,V_{h3}),$$

where $V_{h3}=\tau\Omega_{h3}\tau^{'}$ and $\Omega_{h3}=V\left\{r\pi^{-1}(U-h^*)+h^*\right\}$.

By Theorem 3.5.1, the asymptotic variance of $\hat{\theta}_{h3}$ using nonparametric $\hat{\pi}_{i,H}$ is equal to the semiparametric lower bound in (3.17). Note that the linearization in (3.26) does not depend on the choice of h in (3.25). This means that the same result (3.26) can be achieved for different choices of h. This is because, according to (F.9) in the proof of Theorem 3.5.1 in Section F of Appendix B, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_{i,H}}\tilde{h}_i-\frac{1}{n}\sum_{i=1}^{n}\tilde{h}_i=o_p(n^{-1/2}),$$

where $\tilde{h}=h-\mu_h$ and h is any arbitrary function of x with some moment conditions described in section A of Appendix B. In addition, the second constraint in (3.25) can be written as $\sum_{r_i=1}\omega_i\hat{\pi}_{i,H}^{-1}(\tilde{h}_i-n^{-1}\sum_{i=1}^{n}\tilde{h}_i)=0$, which implies that we can safely remove it. Thus, using the nonparametric estimator (3.24) of the response probability, the EL solution can be written as maximizing (3.12) subject to

$$\sum_{r_i=1} \omega_i = 1, \quad \sum_{r_i=1} \omega_i \hat{\pi}_{i,H}^{-1} U_i(\theta) = 0, \tag{3.27}$$

which is equivalent to obtain $\hat{\theta}_{h3}$ by solving $n^{-1} \sum_{i=1}^{n} r_i \hat{\pi}_{i,H}^{-1} U_i(\theta) = 0$.

According to Theorem 3.5.1, a consistent variance estimator for V_{h3} is $\hat{V}_{h3} = \hat{\tau} \hat{\Omega}_{h3} \hat{\tau}'$, where $\hat{\tau} = -\left\{ n^{-1} \sum_{i=1}^{n} r_i \hat{\pi}_{i,H}^{-1} \partial U_i(\hat{\theta}_{h3}) / \partial \theta \right\}^{-1}$ and $\hat{\Omega}_{h3} = (n-1)^{-1} \sum_{i=1}^{n} (\eta_i - \bar{\eta})^2$, where $\eta_i = r_i \hat{\pi}_{i,H}^{-1} (U_i(\hat{\theta}_{h3}) - \hat{h}_i^*) + \hat{h}_i^*$, where $\hat{h}_i^*$ is a consistent estimator of $E(U|X)$ under the working model. In the special case when θ_0 satisfies $E\{U(\theta_0)|X\} = 0$, then a version of Wilk's Theorem can be established as below.

Theorem 3.5.2 *Assume that the regularity conditions in Theorem 3.5.1 hold and* θ_0 *satisfies* $E\{U(\theta_0)|X\} = 0$. *Let* $R_n(\theta_0) = 2\left\{ l(\hat{\theta}_{h3}) - l(\theta_0) \right\}$, *where* $l(\theta) = \sum_{r_i=1} \log\{\omega_i(\theta)\}$ *with* $\omega_i(\theta)$ *obtained by maximizing (3.12) subject to (3.27). Then, as* $n \to \infty$,

$$R_n(\theta_0) \to^d \chi_p^2,$$

where p *is the dimension of* θ.

According to Theorem 3.5.2, we can construct a Wilk-type confidence region for θ_0 without calculating the variance estimator, if $E\{U(\theta_0)|X\} = 0$ holds. For example, if $E(Y|X) = X^T\theta$, then $U = (Y - X^T\theta)X$ satisfies $E\{U(\theta_0)|X\} = 0$. If $E(U|X) \neq 0$ but we know $E(U|X)$, the result for Theorem 3.5.2 holds by replacing U with $U^* = U - E(U|X)$. Alternatively, a resampling method, such as bootstrap or jackknife, can be used to construct a confidence region for θ_0.

3.6 Extension to two-phase sampling

In this section, we briefly discuss an extension of the proposed method to two-phase sampling. In two-phase sampling, the first-phase sample is selected from a probability sampling design and an auxiliary variable X is observed from the first-phase sample. The second-phase sample is selected from the first-phase sample based on the conditional sampling design. An important application of the two-phase sampling is that missing data in complex sampling problem can be treated as a special case of two-phase sampling. Let A_1 be the set of sample indices in the first-phase sample obtained by a probability sampling design whose first-order inclusion probabilities are given by $\pi_{1i} = Pr\,(i \in A_1)$. Let $A_2 \subset A_1$ be the set of sample indices in the second-phase sample obtained by another probability sampling design whose first-order inclusion probabilities are given by $\pi_{2i|1i} = Pr\,(i \in A_2 \mid i \in A_1)$. We observe X_1 throughout the finite population, observe X_2 in the first-phase sample, and observe Y from the second-phase sample. Table 3.1 presents the data structure for two-phase sampling. We are interested in estimating θ that is the solution to $\sum_{i\in U} U(\theta; x_i, y_i) = 0$ where $x_i = (x_{1i}, x_{2i})$. The direct expansion estimator obtained by solving

$$\sum_{i\in A_2} \frac{1}{\pi_{1i}} \frac{1}{\pi_{2i|A_1}} U(\theta; x_i, y_i) = 0$$

is consistent but does not incorporate all available information.

Table 3.1 Data structure for two-phase sampling

Set	Size	Observation
Population (U)	N	x_{1i}
First-phase sample (A_1)	n_1	$x_i = (x_{1i}, x_{2i})$
Second-phase sample (A_2)	n_2	x_i, y_i

In this case, the proposed empirical likelihood can be formulated by maximizing

$$l_2 = \sum_{i\in A_2} \log\,(\omega_{2i})$$

subject to

$$\sum_{i\in A_2} \omega_{2i} = 1, \quad \sum_{i\in A_2} \omega_{2i} \frac{1}{\pi_{1i}} \frac{1}{\pi_{2i|A_1}} \left\{ h_1(x_{1i}) - \bar{h}_1 \right\} = 0, \quad \sum_{i\in A_2} \omega_{2i} \frac{1}{\pi_{1i}} \frac{1}{\pi_{2i|A_1}} \left\{ h_2(x_{2i}) - \bar{h}_{2,EL} \right\} = 0$$

and

$$\sum_{i\in A_2} \omega_{2i} \frac{1}{\pi_{1i}} \frac{1}{\pi_{2i|A_1}} U(\theta; x_i, y_i) = 0,$$

where $\bar{h}_1 = N^{-1} \sum_{i=1}^{N} h_1(x_{1i})$, and $\bar{h}_{2,EL}$ is an EL estimator of $\bar{h}_2 = N^{-1} \sum_{i=1}^{N} h_2(x_{2i})$. The EL estimator of $\bar{h}_2$ can be obtained by

$$\bar{h}_{2,EL} = \frac{\sum_{i\in A_1} \omega_{1i}^{*} \pi_{1i}^{-1} h_2(x_{2i})}{\sum_{i\in A_1} \omega_{1i}^{*} \pi_{1i}^{-1}}$$

where ω_{1i}^* are obtained by maximizing

$$l_1 = \sum_{i \in A_1} \log(\omega_{1i})$$

subject to

$$\sum_{i \in A_1} \omega_{1i} = 1, \quad \sum_{i \in A_1} \omega_{1i} \frac{1}{\pi_{1i}} \left\{ h_1(x_{1i}) - \bar{h}_1 \right\} = 0.$$

Thus, the proposed EL method can be performed in two-steps, which is quite attractive in practice as it can be easily extended to multi-phase sampling which is often the case in real complex sampling. Let $\pi_{2i} = \pi_{1i}\pi_{2i|A_1}$. The asymptotic properties of the proposed estimator are obtained by the following theorem.

Theorem 3.6.1 *Let $\hat{\theta}_{EL}$ be the proposed empirical likelihood estimator by the above procedures, under certain regularity conditions, we have*

$$\hat{\theta}_{EL} - \theta_0 = -S_{11}^{-1} \left\{ \bar{U}_{HT,2} - \bar{V}_{HJ} \right\} + o_p(n^{-1/2}), \tag{3.28}$$

where $S_{11} = N^{-1} \sum_{i=1}^{N} \partial U_i / \partial \theta$, $\bar{U}_{HT,2} = N^{-1} \sum_{i \in A_2} \pi_{1i}^{-1} \pi_{2i|A_1}^{-1} U_i(\theta_0)$ *and*

$$\bar{V}_{HJ} = B_1(\bar{h}_{HJ,1,2} - \bar{h}_1) + B_2(\bar{h}_{HJ,2,2} - \bar{h}_2) + B_3(\bar{h}_{HJ,2,1} - \bar{h}_2) + B_4(\bar{h}_{HJ,1,1} - \bar{h}_1),$$

with

$$\bar{h}_{HJ,s,t} = \frac{\sum_{i \in A_t} \pi_{ti}^{-1} h_{si}}{\sum_{i \in A_t} \pi_{ti}^{-1}}, \quad s, t \in (1, 2)$$

$(B_1, B_2) = S_{14} S_{24}^{-1}$, $S_{14} = (a, b)$, $B_3 = (aA^{-1}B - b)(D - CA^{-1}B)^{-1}$ *and* $B_4 = -(aA^{-1}B - b)(D - CA^{-1}B)^{-1} S_{33} S_{43}^{-1}$, *with*

$$a = N^{-1} \sum_{i=1}^{N} \pi_{2i}^{-1} U_i (h_{1i} - \bar{h}_1)', \quad b = N^{-1} \sum_{i=1}^{N} \pi_{2i}^{-1} U_i (h_{2i} - \bar{h}_2)',$$

$$S_{24} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} N^{-1} \sum_{i=1}^{N} \pi_{2i}^{-1} (h_{1i} - \bar{h}_1)^{\otimes 2} & N^{-1} \sum_{i=1}^{N} \pi_{2i}^{-1} (h_{1i} - \bar{h}_1)(h_{2i} - \bar{h}_2)' \\ N^{-1} \sum_{i=1}^{N} \pi_{2i}^{-1} (h_{2i} - \bar{h}_2)(h_{1i} - \bar{h}_1)' & N^{-1} \sum_{i=1}^{N} \pi_{2i}^{-1} (h_{1i} - \bar{h}_1)^{\otimes 2} \end{bmatrix},$$

$$S_{33} = N^{-1} \sum_{i=1}^{N} \pi_{1i}^{-1} (h_{2i} - \bar{h}_2)(h_{2i} - \bar{h}_2)', \quad S_{43} = N^{-1} \sum_{i=1}^{N} \pi_{1i}^{-1} (h_{1i} - \bar{h}_1)^{\otimes 2}.$$

In the missing data problem, we do not know $\pi_{2i|1i}$ but a consistent estimator $\hat{\pi}_{2i|1i}$ is available. In this case, we have only to replace $\pi_{2i|1i}$ by $\hat{\pi}_{2i|1i}$ in the proposed EL method.

3.7 Simulation Study

3.7.1 Simulation One

Two simulation studies were performed to compare the estimators. In the first simulation study, the following two models were considered to generate the samples:

[A] $x_i \stackrel{iid}{\sim} N(1,1)$, $z_i \stackrel{iid}{\sim} N(0,1)$, $e_i \sim^{iid} exp(1) - 1$, and $y_i = 0.5 + 0.5x_i + 0.5z_i + e_i$.

[B] (x_i, z_i, e_i) are the same as in [A] and $y_i = 0.8(x_i - 0.5)^2 + 0.8e_i$.

For each model, $B = 2,000$ Monte Carlo samples of size $n = 200$ were independently generated. In addition, two different response mechanisms were used to generate r_i, the response indicator function for y_i. The response mechanisms are

[M1] $P(r = 1|x, y) = \exp(-0.5 + x)/[1 + \exp(-0.5 + x)]$

[M2] $P(r = 1|x, y) = (0.3 + 0.175|x|)I(|x| < 1.5) + I(|x| \geq 1.5)$

Thus, the two models are ignorable and the response rate is about 0.6 in both response mechanisms. We are interested in estimating $\theta_0 = E(Y)$, which is the population mean of Y. Thus, we use $U(\theta) = Y - \theta$. We assume that the working model for $E(y \mid x, z)$ is linear in x and z. That is, $E(y \mid x, z) = \beta_0 + \beta_1 x + \beta_2 z$. Also, the working model for $\pi(x) = E(r \mid x)$ is the logistic regression model with $\text{logit}\{\pi(x)\} = \phi_0 + \phi_1 x$. That is, even when the true response mechanism is [M2], we use the logistic regression model to obtain $\hat{\pi}_i = \exp(\hat{\phi}_0 + \hat{\phi}_1 x_i)/\left\{1 + \exp(\hat{\phi}_0 + \hat{\phi}_1 x_i)\right\}$. Thus, we have the following four possible scenarios:

1: Both working models are correct. That is, the samples are generated by [A] and [M1].

2: Only the outcome regression model, the model for $E(y \mid x)$, is correct. That is, the samples are generated by [A] and [M2].

3: Only the response probability model, the model for $E(r \mid x)$, is correct. That is, the samples are generated by [B] and [M1].

4: Both models are incorrect. That is, the samples are generated by [B] and [M2].

Under this setup, we considered eight estimators of θ_0.

1. QZ: The EL estimator of Qin and Zhang (2007) using $h_i = (x_i, z_i)'$ in (3.8).

2. CLQ: The EL estimator of Chen et al. (2008) using $h_i = (x_i, z_i)'$ in (3.9).

3. QZL1: The EL estimator of Qin et al. (2009), which is obtained by maximizing $\sum_{i=1}^{n} \log(\omega_i)$, subject to $\sum_{i=1}^{n} \omega_i = 1$, $\sum_{i=1}^{n} \omega_i r_i \hat{\pi}_i^{-1}(y_i - \theta) = 0$ and $\sum_{i=1}^{n} \omega_i (r_i \hat{\pi}_i^{-1} - 1) h_i = 0$, with $h_i = (x_i, z_i)'$ and $\hat{\pi}_i$ is computed by the MLE of ϕ.

4. QZL2: The EL estimator of Qin et al. (2009), which is obtained by maximizing $\sum_{i=1}^{n} \log(\omega_i)$, subject to (3.10) and (3.11) with $h_i = (x_i, z_i)'$.

5. NEW (MLE): The proposed EL estimator using $h_i = (x_i, z_i)'$ in (3.19), where $\hat{\pi}_i$ is computed by the MLE of ϕ.

6. NEW (CAL): The proposed EL estimator using $h_i = (x_i, z_i)'$ in (3.19), where $\hat{\pi}_i$ is computed by the calibration method on $(1, x)$. That is, $\hat{\phi}$ is computed by solving (3.22) with $q_i = (1, x_i)$.

7. NEW (NP1): The proposed EL estimator using nonparametric estimator (3.24) of $\pi(x) = P(r = 1 \mid x)$ and $h_i = (x_i, z_i)'$ in (3.19). In addition, we used Gaussian kernel and the reference bandwidth $H = 1.06\hat{\sigma}_x n^{-1/5}$, where $\hat{\sigma}_x$ is the estimated standard deviation of x_i in the sample.

8. NEW (NP2): The proposed EL estimator using nonparametric estimator (3.24) of $\pi(x) = P(r = 1 \mid x)$ without using $h_i = (x_i, z_i)'$ in (3.19). We used the same kernel density and bandwidth as NEW (NP1).

Table 3.2 presents the Monte Carlo biases, variances, and mean square errors of the eight estimators under the four difference scenarios. Under Scenario 1, when both the outcome regression model and the response probability model are correct, the simulation results in Table 3.2 show that all the estimators are comparable since they all achieve the semiparametric lower bound except for CLQ, as discussed in Remark 3.3.1 and Remark 3.4.1. The NP2 method also shows some efficiency loss because the nonparametric propensity estimator does not make use of z_i information. Under Scenario 2, when only the outcome regression model is correct, the CLQ estimator shows large bias, suggesting that the CLQ estimator is not robust against the failure of the response model. In terms of efficiency, the QZ method, QZL1 method, and the proposed EL estimators show the smallest variances. Under Scenario 3, when only the response probability model is correct, the biases are all negligible. The QZL estimator is more efficient than the proposed EL estimator using MLE, which is discussed in Remark 3.3.1. The nonparametric estimators, NEW (NP1) and NEW (NP2), show good efficiency because they automatically achieve the lower bound in (3.17) without correctly specifying the outcome regression model, which is consistent with the theory in Theorem 3.5.1. The NEW (NP2) is slightly more efficient than NEW (NP1) because it does not use calibration on the wrong outcome regression model. When both models are incorrect, as in Scenario 4, the nonparametric estimators still show negligible bias because they estimate the response probability consistently. In terms of efficiency, the nonparametric estimators are quite comparable because they achieve the semi-parametric lower bound.

3.7.2 Simulation Two

In this simulation, we assume the finite population of size $N = 1,000$ is generated from the following model:

$$y_i = 1 + 0.8(z_i - 2) + 1.5(x_{1i} - 2) + 0.5(x_{2i} - 2)^2 + z_i/5e_i, \quad i = 1, 2, \dots, N,$$

where (x_{1i}, x_{2i}) is generated from a bivariate normal distribution with parameter

$$(\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}) = (2, 2, 1, 0.6, 1),$$

$z_i \sim \chi^2(1) + 1$ and $e_i \sim \chi^2(1) - 1$. The first phase sample is selected by simple random sampling without replacement with sample size $n_1 = 200$, so we have $\pi_{1i} = n_1/N$. The second phase sample is obtained via a Poisson sampling with the selection probability $\pi_{2i|A_1} = n_2 z_i / \sum_{i \in A_1} z_i$, where $n_2 = 50$. Assume x_{1i} is observed for all elements in the population, (x_{1i}, x_{2i}) is observed in A_1 and (x_{1i}, x_{2i}, y_i) is observed in A_2. Assume the parameter of interest is $\theta_0 = \bar{Y}_N$. The Monte Carlo sample size for this study is $B = 5,000$. The following estimators were computed:

1. Hájek double expansion estimator (HJ):

$$\hat{\theta}_{HJ} = \frac{\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1} y_i}{\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1}}.$$

2. Horvitz-Thompson double expension estimator (HT): $\hat{\theta}_{HT} = N^{-1}\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1} y_i$.

3. Two-phase regression estimator (REG):

$$\hat{\theta}_{REG} = \bar{y}_{\pi,2} - \hat{B}_1(\bar{x}_{\pi,1,2} - \bar{X}_{1,N}) - \hat{B}_2(\bar{x}_{\pi,2,2} - \bar{x}_{\pi,2,1}),$$

where

$$\bar{y}_{\pi,2} = \hat{N}^{-1}\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1} y_i, \quad \bar{x}_{\pi,1,2} = \hat{N}^{-1}\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1} x_{1i},$$
$$\bar{x}_{\pi,2,2} = \hat{N}^{-1}\sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1} x_{2i}, \quad \bar{x}_{\pi,2,1} = \hat{N}^{-1}\sum_{i\in A_1} \pi_{1i}^{-1} x_{2i},$$

with $\hat{N} = \sum_{i\in A_2} \pi_{1i}^{-1}\pi_{2i|A_1}^{-1}$ and

$$\begin{aligned}
\hat{B} &= (\hat{B}_1, \hat{B}_2) \\
&= \left(\sum_{i\in A_2} \pi_{2i}^{-1} y_i(x_{1i} - \bar{x}_{\pi,1,2}), \sum_{i\in A_2} \pi_{2i}^{-1} y_i(x_{2i} - \bar{x}_{\pi,2,2})\right) \\
&\times \begin{bmatrix} \sum_{i\in A_2} \pi_{2i}^{-1}(x_{1i} - \bar{x}_{\pi,1,2})^{\otimes 2} & \sum_{i\in A_2} \pi_{2i}^{-1}(x_{1i} - \bar{x}_{\pi,1,2})(x_{2i} - \bar{x}_{\pi,2,2})' \\ \sum_{i\in A_2} \pi_{2i}^{-1}(x_{2i} - \bar{x}_{\pi,2,2})(x_{1i} - \bar{x}_{\pi,1,2})' & \sum_{i\in A_2} \pi_{2i}^{-1}(x_{2i} - \bar{x}_{\pi,2,2})^{\otimes 2} \end{bmatrix}^{-1},
\end{aligned}$$

and $\pi_{2i} = \pi_{1i}\pi_{2i|A_1}$.

4. Proposed EL estimator (New) defined in (3.28).

Table 3.3 presents the Monte Carlo biases, variances, and mean square errors of the eight estimators under the four difference scenarios. Both regression estimator and the proposed EL estimator show smaller variances than the two double expansion estimators. The proposed EL estimator is more efficient than the regression estimator because it uses an improved estimator for $\bar{x}_{2,N}$ using the first-step EL method. In fact, the proposed EL estimator is asymptotically equivalent to the two-step regression estimator

$$\hat{\theta}_{REG2} = \bar{y}_{\pi,2} - \hat{B}_1(\bar{x}_{\pi,1,2} - \bar{X}_{1,N}) - \hat{B}_2(\bar{x}_{\pi,2,2} - \bar{x}_{reg,2}),$$

where $\bar{x}_{reg,2} = \bar{x}_{\pi,2,1} - \hat{C}_1(\bar{x}_{\pi,1,1} - \bar{X}_{1,N})$ is the regression estimator of $\bar{X}_{2,N}$ using $\bar{X}_{1,N}$ information. The two-step regression estimator under two-phase sampling was originally considered by Dupont (1995).

Table 3.2 Biases, Variances and Mean squared errors (MSE) of the estimators under four different scenarios in simulation one.

Scenario	Method	Bias	Var	MSE
1	QZ	0.00	0.0129	0.0130
	CLQ	-0.00	0.0147	0.0147
	QZL1	0.00	0.0130	0.0131
	QZL2	0.00	0.0137	0.0138
	NEW(MLE)	0.00	0.0130	0.0130
	NEW(CAL)	0.00	0.0130	0.0130
	NEW(NP1)	0.00	0.0130	0.0131
	NEW(NP2)	0.04	0.0129	0.0147
2	QZ	0.00	0.0119	0.0119
	CLQ	-0.16	0.0189	0.0467
	QZL1	0.01	0.0116	0.0118
	QZL2	-0.02	0.0148	0.0155
	NEW(MLE)	0.00	0.0119	0.0119
	NEW(CAL)	0.00	0.0119	0.0119
	NEW(NP1)	0.00	0.0120	0.0120
	NEW(NP2)	0.01	0.0128	0.0130
3	QZ	-0.03	0.0231	0.0247
	CLQ	-0.01	0.0210	0.0213
	QZL1	-0.02	0.0197	0.0202
	QZL2	-0.01	0.0173	0.0174
	NEW(MLE)	-0.02	0.0220	0.0227
	NEW(CAL)	-0.02	0.0216	0.0221
	NEW(NP1)	-0.05	0.0169	0.0196
	NEW(NP2)	-0.00	0.0156	0.0157
4	QZ	0.26	0.0307	0.0988
	CLQ	0.39	0.0602	0.2140
	QZL1	0.22	0.0240	0.0762
	QZL2	0.03	0.0564	0.0576
	NEW(MLE)	0.29	0.0285	0.1150
	NEW(CAL)	0.30	0.0319	0.1255
	NEW(NP1)	0.03	0.0169	0.0180
	NEW(NP2)	0.04	0.0168	0.0189

Table 3.3 The Monte Carlo biases, variances, and the mean squared errors (MSE) of the point estimators in simulation two.

Method	Bias	Var	MSE
HJ	0.0100	0.1038	0.1039
HT	-0.0077	0.1211	0.1211
REG	0.0140	0.0393	0.0395
NEW	0.0117	0.0388	0.0389

CHAPTER 4. POPULATION EMPIRICAL LIKELIHOOD FOR NONPARAMETRIC INFERENCE IN SURVEY SAMPLING

A paper submitted to the *Statistica Sinica(revision invited)*

Sixia Chen and Jae Kwang Kim

Abstract

Empirical likelihood is a popular tool for incorporating auxiliary information and constructing nonparametric confidence intervals. In survey sampling, sample elements are often selected by using an unequal probability sampling method and the empirical likelihood function needs to be modified to account for the unequal probability sampling. Wu and Rao (2006) proposed a way of constructing confidence regions using the pseudo empirical likelihood of Chen and Sitter (1999).

In this paper, we propose a novel approach of using empirical likelihood in survey sampling based on the so-called population empirical likelihood (POEL). In the POEL approach, a single empirical likelihood is defined for the finite population. The sampling design can be incorporated into the constraint in the optimization of the POEL. For some special sampling designs, the proposed method leads to optimal estimation and does not require artificial adjustment for constructing the likelihood ratio confidence intervals. Furthermore, because a single empirical likelihood is defined for the finite population, it naturally incorporates auxiliary information obtained from multiple surveys. Results from two simulation studies are presented to show the finite sample performance of the proposed method.

Key Words: Calibration estimation, Optimal estimation, Regression estimation, Wilk's the-

orem.

4.1 Introduction

The empirical likelihood method, proposed by Owen (1988, 1990), provides a useful tool for obtaining nonparametric confidence regions for statistical functionals. Even though the empirical likelihood method is a nonparametric approach in the sense that it does not require a parametric model for the underlying distribution of the sample observations, the empirical likelihood method shares most of the desirable properties of the likelihood-based method. Using a nonparametric likelihood function, the empirical likelihood method can easily incorporate both known constraints on parameters and also prior information about parameters obtained from other sources. For example, Chen and Qin (1993), Qin (2000), and Chaudhuri, Handcock and Rendall (2008) discussed combining information using empirical likelihood. Qin and Lawless (1994) considered the situation when the parameter of interest is the solution to an estimating equation. A comprehensive overview of the empirical likelihood method is provided by Owen (2001).

When the sample is selected by an unequal probability sampling method from the finite population, the empirical likelihood needs to be modified to incorporate the sampling design. Chen and Sitter (1999) considered the pseudo empirical likelihood estimator that uses the sampling weight in the empirical log-likelihood function. Kim (2009) considered an alternative empirical likelihood function based on the biased sampling likelihood of Vardi (1985) and Qin (1993). In either case, the resulting empirical likelihood estimator naturally incorporates the available population information and achieves optimality under some limited situations. Because the empirical likelihood function is changed to incorporate the unequal probability sampling design, the resulting confidence interval based on the likelihood ratio does not have a limiting chi-square distribution and often extra computations, as discussed in Wu and Rao (2006), are required to obtain a Wilk-type confidence region. Furthermore, the sample-based empirical likelihood approach can be problematic if we want to combine information from two independent surveys, since there are different empirical likelihood functions associated with each sample.

In this paper, we propose a novel approach for the empirical likelihood in survey sampling based on the so-called population empirical likelihood (POEL). In this POEL approach, a single empirical likelihood is defined for the finite population and the sampling design can be incorporated as a constraint in the empirical likelihood. For some sampling designs, such as the Poisson sampling or the rejective Poisson sampling of Fuller (2009a), the proposed method leads to optimal estimation and the likelihood ratio follows a chi-square distribution in the limit if the sampling rate is negligible. Thus, unlike the pseudo empirical likelihood method, a Wilk-type confidence interval based on the POEL can be constructed without any artificial adjustment. Furthermore, because a single empirical likelihood is defined for the entire finite population, it naturally incorporates the setup of combining multiple surveys. The resulting empirical likelihood estimator is asymptotically equivalent to the optimal estimator obtained by the generalized method of moments (GMM), but it avoids the burden of computing the variance-covariance matrix for the GMM computation.

The rest of this paper is organized as follows. In Section 2, basic setup is introduced and the population empirical likelihood method is presented. In Section 3, some asymptotic properties of the proposed estimator are discussed under Poisson sampling. The proposed method is extended to the rejective Poisson sampling in Section 4 and also extended to the problem of combining independent surveys in Section 5. Results from two limited simulation studies are presented in Section 6. Concluding remarks are made in Section 7. All the technical details are given in Appendix C.

4.2 Population empirical likelihood

Consider a finite population (x_i, y_i) of size N, and we assume N is known. Suppose we are interested in estimating parameter θ_0 that is defined by solving

$$\sum_{i=1}^{N} U(x_i, y_i; \theta) = 0, \tag{4.1}$$

for θ. Many finite-population parameters can be defined as the solutions to (4.1). If the parameter of interest is the population total $Y = \sum_{i=1}^{N} y_i$, we can define $\theta = N\mu_y$ and μ_y through

(4.1) with $U(X,Y;\mu_y) = (Y - \mu_y)$. For simplicity, we assume that the solution θ_0 to (4.1) is unique.

Suppose now that a sample of size n is selected from the population using a probability sampling design. Let s be the index set of the sample and $\pi_i = Pr\,(i \in s)$, the first-order inclusion probabilities of unit i, are known for all units in the population. Let $d_i = \pi_i^{-1}$ be the design weight of unit i in the sample. A design-consistent estimator of θ_0 can be obtained by solving the following estimating equation

$$N^{-1} \sum_{i \in s} d_i U\left(x_i, y_i; \theta\right) = 0, \tag{4.2}$$

for θ. Binder (1983) discussed estimators defined as the solution to the estimating equation (4.2).

If we know information on x, for example, the population mean $\bar{X}_N$, then we can incorporate the auxiliary information into estimation to improve the efficiency of the resulting estimator of θ_0. One way to achieve this efficiency is through calibration. That is, instead of solving (4.2), consider solving

$$\sum_{i \in s} d_i \omega_i U\left(x_i, y_i; \theta\right) = 0, \tag{4.3}$$

where ω_i is determined to minimize $\sum_{i \in s} d_i \left(\omega_i - 1\right)^2$ subject to the calibration constraint applied to $(1, x_i')'$:

$$\sum_{i \in s} d_i \omega_i (1, x_i')' = (1, \bar{X}_N')'. \tag{4.4}$$

For the special case of $\theta_0 = \bar{Y}_N = N^{-1} \sum_{i=1}^{N} y_i$, Deville and Särndal (1992) discussed the choice of objective functions that lead to calibration estimators asymptotically equivalent to the generalized regression (GREG) estimator:

$$\hat{\theta}_{GREG} = \bar{y}_d - \hat{B}\left(\bar{x}_d - \bar{X}_N\right), \tag{4.5}$$

where

$$\left(\bar{x}_d', \bar{y}_d\right)' = \left(\sum_{i \in s} d_i\right)^{-1} \sum_{i \in s} d_i \left(x_i', y_i\right)',$$

and

$$\hat{B} = \sum_{i \in s} d_i y_i \left(x_i - \bar{x}_d\right)' \left\{\sum_{i \in s} d_i \left(x_i - \bar{x}_d\right)\left(x_i - \bar{x}_d\right)'\right\}^{-1}.$$

Chen and Sitter (1999) considered using the pseudo empirical likelihood function

$$l_p(\omega) = \sum_{i \in s} d_i \log(\omega_i), \tag{4.6}$$

as an objective function for the calibration estimation with constraints (4.4). The resulting pseudo empirical likelihood calibration estimator for $\theta_0 = \bar{Y}_N$ is asymptotically equivalent to the GREG estimator in (4.5). The GREG estimator has certain optimal properties under the model where the finite population is a realization of a linear regression model

$$y_i = x_i'\beta + e_i, \tag{4.7}$$

with $E_\zeta(e_i) = 0$ and $V_\zeta(e_i) = \sigma^2$, where E_ζ and V_ζ are the expectation and variance under the super-population model. If the linear regression model (4.7) does not hold, then the GREG estimator is no longer optimal.

The design optimal regression estimator that minimizes the design variance among the linear class of the following form

$$\hat{\theta} = \bar{y}_{HT} - B(\bar{q}_{HT} - \bar{q}_N)$$

can be obtained by

$$\hat{\theta}_{opt} = \bar{y}_{HT} - \hat{B}_{opt}(\bar{q}_{HT} - \bar{q}_N), \tag{4.8}$$

where

$$(\bar{q}_{HT}', \bar{y}_{HT})' = (N^{-1}\sum_{i \in s} d_i q_i', N^{-1}\sum_{i \in s} d_i y_i)',$$

$q_i = (1, x_i')'$ and $\hat{B}_{opt}$ is a consistent estimator of $B_{opt} = Cov(\bar{y}_{HT}, \bar{q}_{HT})\{Var(\bar{q}_{HT})\}^{-1}$. The design optimal regression estimator has been discussed by Fuller and Isaki (1981), Montanari (1987), and Rao (1994).

In this paper, we consider an empirical-likelihood-type estimator that leads to the solution asymptotically equivalent to the design optimal regression estimator in (4.8). To achieve this goal, instead of assigning weights only for the sample, we propose using the population-level log-likelihood

$$l = \sum_{i=1}^{N} \log(\omega_i), \tag{4.9}$$

where $\sum_{i=1}^{N} \omega_i = 1$, as the objective function for the calibration estimation. Because the final estimator is obtained by solving (4.3) for θ, the final weights $d_i\omega_i$ in (4.3) are used to compute the design optimal estimator from the sample observation. Unlike the pseudo empirical likelihood, the proposed likelihood (4.9), called the population empirical likelihood (POEL), is defined at the population level. To incorporate the auxiliary information into the estimation, we use

$$\sum_{i \in s} d_i \omega_i (1, x_i')' = (1, \bar{X}_N')',$$

which are the same constraints in (4.4). For rejective Poisson sampling, in order to remove the effect of sampling design, we can incorporate additional constraints in the sampling design, which will be discussed in Section 4.

There are several advantages of the proposed method. First, it naturally incorporates additional information. For example, if $\bar{h}_N = N^{-1} \sum_{i=1}^{N} h(x_i)$ is known, where $h(x)$ is an arbitrary function of x, then we can add the constraint

$$\sum_{i \in s} d_i \omega_i h(x_i) = \bar{h}_N,$$

into the optimization using the POEL. Thus, it is directly applicable in the calibration problem of survey sampling. Secondly, given the constraints, it achieves the lower bound for the asymptotic design variance under some sampling designs. That is, for example, if $\theta_0 = \bar{Y}_N$ and $h(x) = (1, x')'$, we show that the proposed estimator is asymptotically equal to design optimal regression estimator (4.8) when the sampling rate is negligible. In addition, under some regularity conditions, the POEL enables us to obtain the likelihood ratio confidence intervals using chi-square quantiles. Furthermore, we can combine all sources of information from several surveys by using a single POEL to obtain the optimal estimator, which will be discussed in Section 5.

4.3 Main results

We first consider a Poisson sampling setup where independent Bernoulli trials are used to select the sample. Let I_i be the sample selection indicator that takes the value one if unit i

is selected in the sample and takes the value zero otherwise. In the Poisson sampling, I_i are independent Bernoulli (π_i) random variables, where π_i are known.

Under Poisson sampling, the proposed POEL approach discussed in Section 2 can be formulated as maximizing

$$l = \sum_{i=1}^{N} \log(\omega_i), \tag{4.10}$$

subject to

$$\sum_{i=1}^{N} \omega_i = 1, \quad \sum_{i=1}^{N} \omega_i \frac{I_i}{\pi_i} U_i(\theta) = 0. \tag{4.11}$$

Thus, without extra information, we get $\omega_i = N^{-1}$ and the POEL estimator $\hat{\theta}_{POEL}$ is the same as that obtained from the solution of (4.2). In order to incorporate the known population size information, we add the following constraint

$$\sum_{i=1}^{N} \omega_i (\frac{I_i}{\pi_i} - 1) = 0. \tag{4.12}$$

Note that in constraints (4.11) and (4.12), the observed values of x_i in the units with $I_i = 0$ are not used. To incorporate the auxiliary information associated with non-sampled part of x_i, we can impose

$$\sum_{i=1}^{N} \omega_i \frac{I_i}{\pi_i} h_i = \bar{h}_N, \tag{4.13}$$

for some function $h_i = h(x_i)$, where $\bar{h}_N = N^{-1} \sum_{i=1}^{N} h(x_i)$. By (4.11) and (4.12), constraint (4.13) can be written as

$$\sum_{i=1}^{N} \omega_i \frac{I_i}{\pi_i} \left(h_i - \bar{h}_N\right) = 0. \tag{4.14}$$

To solve for the above optimization problem, by the Lagrange multiplier method, the following two-step algorithm can be used. In the first step, the optimal weight that maximizes (4.10) subject to $\sum_{i=1}^{N} \omega_i = 1$, (4.12) and (4.14) can be expressed as

$$\hat{\omega}_i = \frac{1}{N} \frac{1}{1 + \hat{\lambda}' g_i},$$

where $g_i = ((I_i \pi_i^{-1} - 1), I_i \pi_i^{-1} (h_i - \bar{h}_N)')'$ and $\hat{\lambda}$ is the solution to

$$\frac{1}{N} \sum_{i=1}^{N} \frac{g_i}{1 + \hat{\lambda}' g_i} = 0.$$

In the second step, we can get the resulting POEL estimator $\hat{\theta}_{POEL}$ by solving

$$\sum_{i=1}^{N} \hat{\omega}_i \frac{I_i}{\pi_i} U_i(\theta) = 0. \tag{4.15}$$

Because the control function h_i in (4.13) does not depend on θ, the POEL estimator was obtained by the two-step algorithm above. Such two-step algorithm was discussed in Chaudhuri, Handcock and Rendall (2008). If the control function h_i depends on unknown parameter θ, say $h_i = h(x_i;\theta)$, then the optimization is computationally more challenging. In this case, using $\hat{h}_i = h(x_i;\hat{\theta})$, where $\hat{\theta}$ is any $\sqrt{n}$-consistent estimator of θ, in (4.13) leads to the same two-step algorithm for optimization and the two-step solution is asymptotically equivalent to the original solution.

To discuss the asymptotic properties, we first assume a sequence of finite populations and the samples satisfying the following regularity conditions:

(C1) Parameter $\theta_0 \in \Theta$ is the unique solution to $N^{-1}\sum_{i=1}^{N} U(X_i, Y_i;\theta_0) = 0$, Θ is a compact set in p-dimensional Euclidean space, and $U(X, Y;\theta)$ is uniformly continuous in Θ.

(C2) The partial derivative $\dot{U}(\theta) = \partial U(X, Y;\theta)/\partial\theta$ is a continuous function of θ in the neighborhood of θ_0 almost everywhere. Also, $\partial U(\theta_0)/\partial\theta$ is nonsingular.

(C3) Writing $g_i = \left(I_i\pi_i^{-1} - 1, I_i\pi_i^{-1}(h_i - \bar{h}_N)'\right)'$, as $n_B \to \infty$,

$$n_B^{1/2}(N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}U_i'(\theta_0), N^{-1}\sum_{i=1}^{N} g_i')' \to^d N(0, V),$$

where $n_B = E(n)$ and V is a positive definite matrix.

(C4) $||\partial U(x, y;\theta)/\partial\theta||$, $||U(x, y;\theta)||^4$ and $||h(x)||^4$ are bounded by $K(x, y)$ in Θ and

$$\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} K(x_i, y_i) = \mu_K,$$

where $\mu_K > 0$.

(C5) $\max_{i\in s} ||h_i|| = o_p(n_B^{1/2})$ and $\max_{i\in s} ||U_i(\theta_0)|| = o_p(n_B^{1/2})$.

(C6) $C_1 < \pi_i N n_B^{-1} < C_2, \quad i = 1, 2, \ldots, N$ for some constants $0 < C_1 < C_2$.

Condition (C1) and (C2) ensure the identifiability of parameter θ_0 and the smoothness properties of function $U(\theta)$. Condition (C3) ensures the asymptotic normality of Horvitz-Thompson type estimator under Poisson sampling. Theorem 1.3.3 of Fuller (2009b) provides sufficient conditions for (C3). Condition (C4) is the usual moment condition in survey sampling. Condition in (C5) is one of the typical conditions to enable $\hat{\lambda} = O_p(n_B^{-1/2})$ and Taylor expansion. Condition (C6) controls the behavior of the first order inclusion probabilities.

The following theorem presents some asymptotic properties of the POEL estimator $\hat{\theta}_{POEL}$ defined in (4.15).

Theorem 4.3.1 *Under the regularity conditions (C1)-(C6) described as above, the population empirical likelihood (POEL) estimator $\hat{\theta}_{POEL}$ which we obtained in (4.15) has the following asymptotic expansion*

$$\hat{\theta}_{POEL} - \theta_0 = -\tau \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} U_i(\theta_0) - B^*\left(\frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} \eta_i - \bar{\eta}_N\right) \right\} + o_p(n_B^{-1/2}), \tag{4.16}$$

where $\tau = \left[N^{-1} \sum_{i=1}^{N} \partial U_i(\theta_0)/\partial\theta\right]^{-1}$, $\eta = (1, (h - \bar{h}_N)')'$, $h = h(x)$, and $B^* = \Omega_1 \Omega_2^{-1}$, where

$$\Omega_1 = \left(\frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1\right) U_i, \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{\pi_i} U_i (h_i - \bar{h}_N)'\right) \tag{4.17}$$

and

$$\Omega_2 = \begin{pmatrix} N^{-2} \sum_{i=1}^{N} (\pi_i^{-1} - 1) & N^{-2} \sum_{i=1}^{N} (\pi_i^{-1} - 1)(h_i - \bar{h}_N)' \\ N^{-2} \sum_{i=1}^{N} (\pi_i^{-1} - 1)(h_i - \bar{h}_N) & N^{-2} \sum_{i=1}^{N} \pi_i^{-1} (h_i - \bar{h}_N)^{\otimes 2} \end{pmatrix}. \tag{4.18}$$

with $X^{\otimes 2} = XX'$. Hence, we have

$$V_h^{-1/2} \left(\hat{\theta}_{POEL} - \theta_0\right) \to^d N(0, I), \tag{4.19}$$

where $\to^d$ denotes the convergence in distribution, $V_h = \tau \Omega_h \tau'$ with

$$\Omega_h = N^{-2} V \left\{ \sum_{i=1}^{N} \frac{I_i}{\pi_i} U_i - B^*\left(\sum_{i=1}^{N} \frac{I_i}{\pi_i} \eta_i - \sum_{i=1}^{N} \eta_i\right) \right\}.$$

Remark 4.3.1 *For $\theta_0 = \bar{Y}_N$, $h = x$, and $U = y - \theta$, (4.16) becomes*

$$\begin{aligned} \hat{\theta}_{POEL} &= \bar{Y}_N + \frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} (y_i - \bar{Y}_N) - B_1^*\left(\frac{\hat{N}}{N} - 1\right) \\ &- B_2^* \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} (x_i - \bar{X}_N) \right\} + o_p(n_B^{-1/2}), \end{aligned} \tag{4.20}$$

where $\hat{N} = \sum_{i=1}^{N} I_i\pi_i^{-1}$, $(B_1^*, B_2^*) = \Omega_1\Omega_2^{-1}$ *with* Ω_1 *and* Ω_2 *defined in (4.17) and (4.18), respectively. If* $n_B/N \to 0$, *under Poisson sampling,*

$$\Omega_1 = Cov(N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}U_i, N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}q_i) + o_p(n_B^{-1})$$

and $\Omega_2 = Var(N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}q_i) + o_p(n_B^{-1})$ *with* $q_i = (1, (x_i - \bar{X}_N)')'$. *Notice that* $\hat{\theta}_{POEL}$ *is obtained by minimizing the first order asymptotic variance of the estimators in the class of (4.20). Thus, it is asymptotically equivalent to the optimal estimator (4.8).*

In Theorem 4.3.1, the sampling design is not necessarily the Poisson sampling. However, the optimality result in Remark 4.3.1 is established under Poisson sampling. By Theorem 4.3.1, the consistent estimator of V_h can be written as $\hat{V}_h = \hat{\tau}\hat{\Omega}_h\hat{\tau}'$ where $\hat{\tau} = \left\{N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}\partial U_i(\hat{\theta})/\partial\theta\right\}^{-1}$, $\hat{\Omega}_h = N^{-2}\sum_{i=1}^{N} I_i(1-\pi_i)\pi_i^{-2}\hat{r}_i^{\otimes 2}$, $\hat{r}_i = U_i(\hat{\theta}) - \hat{B}^*\eta_i$, $\hat{B}^* = \hat{\Omega}_1\hat{\Omega}_2^{-1}$,

$$\hat{\Omega}_1 = \Big(\frac{1}{N^2}\sum_{i=1}^{N} I_i(1-\pi_i)\pi_i^{-2}U_i(\hat{\theta}), \frac{1}{N^2}\sum_{i=1}^{N} I_i\pi_i^{-2}U_i(\hat{\theta})(h_i - \bar{h}_N)'\Big)$$

and

$$\hat{\Omega}_2 = \begin{pmatrix} N^{-2}\sum_{i=1}^{N} I_i(1-\pi_i)\pi_i^{-2} & N^{-2}\sum_{i=1}^{N} I_i(1-\pi_i)\pi_i^{-2}(h_i - \bar{h}_N)' \\ N^{-2}\sum_{i=1}^{N} I_i(1-\pi_i)\pi_i^{-2}(h_i - \bar{h}_N) & N^{-2}\sum_{i=1}^{N} I_i\pi_i^{-2}(h_i - \bar{h}_N)^{\otimes 2} \end{pmatrix},$$

with $\hat{\theta} = \hat{\theta}_{POEL}$.

By Theorem 4.3.1, we can construct a Wald-type confidence interval for θ_0 using the asymptotic normality. The following theorem presents a limiting distribution of the likelihood ratio statistics using the population empirical likelihood.

Theorem 4.3.2 *Under the same assumptions of Theorem 4.3.1, let* $R_n(\theta_0) = 2\left\{l(\hat{\theta}_{POEL}) - l(\theta_0)\right\}$ *where* $l(\theta) = \sum_{i=1}^{N}\log(\omega_i)$ *with* ω_i *satisfying (4.11), (4.12) and (4.14). Then, as* $n_B \to \infty$ *and* $n_B/N \to 0$,

$$R_n(\theta_0) \to^d \chi_p^2,$$

where p *is the dimension of* θ.

According to Theorem 4.3.2, under some regularity conditions, a Wilk-type confidence interval for θ_0 can be constructed with a chi-square distribution as the limiting distribution when the sampling rate n_B/N is negligible.

Instead of population empirical likelihood, if one uses the pseudo empirical likelihood (4.6), the limiting distribution of the likelihood ratio statistic is a scaled chi-square distribution. Wu and Rao (2006) proposed using an adjustment based on the design effect to construct likelihood ratio confidence intervals from the pseudo empirical likelihood. However, the design effect is usually unknown and has to be estimated for each parameter. The proposed likelihood ratio confidence interval based on the population empirical likelihood does not require such extra computation.

The variance of the POEL estimator depends on the choice of the control function h_i in constraint (4.13). Discussion for the optimal choice of h_i requires some superpopulation model for the conditional distribution of y_i on x_i. Because the mode of inference is purely design-based in our paper, we do not pursue this topic here.

Remark 4.3.2 *Instead of maximizing the population likelihood (4.10), we can consider maximizing the sampled part of the population likelihood subject to the same constraints with* $\sum_{i=1}^N \omega_i = 1$ *replaced by* $\sum_{i\in s} \omega_i = 1$*. That is, the sample empirical likelihood (SEL) estimator* $\hat{\theta}_{SEL}$ *can be obtained by maximizing*

$$l_e = \sum_{i\in s} \log(\omega_i),$$

subject to

$$\sum_{i\in s} \omega_i = 1, \quad \sum_{i\in s} \omega_i \pi_i^{-1} U_i(\theta) = 0, \tag{4.21}$$

and

$$\sum_{i\in s} \omega_i \pi_i^{-1} (h_i - \bar{h}_N) = 0. \tag{4.22}$$

Note that the resulting SEL estimator is algebraically equivalent to the nonparametric likelihood estimator proposed by Kim (2009). Furthermore, under certain conditions, it can be shown that

$$R_n(\theta_0) = 2\left\{ l_e(\hat{\theta}_{SEL}) - l_e(\theta_0) \right\} \to^d \chi_1^2.$$

For $\theta = E(Y)$, if $n/N \to 0$, the SEL estimator $\hat{\theta}_{SEL}$ with $h = x$ is asymptotically equivalent to the optimal estimator

$$\hat{\theta}_{opt1} = \bar{y}_d - \hat{B}(\bar{x}_d - \bar{X}_N), \tag{4.23}$$

where

$$(\bar{x}_d^T, \bar{y}_d) = (\sum_{i \in s} \pi_i^{-1} x_i^T, \sum_{i \in s} \pi_i^{-1} y_i) / \sum_{i \in s} \pi_i^{-1},$$

$\hat{B} = \hat{C}(\bar{y}_d, \bar{x}_d)\{\hat{V}(\bar{x}_d)\}^{-1}$, and $\hat{C}(\bar{y}_d, \bar{x}_d)$, $\hat{V}(\bar{x}_d)$ are design consistent estimator of $Cov(\bar{y}_d, \bar{x}_d)$, $Var(\bar{x}_d)$, respectively. Comparing (4.23) with (4.8), the POEL estimator is more efficient than the SEL estimator. Furthermore, it is easier to combine information from multiple surveys for the POEL.

4.4 Extension to rejective Poisson sampling

We now extend the results in Section 3 to other sampling designs. In particular, we consider the rejective Poisson sampling, which covers the simple random sampling and the stratified random sampling as a special case. Rejective Poisson sampling has been studied by Hájek (1964), Hájek (1981) and Fuller (2009a). Hájek (1964) considered the linear design constraint as below

$$\sum_{i=1}^{N} \frac{\delta_i}{p_i} z_i = \sum_{i=1}^{N} z_i, \tag{4.24}$$

with $z_i = p_i$ and $\sum_{i=1}^{N} p_i = n$, where p_i and δ_i are the inclusion probabilities and sampling indicators for the initial sampling design, respectively. Fuller (2009a) considered a rejective sampling with constraints

$$\hat{Q}_{p,n} = (\bar{z}_p - \bar{Z}_N)' V_{\bar{z}\bar{z}}^{-1} (\bar{z}_p - \bar{Z}_N) < \gamma^2, \tag{4.25}$$

for some $\gamma^2 > 0$, where $\bar{z}_p = N^{-1} \sum_{i=1}^{N} \delta_i p_i^{-1} z_i$ and $V_{\bar{z}\bar{z}} = V_{poi}(\bar{z}_p)$, V_{poi} denotes the variance calculated under Poisson sampling design. Since constraint (4.24) is a special case of constraint (4.25), then we will consider constraint (4.25) only. We consider the following rejective Poisson sampling procedure

[Step 1] For $i = 1, \cdots, N$, generate $\delta_i \sim Bernoulli(p_i)$ independently.

[Step 2] Check if (4.25) holds. If it does not hold, then go to [Step 1]. If the constraint is satisfied, then set $(I_1, \cdots, I_N) = (\delta_1, \cdots, \delta_N)$. The final sample consists of elements with $I_i = 1$.

Thus, even if δ_i are generated independently, the realized sampling indicators $I_1, \cdots, I_N$ are no longer independent. The initial selection probabilities $p_i (i = 1, 2, \ldots, N)$ for Poisson sampling are not exactly equal to the target inclusion probabilities $\pi_i (i = 1, 2, \ldots, N)$. The POEL estimator can be obtained by maximizing (4.10) subject to

$$\sum_{i=1}^{N} \omega_i = 1, \qquad \sum_{i=1}^{N} \omega_i (\frac{I_i}{p_i} - 1) = 0, \quad \sum_{i=1}^{N} \omega_i (\frac{I_i}{p_i} - 1) z_i = 0, \tag{4.26}$$

and

$$\sum_{i=1}^{N} \omega_i \frac{I_i}{p_i} (h_i - \bar{h}_N) = 0, \quad \sum_{i=1}^{N} \omega_i \frac{I_i}{p_i} U_i(\theta) = 0. \tag{4.27}$$

In (4.26), constraint $\sum_{i=1}^{N} \omega_i (I_i p_i^{-1} - 1) z_i = 0$ is added to account for the design constraint in (4.25). Suppose the regularity conditions (C1)-(C3) and (C5) in Section 3 hold with π_i, g_i replaced by p_i, g_i^*, respectively, where $g_i^* = ((I_i p_i^{-1} - 1) z_i^*, I_i p_i^{-1} (h_i' - \bar{h}_N'))'$, $z_i^* = (1, z_i')'$. Let $G_N(\gamma^2) = Pr(\hat{Q}_{p,n} \leq \gamma^2), G_{N(i)}(\gamma^2) = Pr(\hat{Q}_{p,n} \leq \gamma^2 | i \in s), G_{N(ij)}(\gamma^2) = Pr(\hat{Q}_{p,n} \leq \gamma^2 | i, j \in s)$. We also assume the following conditions:

(C7) $|n_B N^{-1} p_i^{-1}|, z_i$ are bounded.

(C8) $||\partial M(\theta)/\partial\theta||, ||M(\theta)||^4$ are bounded by $K(x, y)$ in Θ and $\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} K(x_i, y_i) = \mu_K$, for some $\mu_K > 0$, where $M_i(\theta) = (U_i'(\theta), z_i^{*'}, h_i' - \bar{h}_N')'$.

(C9) Suppose $G_{N(i)}(\gamma^2) = G_N(\gamma^2) + g_{1N}(\gamma^2)\gamma^2\eta_i + o_p(n_B^{-1})$, where $\eta_i = n_B N^{-2}(1 - p_i) p_i^{-1} z_i^2$ and $g_{1N}(\gamma^2)$ is a bounded sequence.

(C10) Suppose $G_{N(ij)}(\gamma^2) = G_N(\gamma^2) + g_{1N}(\gamma^2)\gamma^2(\eta_i + \eta_j) + o_p(n_B^{-1})$.

(C11) Suppose $V_{\bar{z}\bar{z}}^{-1/2}(\bar{z}_p - \bar{Z}_N) \to^d N(0, I)$.

Condition (C7) is used to control the behavior of the first order inclusion probabilities and boundness of z_i. Condition (C8) is the usual moment condition in survey sampling. Condition (C9) and (C10) are similar to Assumption 8 in Fuller (2009a). Condition (C11) will hold for Poisson sampling under the moment conditions specified in Theorem 1.3.3 of Fuller (2009b). To discuss some motivation of assumptions (C9) and (C10), without loss of generality, first assume $\bar{z}_N = 0$ and $n_B N^{-2} \sum_{i=1}^{N} (1 - p_i) p_i^{-1} z_i^2 = 1$. After some algebra, we can obtain

$$E(\bar{z}_p - \bar{Z}_N | i \in s) = \frac{1}{N} \frac{1 - p_i}{p_i} z_i, \tag{4.28}$$

$$E(\bar{z}_p - \bar{Z}_N | i, j \in s) = \frac{1}{N}\frac{1-p_i}{p_i} z_i + \frac{1}{N}\frac{1-p_j}{p_j} z_j, \tag{4.29}$$

$$Var(\bar{z}_p - \bar{Z}_N | i \in s) = n_B^{-1} - \frac{1}{N^2}\frac{1-p_i}{p_i} z_i^2, \tag{4.30}$$

and

$$Var(\bar{z}_p - \bar{Z}_N | i, j \in s) = n_B^{-1} - \frac{1}{N^2}\frac{1-p_i}{p_i} z_i^2 - \frac{1}{N^2}\frac{1-p_j}{p_j} z_j^2. \tag{4.31}$$

According to assumption (C11), G_N is the CDF of the Chi-square distribution and $G_{N(i)}, G_{N(ij)}$ are the CDF of the noncentralized Chi-square distribution. According to (4.28)-(4.31), we have

$$E(\hat{Q}_{p,n} | i \in s) = 1 - \eta_i + o_p(n_B^{-1}), \quad E(\hat{Q}_{p,n} | i, j \in s) = 1 - \eta_i - \eta_j + o_p(n_B^{-1}),$$

where $\eta_i = n_B N^{-2}(1-p_i)p_i^{-1} z_i^2$. So, we can express

$$\begin{aligned} G_{N(i)}(\gamma^2) &= Pr(\hat{Q}_{p,n} \le \gamma^2 | i \in s) = Pr\left\{(1-\eta_i)^{-1}\hat{Q}_{p,n} \le (1-\eta_i)^{-1}\gamma^2 | i \in s\right\} \\ &= G_N\left\{(1-\eta_i)^{-1}\gamma^2\right\} + o_p(n_B^{-1}) \\ &= G_N((1+\eta_i)\gamma^2) + o_p(n_B^{-1}) \\ &= G_N(\gamma^2) + g_{1N}(\gamma^2)\gamma^2\eta_i + o_p(n_B^{-1}), \end{aligned}$$

where g_{1N} is the density of the Chi-square distribution. Similarly, we can obtain

$$G_{N(ij)}(\gamma^2) = G_N(\gamma^2) + g_{1N}(\gamma^2)\gamma^2(\eta_i + \eta_j) + o_p(n_B^{-1}).$$

We now provide the following asymptotic results for the proposed POEL estimator under rejective Poisson sampling.

Theorem 4.4.1 *Assume a rejective Poisson sampling with the design constraint in (4.25). Let $\hat{\theta}_{POEL}$ be the population empirical likelihood estimator obtained by maximizing (4.10) subject to constraints (4.26) and (4.27). Under regularity conditions (C1)-(C3), (C5) and (C6) and (C7)-(C11) above, $\hat{\theta}_{POEL}$ has the following asymptotic expansion*

$$\hat{\theta}_{POEL} - \theta_0 = -\tau\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}U_i(\theta_0) - B\left(\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\eta_i - \bar{\eta}_N\right)\right\} + o_p(n_B^{-1/2}), \tag{4.32}$$

where $\tau = \left\{N^{-1}\sum_{i=1}^{N}\partial U_i(\theta_0)/\partial\theta\right\}^{-1}$, $\eta_i = (z_i^{*'}, (h_i - \bar{h}_N)')'$, $h_i = h(x_i)$, $z_i^* = (1, z_i')'$, $B = \Omega_1\Omega_2^{-1}$, *where*

$$\Omega_1 = \left(\frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{1}{\pi_i} - 1\right)U_i z_i^{*'}, \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{\pi_i}U_i(h_i - \bar{h}_N)'\right) \tag{4.33}$$

and

$$\Omega_2 = \left(\begin{array}{cc} N^{-2}\sum_{i=1}^N(\pi_i^{-1}-1)z_i^{*\otimes 2} & N^{-2}\sum_{i=1}^N(\pi_i^{-1}-1)z_i^*(h_i-\bar{h}_N)' \\ N^{-2}\sum_{i=1}^N(\pi_i^{-1}-1)(h_i-\bar{h}_N)z_i^{*\prime} & N^{-2}\sum_{i=1}^N \pi_i^{-1}(h_i-\bar{h}_N)^{\otimes 2} \end{array} \right). \tag{4.34}$$

Hence, we have

$$V_h^{-1/2}\left(\hat{\theta}_{POEL}-\theta_0\right) \to^d N(0,I), \tag{4.35}$$

where $V_h = \tau\Omega_h\tau'$ *with*

$$\begin{aligned} \Omega_h &= N^{-2}V\left\{\sum_{i=1}^N \frac{I_i}{\pi_i}U_i - B(\sum_{i=1}^N \frac{I_i}{\pi_i}\eta_i - \sum_{i=1}^N \eta_i)\right\} \\ &= V_{poi}(\hat{\bar{e}}_p), \end{aligned}$$

and V_{poi} *denotes the variance under Poisson sampling design and* $\hat{\bar{e}}_p = N^{-1}\sum_{i=1}^N I_i p_i^{-1} e_i$ *with* $e_i = U_i - B\eta_i$.

Remark 4.4.1 *For* $\hat{\theta}_0 = \bar{Y}_N$ *and* $h = x$, *(4.32) simplifies to*

$$\begin{aligned} \hat{\theta}_{POEL} &= \bar{Y}_N + \frac{1}{N}\sum_{i=1}^N \frac{I_i}{\pi_i}(y_i - \bar{Y}_N) - B_1(\frac{1}{N}\sum_{i=1}^N \frac{I_i}{\pi_i} - 1)z_i^* \\ &- B_2\{\frac{1}{N}\sum_{i=1}^N \frac{I_i}{\pi_i}(x_i - \bar{X}_N)\} + o_p(n_B^{-1/2}), \end{aligned}$$

where $(B_1, B_2) = \Omega_1\Omega_2^{-1}$ *with* Ω_1 *and* Ω_2 *defined in (4.33) and (4.34), respectively. If we choose* $\gamma = o(1)$ *in (4.25), then*

$$\bar{z}_{HT} - \bar{Z}_N = o_p(n_B^{-1/2}), \tag{4.36}$$

with $\bar{z}_{HT} = N^{-1}\sum_{i=1}^N I_i\pi_i^{-1}z_i$. *When* $n_B/N \to 0$, *by (4.36), we have*

$$\begin{aligned} \hat{\theta}_{POEL} &= \bar{Y}_N + \frac{1}{N}\sum_{i=1}^N \frac{I_i}{\pi_i}(y_i - \bar{Y}_N) - B_1^*(\frac{1}{N}\sum_{i=1}^N \frac{I_i}{\pi_i} - 1) \\ &- B_2^*\{\frac{1}{N}\sum_{i=1}^N \frac{I_i}{\pi_i}(x_i - \bar{X}_N)\} + o_p(n_B^{-1/2}), \end{aligned}$$

where $z_i^* = (1, z_i')'$, $(B_1^*, B_2^*) = \Omega_1^*\Omega_2^{*-1}$,

$$\begin{aligned} \Omega_1^* &= N^{-2}\sum_{i=1}^N(1-\pi_i)\pi_i^{-1}(y_i - \bar{Y}_N)q_i' \\ &- \{N^{-2}\sum_{i=1}^N(1-\pi_i)\pi_i^{-1}(y_i-\bar{Y}_N)z_i'\}\{\sum_{i=1}^N(1-\pi_i)\pi_i^{-1}z_iz_i'\}^{-1}\{\sum_{i=1}^N(1-\pi_i)\pi_i^{-1}z_iq_i'\} \end{aligned}$$

and

$$\begin{aligned}\Omega_2^* &= N^{-2}\sum_{i=1}^{N}(1-\pi_i)\pi_i^{-1}q_iq_i' \\ &- \{N^{-2}\sum_{i=1}^{N}(1-\pi_i)\pi_i^{-1}q_iz_i'\}\{\sum_{i=1}^{N}(1-\pi_i)\pi_i^{-1}z_iz_i'\}^{-1}\{\sum_{i=1}^{N}(1-\pi_i)z_iq_i'\},\end{aligned}$$

with $q_i = (1,(x_i-\bar{X}_N)')'$. *Under some regularity conditions, it can be shown that*

$$\Omega_1^* = Cov(N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}U_i, N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}q_i) + o_p(n_B^{-1})$$

and

$$\Omega_2^* = Var(N^{-1}\sum_{i=1}^{N} I_i\pi_i^{-1}q_i) + o_p(n_B^{-1}).$$

Thus, by using a similar argument as in Remark 4.3.1, we have

$$\hat{\theta}_{POEL} = \hat{\theta}_{opt} + o_p(n_B^{-1/2}),$$

and $\hat{\theta}_{opt}$ *is defined in (4.8).*

A consistent variance estimator of $\hat{\theta}_{POEL}$ can be constructed by $\hat{V}_h = \hat{\tau}\hat{\Omega}_h\hat{\tau}'$,

$$\hat{\tau} = \left\{N^{-1}\sum_{i\in s} p_i^{-1}\partial U_i(\hat{\theta})/\partial\theta\right\}^{-1}, \quad \hat{\Omega}_h = N^{-2}\sum_{i=1}^{N} I_i(1-p_i)p_i^{-2}\hat{r}_i^{\otimes 2},$$

where $\hat{r}_i = U_i(\hat{\theta}) - \hat{B}^*\eta_i$, $\hat{B}^* = \hat{\Omega}_1\hat{\Omega}_2^{-1}$, and

$$\hat{\Omega}_1 = \Big(\frac{1}{N^2}\sum_{i=1}^{N} I_i(1-p_i)p_i^{-2}U_i(\hat{\theta})z_i^{*'}, \frac{1}{N^2}\sum_{i=1}^{N} I_ip_i^{-2}U_i(\hat{\theta})(h_i-\bar{h}_N)'\Big)$$

and

$$\hat{\Omega}_2 = \begin{pmatrix} N^{-2}\sum_{i=1}^{N} I_i(1-p_i)p_i^{-2}z_i^{*\otimes 2} & N^{-2}\sum_{i=1}^{N} I_i(1-p_i)p_i^{-2}z_i^{*}(h_i-\bar{h}_N)' \\ N^{-2}\sum_{i=1}^{N} I_i(1-p_i)p_i^{-2}(h_i-\bar{h}_N)z_i^{*'} & N^{-2}\sum_{i=1}^{N} I_ip_i^{-2}(h_i-\bar{h}_N)^{\otimes 2} \end{pmatrix}.$$

The following theorem presents a limiting distribution of the likelihood ratio statistics using the population empirical likelihood.

Theorem 4.4.2 *Assume that the sample is obtained by the rejective Poisson sampling design and assume the regularity conditions in Theorem 4.4.1. Let* $R_n(\theta_0) = 2\left\{l(\hat{\theta}_{POEL}) - l(\theta_0)\right\}$ *where* $l(\theta) = \sum_{i=1}^{N}\log(\omega_i)$ *with* ω_i *satisfying (4.26) and (4.27). Then, as* $n_B \to \infty$, *and* $n_B/N \to 0$

$$R_n(\theta_0) \to^d \chi_p^2,$$

where p *is the dimension of* θ.

4.5 Combining information from two independent surveys

Consider two independent surveys, survey 1 and survey 2, from the same finite population, and the auxiliary variable x_i is observed in common in both surveys. In addition, we observe (z_{1i}, z_{2i}) throughout the population, where z_{1i} is observed in the survey 1 sample and z_{2i} is observed in the survey 2 sample. Thus, we observe (z_{i1}, x_i) from the survey 1 sample and observe (z_{i2}, x_i, y_i) from the survey 2 sample. Suppose that an intercept is included in z_{1i} and z_{2i}. This type of sampling design is often called non-nested two-phase sampling design (Hidiroglou, 2001). Zieschang (1990), Renssen and Nieuwenbroek (1997), and Merkouris (2004) considered using GREG-type estimators to combine information from different surveys. Wu (2004) considered the pseudo empirical likelihood method to solve such problems and showed that the pseudo empirical likelihood estimator is asymptotically equivalent to the GREG estimator.

We propose using the population empirical likelihood method in Section 3 to combine information from non-nested two-phase sampling. The proposed population-level empirical likelihood method is different from the sample-level empirical likelihood method of Wu (2004) in that we use all the available information and the proposed estimator is optimal. In addition, under some regularity conditions, we can construct likelihood ratio type confidence intervals with a chi-square limiting distribution. The proposed method can be easily extended to combining more than two surveys.

For simplicity, assume that the sampling designs in two surveys are independent Poisson sampling designs. We can easily extend our results to other sampling designs, like the rejective Poisson sampling, by using similar arguments as in Section 4. Let I_{1i} and I_{2i} be the sample selection indicators for survey 1 and survey 2, respectively, and let π_{1i} and π_{2i} be the first order inclusion probabilities for survey 1 and for survey 2, respectively.

We are interested in estimating the general parameter defined by (4.1). The proposed POEL procedure for combining two surveys can be formulated as maximizing

$$l = \sum_{i=1}^{N} \log(\omega_i),$$

subject to

$$\sum_{i=1}^{N} \omega_i = 1, \quad \sum_{i=1}^{N} \omega_i (I_{1i}\pi_{1i}^{-1} - 1) z_{1i} = 0, \quad \sum_{i=1}^{N} \omega_i (I_{2i}\pi_{2i}^{-1} - 1) z_{2i} = 0,$$

and

$$\sum_{i=1}^{N} \omega_i (I_{1i}\pi_{1i}^{-1} - I_{2i}\pi_{2i}^{-1}) h_i = 0, \quad \sum_{i=1}^{N} \omega_i I_{2i}\pi_{2i}^{-1} U_i(\theta) = 0.$$

Under the regularity conditions of Theorem 4.3.1 at each survey, if $n/N \to 0$, it can be shown that our proposed estimator $\hat{\theta}_{POEL}$ is asymptotically equivalent to the optimal estimator that minimizes

$$Q(\bar{h}_N, \theta) = \begin{pmatrix} \bar{z}_{HT,1} - \bar{Z}_1 \\ \bar{h}_{HT,1} - \bar{h}_N \\ \bar{h}_{HT,2} - \bar{h}_N \\ \bar{z}_{HT,2} - \bar{Z}_2 \\ \bar{U}_{HT,2}(\theta) \end{pmatrix}' V_Q^{-1} \begin{pmatrix} \bar{z}_{HT,1} - \bar{Z}_1 \\ \bar{h}_{HT,1} - \bar{h}_N \\ \bar{h}_{HT,2} - \bar{h}_N \\ \bar{z}_{HT,2} - \bar{Z}_2 \\ \bar{U}_{HT,2}(\theta) \end{pmatrix}, \tag{4.37}$$

with respect to $\bar{h}_N$ and θ, where $(\bar{z}_{HT,1}, \bar{h}_{HT,1}) = N^{-1}\sum_{i=1}^{N} I_{1i}\pi_{1i}^{-1}(z_{1i}, h_i)$, $(\bar{z}_{HT,2}, \bar{h}_{HT,2}) = N^{-1}\sum_{i=1}^{N} I_{2i}\pi_{2i}^{-1}(z_{2i}, h_i)$, $(\bar{Z}_1, \bar{Z}_2) = N^{-1}\sum_{i=1}^{N}(z_{1i}, z_{2i})$,

$$\bar{h}_N = N^{-1}\sum_{i=1}^{N} h_i, \quad \bar{U}_{HT,2}(\theta) = N^{-1}\sum_{i=1}^{N} I_{i2}\pi_{2i}^{-1} U_i(\theta),$$

and V_Q is the estimated variance-covariance matrix of $(\bar{z}_{HT,1}, \bar{h}_{HT,1}, \bar{h}_{HT,2}, \bar{z}_{HT,2}, \bar{U}_{HT,2}(\theta))$. The optimal estimator obtained by minimizing (4.37) is called the generalized method of moment (GMM) estimator (Hansen, 1982). The GMM estimator is a popular tool for combining information from several sources in the econometrics literature (Imbens and Lancaster, 1994; Hirano, Imbens, Ridder, and Rubin, 1998). Imbens (2002) showed the asymptotic equivalence between the empirical likelihood estimator and the GMM estimator under the single sample setup. To compute the GMM estimator from (4.37), we need to estimate the variance-covariance matrix. The empirical likelihood approach avoids the computation for the variance-covariance matrix.

For the special case of $\theta_0 = \bar{Y}_N$ and $h_i = x_i$, the optimal estimator of θ_0 minimizing (4.37) can be written as

$$\hat{\theta}_{opt} = \bar{y}_{HT,2} + \hat{B}_{1opt}(\bar{Z}_1 - \bar{z}_{HT,1}) + \hat{B}_{2opt}(\bar{Z}_2 - \bar{z}_{HT,2}) + \hat{B}_{3opt}(\bar{x}_{HT,1} - \bar{x}_{HT,2}),$$

where $\bar{y}_{HT,2} = N^{-1}\sum_{i=1}^{N} I_{2i}\pi_{2i}^{-1}y_i$,

$$\bar{x}_{HT,t} = N^{-1}\sum_{i=1}^{N} I_{ti}\pi_{ti}^{-1}x_i, \quad \bar{z}_{HT,t} = N^{-1}\sum_{i=1}^{N} I_{ti}\pi_{ti}^{-1}z_{ti}, \quad t = 1, 2,$$

and

$$\hat{B}_{opt} = (\hat{B}_{1opt}, \hat{B}_{1opt}, \hat{B}_{1opt}) = \hat{C}(\bar{y}_{HT,2}, \bar{S}_{HT})\{\hat{V}(\bar{S}_{HT})\}^{-1},$$

with $\bar{S}_{HT} = (\bar{z}'_{HT,1} - \bar{Z}'_1, \bar{z}'_{HT,2} - \bar{Z}'_2, \bar{x}'_{HT,2} - \bar{x}'_{HT,1})'$, $\hat{C}(\bar{y}_{HT,2}, \bar{S}_{HT})$ and $\hat{V}(\bar{S}_{HT})$ are consistent estimators of $Cov(\bar{y}_{HT,2}, \bar{S}_{HT})$ and $Var(\bar{S}_{HT})$, respectively. Also, under some regularity conditions for both surveys, we can get

$$2[l(\hat{\theta}_{POEL}) - l(\theta_0)] \rightarrow^{d} \chi_1^2$$

which will be very useful for constructing a Wilk-type confidence interval.

4.6 Simulation Study

4.6.1 Simulation One

To test our theory, we performed two limited simulation studies. The first simulation study can be described as a $2 \times 3 \times 4$ factorial design with three factors. The first factor is the model for generating the finite population. The second is the sampling design, and the third is the estimation method. Two finite populations of (x_i, y_i, z_i), population A and population B, with size $N = 10,000$ were generated. In population A, the population elements were generated by $z_i \sim \chi^2(2) + 1, \quad x_i = a_i + z_i, \quad y_i = 1 + 1.2(x_i - 3) + (x_i/4)e_i$, where $a_i \sim N(0, 1)$, independent of z_i, and $e_i \sim \chi^2(1) - 1$, independent of (a_i, z_i). In population B, (x_i, z_i) were the same as in population A and $y_i = 0.2(x_i - 1)^2 + (x_i/4)e_i$. From each population, $n = 200$ sample elements were selected repeatedly for $B = 2,000$ times. For the sampling design, three sampling designs were considered: simple random sampling (SRS) without replacement, Poisson sampling, and the rejective Poisson sampling. For the Poisson sampling, we used $\pi_i = nz_i/(\sum_{i=1}^{N} z_i)$. In the rejective Poisson sampling, the fixed-size constraint $\sum_{i=1}^{N} I_i = n$ was used with the initial sample selection probability $p_i = nz_i/(\sum_{i=1}^{N} z_i)$. The parameter of interest is the population mean of y. From each sample, the following six point estimators were computed.

1. Hájek (HJ) estimator: $\hat{\theta}_{HJ} = \sum_{i\in s} \pi_i^{-1} y_i / \sum_{i\in s} \pi_i^{-1}$.

2. Horvitz-Thompson (HT) estimator: $\hat{\theta}_{HT} = N^{-1} \sum_{i\in s} \pi_i^{-1} y_i$.

3. Proposed population-level empirical likelihood (POEL1) method without using x information. That is, it is obtained by maximizing $l = \sum_{i=1}^{N} \log(\omega_i)$ subject to (4.11), (4.12) and $\sum_{i=1}^{N} \omega_i (I_i - p_i) = 0$ (for SRSWOR and rejective Poisson sampling) with $U = y - \theta$.

4. Pseudo-empirical likelihood (PEL) method with constraint (4.4).

5. Proposed sample-level empirical likelihood (SEL) method in Remark 4.3.2 by using constraints (4.21), (4.22) and design constraint (for SRSWOR and rejective Poisson sampling) $\sum_{i\in s} \omega_i p_i^{-1} (p_i - \bar{p}_N) = 0$ with $U = y - \theta$ and $h = x$.

6. Proposed population-level empirical likelihood (POEL2) method by using constraints (4.11), (4.12), (4.13) and $\sum_{i=1}^{N} \omega_i (I_i - p_i) = 0$ (for SRSWOR and rejective Poisson sampling) with $U = y - \theta$ and $h = x$.

Thus, the first three estimators are computed without using x information while the next three estimators incorporate the population mean of x. Based on $B = 2,000$ Monte Carlo samples, we have computed the biases, variances, and mean squared errors of the six estimators. Table 4.1 presents the simulation results of the six point estimators. HJ estimator and HT estimator are identical under SRS, but HT estimator is more efficient than HJ estimator under other designs. POEL1 estimator has the same efficiency as HJ and HT estimators under SRS, but it performs better under other designs because it effectively uses the population size (N) information. The three empirical likelihood methods (PEL, SEL, POEL2) using x information show similar performances in both populations under SRS, but the SEL and POEL2 are more efficient than the PEL estimator for other designs because SEL and POEL2 methods incorporate the design information more efficiently than the PEL method.

In addition to point estimators, we also computed interval estimators for the POEL2 method with a 95% nominal coverage. The interval estimators were computed by the likelihood ratio method based on the results in Theorem 4.3.2 and Theorem 4.4.2. Table 4.2 presents the simulation results of the interval estimators. In Table 4.2, Wald-type confidence intervals were constructed by $(\hat{\theta} - 2\sqrt{\hat{V}}, \hat{\theta} + 2\sqrt{\hat{V}})$, where $\hat{V}$ was computed by the plug-in method described after Theorem 4.3.1 and Theorem 4.4.1. The Wilk-type confidence intervals are computed by the method in Theorem 4.3.2 and Theorem 4.4.2. The actual coverage rates of the Wilk-type confidence intervals are very close to the nominal coverage rates in the simulation study. In general, the Wilk-type confidence intervals show better coverage properties than the Wald-type confidence intervals in terms of coverage rates. We found that similar results hold for SEL method.

4.6.2 Simulation Two

In the second simulation study, we consider combining information from the two independent surveys discussed in Section 5. In this simulation, an artificial finite population of size $N = 10,000$ was generated from

$$y_i = 1 + 0.8(z_i - 3) + 1.5x_i + (z_i/5)e_i,$$

where z_i are generated from $\chi^2(2) + 1$, $e_i \sim \chi^2(1) - 1$, and $x_i \sim N(2, 1)$. From the finite population, we repeatedly generated two independent samples, A_1 and A_2, with sample sizes $n_1 = 500$ and $n_2 = 200$, respectively, and $B = 2,000$ times. The sampling design for survey 1 is the simple random sampling without replacement with sample size $n_1 = 500$. From the survey 1 sample, we only observe x_i. The sampling design for survey 2 is the rejective Poisson sampling with fixed sample size. For the rejective Poisson sampling, we used $\pi_{i2} = n_2 z_i / \sum_{i=1}^{N} z_i$ for the initial selection probability. From survey 2 sample, we observe x_i and y_i. The parameter of interest is the population mean of y.

From each sample pair generated as above, we computed four point estimates:

1. Pseudo empirical likelihood estimator (Wu 2004), which is denoted as $\hat{\theta}_{PEL}$, and $\hat{\theta}_{PEL} = \sum_{j \in s_2} \hat{q}_j y_j$, where $\hat{q}_j$ is obtained by maximizing $l = \sum_{i \in s_1} d_{1i} \log(p_i) + \sum_{j \in s_2} d_{2j} \log(q_j)$, subject to $\sum_{i \in s_1} p_i = \sum_{j \in s_2} q_j = 1$ and $\sum_{i \in s_1} p_i x_i = \sum_{j \in s_2} q_j x_j$.

2. The naive optimal estimator, denoted as $\hat{\theta}_{opt1}$, which can be written as

$$\hat{\theta}_{opt1} = \bar{y}_{d,2} + (\bar{x}_1 - \bar{x}_{d,2})\,\hat{B}_{opt},$$

with $\bar{x}_1 = n_1^{-1} \sum_{i \in s_1} x_i$, $(\bar{x}_{d,2}, \bar{y}_{d,2}) = \left(\sum_{i \in s_2} \pi_{i2}^{-1}\right)^{-1} \sum_{i \in s_2} \pi_{i2}^{-1}(x_i, y_i)$ and

$$\hat{B}_{opt} = \left\{\hat{V}(\bar{x}_1) + \hat{V}(\bar{x}_{d,2})\right\}^{-1} \hat{Cov}(\bar{y}_{d,2}, \bar{x}_{d,2})$$

.

3. The augmented optimal estimator, denoted as $\hat{\theta}_{opt2}$, which can be written as

$$\hat{\theta}_{opt2} = \bar{y}_{d,2} + (\bar{x}_1 - \bar{x}_{d,2})\,\hat{B}_{opt1} + (\bar{\pi}_{2N} - \bar{\pi}_{d,2})\,\hat{B}_{opt2},$$

   where $\hat{B}_{opt} = (\hat{B}'_{opt1}, \hat{B}'_{opt2})' = \hat{V}^{-1}(\bar{S}_d)\hat{Cov}(\bar{y}_{d,2}, \bar{S}_d)$, $\bar{S}_d = [(\bar{x}_{d,2} - \bar{x}_1), (\bar{\pi}_{d,2} - \bar{\pi}_{2N})]'$, $\bar{\pi}_{d,2} = \sum_{i=1}^N I_{2i}\pi_{2i}^{-1}\pi_{2i} / \sum_{i=1}^N I_{2i}\pi_{2i}^{-1}$, $\bar{\pi}_{2N} = N^{-1}\sum_{i=1}^N \pi_{2i}$.

4. Proposed POEL estimator $\hat{\theta}_{POEL}$ using constraints $\sum_{i=1}^N \omega_i = 1$,

$$\sum_{i=1}^N \omega_i I_{1i}\pi_{1i}^{-1} = \sum_{i=1}^N \omega_i I_{2i}\pi_{2i}^{-1} = 1, \quad \sum_{i=1}^N \omega_i I_{1i}\pi_{1i}^{-1} x_i = \sum_{i=1}^N \omega_i I_{2i}\pi_{2i}^{-1} x_i,$$

   and two design constraints $\sum_{i \in s_1} \omega_i = n_1/N$ and $\sum_{i=1}^N \omega_i I_{2i} = \sum_{i=1}^N \omega_i \pi_{2i}$.

The augmented optimal estimator is included to show the effect of incorporating the inclusion probability into the estimation. Table 4.3 presents the biases, variances, and the mean squared errors of the four point estimates. The proposed POEL estimator is more efficient than the naive optimal estimator because it incorporates additional information associated with a fixed sample size for survey 2. The performance of the augmented optimal estimator is close to the proposed POEL estimator, which confirms our theory in Section 5.

4.7 Concluding remarks

We have considered a new empirical-likelihood-type estimator that incorporates the population level information effectively. Instead of using a sample-level likelihood for optimization, we propose using the population level objective function (4.9) for constrained optimization. The objective function (4.9) can be viewed as a population-level nonparametric likelihood when the finite population is treated as a random sample from a superpopulation model. In the purely design-based approach, superpopulation model is not assumed and the objective function in

(4.9) is regarded as the negation of a distance function

$$\sum_{i=1}^{N}\left(\frac{1}{N}\right)\log\left(\frac{1/N}{\omega_i}\right)$$

where the distance is the Kullback-Leibler divergence from $(N^{-1},\cdots,N^{-1})$ to $(\omega_1,\cdots,\omega_N)$. The sampling design is incorporated into the constraints, rather than into the objective function for optimization, when solving the population empirical likelihood estimator. Auxiliary information for the population can also be incorporated into the constraint of the population empirical likelihood method.

The optimality of the proposed estimator follows under the assumption when the sampling fraction, n/N, is negligible. If the sampling rate is not negligible, then, instead of (4.13), we can use $\sum_{i=1}^{N}\omega_i(I_i/\pi_i-1)h_i=0$ in the constraint, as suggested by Qin, Zhang, and Leung (2009) in the context of missing data problems. In this case, the calibration condition holds only asymptotically, but not exactly. Population size N is needed to implement the population empirical likelihood method. If N is unknown, the sample empirical likelihood method discussed in Remark 4.3.2 or the new approach proposed by Berger and De La Riva Torres (2012) can be used. Further extension of the proposed method, including extension to other complex sampling designs and variable selection for calibration, can be a topic of future research.

Table 4.1 Monte Carlo biases, variances, and mean squared errors of the point estimators.

Population	Design	Method	Bias	Var	MSE
A	SRSWOR	HJ	-0.006	0.046	0.046
		HT	-0.006	0.046	0.046
		POEL1	-0.006	0.046	0.046
		PEL	-0.003	0.010	0.010
		SEL	-0.001	0.009	0.009
		POEL2	-0.001	0.009	0.009
	Poisson	HJ	0.011	0.043	0.043
		HT	0.004	0.035	0.035
		POEL1	0.004	0.035	0.035
		PEL	0.001	0.008	0.008
		SEL	0.003	0.007	0.007
		POEL2	0.003	0.007	0.007
	Rejective Poisson	HJ	0.000	0.039	0.039
		HT	-0.004	0.028	0.028
		POEL1	-0.002	0.016	0.0165
		PEL	-0.005	0.008	0.008
		SEL	-0.002	0.006	0.006
		POEL2	-0.002	0.006	0.006
B	SRSWOR	HJ	-0.005	0.070	0.070
		HT	-0.005	0.070	0.070
		POEL1	-0.005	0.070	0.070
		PEL	-0.005	0.024	0.024
		SEL	-0.003	0.024	0.024
		POEL2	-0.003	0.024	0.024
	Poisson	HJ	0.007	0.038	0.038
		HT	0.000	0.034	0.034
		POEL1	0.000	0.030	0.030
		PEL	-0.001	0.022	0.022
		SEL	-0.001	0.016	0.016
		POEL2	-0.002	0.016	0.016
	Rejective Poisson	HJ	0.003	0.037	0.037
		HT	-0.002	0.019	0.019
		POEL1	-0.003	0.014	0.014
		PEL	-0.001	0.022	0.022
		SEL	-0.004	0.013	0.013
		POEL2	-0.004	0.013	0.013

HJ: Hájek estimator, HT: Horvitz-Thompson estimator, PEL: Pseudo Empirical Likelihood estimator, SEL: Proposed sample EL estimator, POEL1: Proposed population EL estimator(without using x information), POEL2: Proposed population EL estimator incorporating x information, SRSWOR: Simple Random Sampling Without Replacement.

Table 4.2 Coverage rate and average length comparison for Wald's and Wilk's type 95% confidence intervals of proposed POEL2 method.

Population	Sampling design	Method	Coverage rate	Average length
A	SRSWOR	Wald	0.923	0.362
		Wilk	0.934	0.379
	Poisson	Wald	0.931	0.313
		Wilk	0.942	0.327
	Rejective Poisson	Wald	0.932	0.309
		Wilk	0.944	0.322
B	SRSWOR	Wald	0.923	0.580
		Wilk	0.938	0.598
	Poisson	Wald	0.935	0.486
		Wilk	0.944	0.503
	Rejective Poisson	Wald	0.936	0.450
		Wilk	0.949	0.471

Table 4.3 The Monte Carlo biases, variances, and the mean squared errors (MSE) of the point estimators in Simulation Two.

Method	Bias	Var	MSE
Pseudo EL	0.009	0.019	0.019
Naive Optimal	0.008	0.017	0.017
Augmented Optimal	-0.002	0.006	0.006
Proposed POEL	0.002	0.006	0.006

CHAPTER 5. TWO-PHASE SAMPLING FOR PROPENSITY SCORE ESTIMATION IN VOLUNTARY SAMPLES

A paper submitted to the *Journal of the American Statistical Association*

Sixia Chen and Jae Kwang Kim

Abstract

Voluntary sampling is a non-probability sampling design with unknown sample inclusion probabilities. When the sample inclusion probability depends on the study variables, propensity score adjustment using auxiliary information may lead to biased estimation. In this paper, we propose a novel application of two-phase sampling to estimate the parameters in the propensity model. The proposed method includes an experiment in which data are collected again from a subset of the original voluntary sample. With this two-phase sampling experiment, we can estimate the parameters in a propensity score model consistently. Then the propensity score adjustment can be applied to the original voluntary sample to estimate the population parameters. The proposed method can be extended to non-nested two-phase samples. Results from simulation studies are presented and the proposed method applied to the 2012 Iowa Caucus Survey.

Key Words: Capture-recapture; Nonignorable missing; Self-selected sample; Web surveys.

5.1 Introduction

Voluntary sampling, or self-selected sampling, is sampling where the inclusion of sampling units is determined by the units themselves agreeing to participate in the study. When survey units are chosen by surveyors, but these units nonetheless elect not to participate, self-selection occurs. If many units elect not to participate, the representativeness of the observed sample can be called into question. In non-probability sampling, such as web surveys, the contact probabilities are unknown, and the participation probabilities are unknown. A valid analysis of voluntary sampling is extremely difficult when survey participation is related to survey items. There exist sociological theories, such as the leverage-saliency theory (Groves et al, 2000), that try to identify psychological factors influencing survey participation, but it is not clear how to use those theories to analyze observed data.

In spite of the danger of selection bias in the voluntary sample, voluntary sampling is increasingly popular, reflecting the fact that complete control of survey participation is often not achievable. There are procedures for reducing the bias of estimators from voluntary samples. Propensity score weighting is a common method. Rosenbaum and Rubin (1983) as well as Rosenbaum (1987) proposed the basic theory of using propensity scores to estimate treatment effects in observational studies. Duncan and Stasny (2001) used the propensity score method to control coverage bias in telephone surveys. Lee (2006) applied the propensity score method to a volunteer panel web survey. Lee and Valliant (2009) and Valliant and Dever (2011) considered the propensity score method for a web-based voluntary sample. All of these studies assumed an ignorable selection mechanism. That is, it was assumed that the sample inclusion probability depends on one or more auxiliary variables with known or estimated marginal distributions. In other words, the selection mechanism was assumed to be missing at random in the sense of Rubin (1976). If that is the case, propensity scores can be consistently estimated and the resulting analysis is valid under the assumed propensity score model.

In voluntary sampling, the ignorable selection mechanism assumption is not realistic because it is well known that survey participation is related to the survey topic of interest (Groves et al, 2004). As a corollary, the propensity model using only demographic auxiliary variables may lead to biased estimation. In this paper, we consider the nonignorable selection mechanism in the propensity model for survey participation. To estimate the parameters of the propensity model consistently, we propose a novel application of the capture-recapture experiment in a voluntary survey with voluntary respondents contacted twice. Unlike regular capture-recapture sampling, only the respondents to the first survey are contacted again. Thus, the overall sampling follows a two-phase sampling scheme, because the second-phase sample is nested within the first-phase sample. In Section 4, we discuss an extension to non-nested two-phase sampling scheme where the two voluntary samples are selected independently.

Our paper is motivated by a telephone survey for the 2012 Iowa Caucus. In this survey, the individuals obtained from a probability sampling procedure were asked about their intention to vote in the 2012 Iowa Caucus. Because of the low response rate (15%), the sample of respondents cannot be viewed as a probability sample. Thus, it is reasonable to assume that selection probability depends on the study variables (intention to vote for given candidates). In November 2011, the first-phase voluntary sample was obtained and then the second voluntary sample was obtained from the first voluntary sample in the next month. Because the survey questions for both surveys were quite similar, we treat the two voluntary sampling mechanisms as identical, up to overall response rates. The model parameters are estimated from the two-phase sample and the final estimates for voting intention are computed using the estimated propensity. Further details are presented in Section 6.

5.2 Basic Setup

Let U be a finite population and $A_1(\subset U)$ be a voluntary sample obtained by an unknown sampling mechanism. In sample A_1, we observe $(\mathbf{x}_i', y_{1i})$, where $\mathbf{x}_i$ is the vector of auxiliary variables and y_{1i} is the realized value of the study variable of interest at the time of observing elements in A_1. We assume that the population size N is known. We also assume that the probability of being included in the sample is a function of $\mathbf{x}$ and y. We define the sampling model for A_1 to be

$$\pi_{1i}(\phi) = Pr\left(\delta_{1i} = 1 \mid \mathbf{x}_i, y_{1i}\right) = \frac{\exp(\phi_0 + \phi_1'\mathbf{x}_i + \phi_2 y_{1i})}{1 + \exp(\phi_0 + \phi_1'\mathbf{x}_i + \phi_2 y_{1i})}, \tag{5.1}$$

where δ_{1i} is the indicator function for element i to be in sample A_1.

To estimate the parameters in (5.1), we subject the respondents in the first-phase sample to similar survey questions and obtain a second voluntary sample A_2 from A_1. That is, we perform two-phase sampling under the same voluntary sampling mechanism. The sampling model for A_2 is

$$\pi_{2i}(\phi^*) = Pr\left(\delta_{2i} = 1 \mid \mathbf{x}_i, y_{2i}, \delta_{1i} = 1\right) = \frac{\exp(\phi_0^* + \phi_1'\mathbf{x}_i + \phi_2 y_{2i})}{1 + \exp(\phi_0^* + \phi_1'\mathbf{x}_i + \phi_2 y_{2i})}, \tag{5.2}$$

where (ϕ_1', ϕ_2) is defined in (5.1). Thus, we assume that the conditional odds for the first-phase selection and for the second-phase selection are the same . In addition, we assume that the population size, N, is available from an external source. Here, we allow the study item Y can be time-dependent; in other words, the value of Y can change over time. Thus, y_{2i} is the measurement of Y at the time of selecting A_2. We are interested in estimating $\theta_1 = E(Y_1)$ and $\theta_2 = E(Y_2)$ from the two-phase sample.

We now discuss parameter estimation for the propensity models. To estimate the parameters, note that we observe $(\mathbf{x}_i', y_{1i}, y_{2i})$ in A_2. Thus, we can construct the following estimating equation to estimate the parameters in (5.2).

$$\sum_{i \in A_1} \left\{ \frac{\delta_{2i}}{\pi_{2i}(\phi^*)} - 1 \right\} \mathbf{h}_{1i} = 0, \tag{5.3}$$

where $\mathbf{h}_{1i} = (1, \mathbf{x}_i', y_{1i})'$. Once $\hat{\phi}^* = (\hat{\phi}_0^*, \hat{\phi}_1', \hat{\phi}_2)'$ is computed, we can use

$$\sum_{i \in A_1} \frac{1}{\pi_{1i}(\phi_0, \hat{\phi}_1, \hat{\phi}_2)} = N$$

to estimate ϕ_0. Equation (5.3) is a calibration equation in the second-phase sample using $\mathbf{h}_{1i}$ as the control variable. Use of calibration for propensity score adjustment has been considered by Fuller et al (1994), Kott (2006), and Kott and Chang (2010).

Once the parameters in (5.1) and (5.2) are estimated, we can use the following propensity-score-adjusted (PSA) estimator

$$\hat{\theta}_1 = \frac{1}{N} \sum_{i \in A_1} \hat{\pi}_{1i}^{-1} y_{1i} \tag{5.4}$$

to estimate $\theta_1 = E(Y_1)$. Also, we can use

$$\hat{\theta}_2 = \frac{1}{N} \sum_{i \in A_2} \hat{\pi}_{1i}^{-1} \hat{\pi}_{2i}^{-1} y_{2i} \tag{5.5}$$

to estimate $\theta_2 = E(Y_2)$. In addition, we want to use the population-level information of $\mathbf{x}$. Variance estimation is also possible with this setup under the assumption that the sample selection models (5.1) and (5.2) are correct.

5.3 Main Results

In this section, we discuss some asymptotic properties of the proposed PSA estimators. To discuss asymptotic properties of $\hat{\theta}_1$ in (5.4), we first define $\Phi = (\phi_0^*, \phi_1', \phi_2, \phi_0)'$,

$$U_1(\Phi) \triangleq \sum_{i \in A_1} \left\{ \frac{\delta_{2i}}{\pi_{2i}(\phi_0^*, \phi_1, \phi_2)} - 1 \right\} (1, \mathbf{x}_i', y_{1i})' = (0, 0, 0)' \quad (5.6)$$

and

$$U_2(\Phi) \triangleq \sum_{i \in A_1} \frac{1}{\pi_{1i}(\phi_0, \phi_1, \phi_2)} - N = 0. \quad (5.7)$$

Thus, equations (5.6) and (5.7) are a system of nonlinear equations that can be solved for Φ. We can write $U_c(\Phi)' = [U_1(\Phi)', U_2(\Phi)']$, and $(\hat{\theta}_1, \hat{\Phi}')'$ can be obtained as the solution to

$$\begin{aligned} U_p(\theta_1, \Phi) &= 0 \\ U_c(\Phi) &= 0, \end{aligned}$$

where $U_p(\theta_1, \Phi) = N^{-1} \sum_{i \in A_1} \{\pi_{1i}(\phi_0, \phi_1, \phi_2)\}^{-1} y_{1i} - \theta_1$. Because $E\{U_p(\theta_1^*, \Phi^*)\} = 0$ and $E\{U_c(\theta_1^*, \Phi^*)\} = 0$, where $(\theta_1^*, \Phi^{*'})'$ is the true parameter values, the solution $(\hat{\theta}_1, \hat{\Phi}')'$ is consistent and has asymptotic variance

$$V\begin{pmatrix} \hat{\theta}_1 \\ \hat{\Phi} \end{pmatrix} \cong \begin{Bmatrix} -1 & E(\partial U_p/\partial \Phi) \\ 0 & E(\partial U_c/\partial \Phi) \end{Bmatrix}^{-1} \begin{Bmatrix} V(U_p) & C(U_p, U_c) \\ C(U_c, U_p) & V(U_c) \end{Bmatrix} \begin{Bmatrix} -1 & E(\partial U_p/\partial \Phi) \\ 0 & E(\partial U_c/\partial \Phi) \end{Bmatrix}^{\prime -1}.$$

Use

$$\begin{bmatrix} -1 & E(\partial U_p/\partial \Phi) \\ 0 & E(\partial U_c/\partial \Phi) \end{bmatrix}^{-1} = \begin{bmatrix} -1 & E(\partial U_p/\partial \Phi)\{E(\partial U_c/\partial \Phi)\}^{-1} \\ 0 & \{E(\partial U_c/\partial \Phi)\}^{-1} \end{bmatrix},$$

then the asymptotic variance of $\hat{\theta}_1$ can be written, using the definition of U_p and U_c, as

$$\begin{aligned} V\left(\hat{\theta}_1\right) &\cong V\left\{U_p - E(\frac{\partial U_p}{\partial \Phi})\left\{E(\frac{\partial U_c}{\partial \Phi})\right\}^{-1} U_c\right\} \\ &= V\left\{\hat{\theta}_1(\Phi) - E\left\{\frac{\partial}{\partial \Phi}\hat{\theta}_1(\Phi)\right\}\begin{bmatrix} E\{\partial U_1(\Phi)/\partial \Phi\} \\ E\{\partial U_2(\Phi)/\partial \Phi\} \end{bmatrix}^{-1} U_c\right\}, \end{aligned}$$

where $\hat{\theta}_1(\Phi) = N^{-1} \sum_{i \in A_1} y_{1i}\{1 + \exp(-\phi_0 - \phi_1' \mathbf{x}_i - \phi_2 y_{1i})\}$. Thus, the asymptotic variance can be written as

$$V\left(\hat{\theta}_1\right) \cong \frac{1}{N^2} V\left[\sum_{i=1}^N \frac{\delta_{1i}}{\pi_{1i}} y_{1i} - B_{1,y}\left\{\sum_{i=1}^N \delta_{1i}\left(\frac{\delta_{2i}}{\pi_{2i}} - 1\right)\mathbf{h}_{1i}\right\} - B_{2,y}\sum_{i=1}^N\left(\frac{\delta_{1i}}{\pi_{1i}} - 1\right)\right], \quad (5.8)$$

and

$$\begin{aligned}
(B_{1,y}, B_{2,y}) &= N \times E\left\{\frac{\partial}{\partial \Phi}\hat{\theta}_1(\Phi)\right\}\left[\begin{array}{c} E\{\partial U_1(\Phi)/\partial\Phi\} \\ E\{\partial U_2(\Phi)/\partial\Phi\} \end{array}\right]^{-1} \\
&= \sum_{i=1}^{N}(1-\pi_{1i})y_{1i}(0,\mathbf{x}_i', y_{1i}, 1)\left(\begin{array}{c} \sum_{i=1}^{N}\pi_{1i}(1-\pi_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i=1}^{N}(1-\pi_{1i})(0,\mathbf{x}_i', y_{1i}, 1) \end{array}\right)^{-1},
\end{aligned}$$

where $\mathbf{h}_{2i} = (1, \mathbf{x}_i', y_{2i})'$ and $\mathbf{0}_{r\times 1}$ is the vector of zeros with dimension $r \times 1$, with $r = 2 + p$, and p is the dimension of $\mathbf{x}_i$. Note that the variance (5.8) can be written as

$$V\left(\hat{\theta}_1\right) = V\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{1i}}{\pi_{1i}}(y_{1i} - B_{2,y})\right\} + V\left\{\frac{1}{N}\sum_{i=1}^{N}B_{1,y}\delta_{1i}(\frac{\delta_{2i}}{\pi_{2i}} - 1)\mathbf{h}_{1i}\right\}. \tag{5.9}$$

Roughly speaking, the first term in (5.9) is the asymptotic variance of the PSA estimator when (ϕ_1', ϕ_2) is known and the second term is the additional variance due to the fact that (ϕ_1', ϕ_2) is estimated from the second-phase sample. For variance estimation, we replace the unknown parameters with their estimators in (5.8) to obtain

$$\hat{V}\left(\hat{\theta}_1\right) = \frac{1}{N^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}(y_{1i} - \hat{B}_{2,y})^2 + \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}(\hat{B}_{1,y}\mathbf{h}_{1i})^2,$$

where

$$(\hat{B}_{1,y}, \hat{B}_{2,y}) = \sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}}y_{1i}(0,\mathbf{x}_i', y_{1i}, 1)\left(\begin{array}{c} \sum_{i\in A_2}\hat{\pi}_{2i}^{-1}(1-\hat{\pi}_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})(0,\mathbf{x}_i', y_{1i}, 1) \end{array}\right)^{-1}.$$

We now discuss the asymptotic properties of $\hat{\theta}_2$ in (5.5), i.e. the direct PSA estimator of θ_2. Using the argument similar to (5.8), we can obtain

$$\begin{aligned}
V\left(\hat{\theta}_2\right) &\cong \frac{1}{N^2}V\left[\sum_{i=1}^{N}\frac{\delta_{1i}\delta_{2i}}{\pi_{1i}\pi_{2i}}y_{2i} - D_{1,y}\left\{\sum_{i=1}^{N}\delta_{1i}\left(\frac{\delta_{2i}}{\pi_{2i}} - 1\right)\mathbf{h}_{1i}\right\} - D_{2,y}\sum_{i=1}^{N}\left(\frac{\delta_{1i}}{\pi_{1i}} - 1\right)\right] \\
&= \frac{1}{N^2}V\left[\sum_{i=1}^{N}\frac{\delta_{1i}}{\pi_{1i}}(y_{2i} - D_{2,y})\right] + \frac{1}{N^2}E\left\{\sum_{i\in A_1}\frac{1-\pi_{2i}}{\pi_{1i}^2\pi_{2i}}(y_{2i} - D_{1,y}\pi_{1i}\mathbf{h}_{1i})^2\right\},
\end{aligned}$$

where

$$(D_{1,y}, D_{2,y}) = N \times E\left\{\frac{\partial}{\partial \Phi}\hat{\theta}_2(\Phi)\right\}\begin{bmatrix} E\{\partial U_1(\Phi)/\partial \Phi\} \\ E\{\partial U_2(\Phi)/\partial \Phi\}\end{bmatrix}^{-1}$$

$$= \sum_{i=1}^{N} y_{2i}\left\{(1-\pi_{1i})(0, \mathbf{x}_i', y_{1i}, 1) + (1-\pi_{2i})(1, \mathbf{x}_i', y_{2i}, 0)\right\}$$

$$\times \begin{pmatrix} \sum_{i=1}^{N} \pi_{1i}(1-\pi_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i=1}^{N}(1-\pi_{1i})(0, \mathbf{x}_i', y_{1i}, 1)\end{pmatrix}^{-1}$$

and

$$\hat{\theta}_2(\Phi) = \frac{1}{N}\sum_{i\in A_2} y_{2i}\left\{1+\exp(-\phi_0 - \phi_1'\mathbf{x}_i - \phi_2 y_{1i})\right\}\left\{1+\exp(-\phi_0^* - \phi_1'\mathbf{x}_i - \phi_2 y_{2i})\right\}.$$

Thus, a consistent estimator for the variance of $\hat{\theta}_2$ in (5.5) is given by

$$\hat{V}(\hat{\theta}_2) = \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2\hat{\pi}_{2i}}(y_{2i} - \hat{D}_{2,y})^2 + \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{1i}^2\hat{\pi}_{2i}^2}(y_{2i} - \hat{D}_{1,y}\hat{\pi}_{1i}\mathbf{h}_{1i})^2,$$

where

$$(\hat{D}_{1,y}, \hat{D}_{2,y}) = \sum_{i\in A_2}\frac{y_{2i}}{\hat{\pi}_{1i}\hat{\pi}_{2i}}\left\{(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1) + (1-\hat{\pi}_{2i})(1, \mathbf{x}_i', y_{2i}, 0)\right\}$$

$$\times \begin{pmatrix} \sum_{i\in A_2} \hat{\pi}_{2i}^{-1}(1-\hat{\pi}_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1)\end{pmatrix}^{-1}. \quad (5.10)$$

Instead of using the direct estimator $\hat{\theta}_2$ in (5.5), we can use a two-phase regression estimator to improve efficiency. The two-phase regression estimator is efficient in that it incorporates auxiliary information obtained from the first-phase sampling. See Hidiroglou and Särndal (1998), Legg and Fuller (2009), and Kim and Yu (2011) for more details about two-phase regression estimators. In our setup, the data vector $\mathbf{h}_{1i} = (1, \mathbf{x}_i', y_{1i})'$ is available for both A_1 and A_2. Thus, the two natural estimators for the population mean $\bar{\mathbf{h}}_{1N} = N^{-1}\sum_{i=1}^{N}\mathbf{h}_{1i}$, $\hat{\mathbf{h}}_{1,1} = N^{-1}\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}\mathbf{h}_{1i}$ and $\hat{\mathbf{h}}_{2,1} = N^{-1}\sum_{i\in A_2}\hat{\pi}_{1i}^{-1}\hat{\pi}_{2i}^{-1}\mathbf{h}_{1i}$ can be computed from A_1 and A_2, respectively, and they are both approximately unbiased for $\bar{\mathbf{h}}_{1N}$. Using $\hat{\mathbf{h}}_{1,1}$ and $\hat{\mathbf{h}}_{2,1}$, the two-phase regression estimator can be constructed by

$$\hat{\theta}_{2,Reg} = \hat{\theta}_2 - \hat{C}_{h_1}(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}), \quad (5.11)$$

where

$$\hat{C}_{h_1} = \sum_{i\in A_2} \hat{\pi}_{1i}^{-1}\hat{\pi}_{2i}^{-1} y_{2i}\mathbf{h}_{1i}' \left\{ \sum_{i\in A_2} \hat{\pi}_{1i}^{-1}\hat{\pi}_{2i}^{-1}\mathbf{h}_{1i}\mathbf{h}_{1i}' \right\}^{-1} . \tag{5.12}$$

Because $E(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}) \cong 0$, the regression estimator in (5.11) is approximately unbiased, regardless of the choice of $\hat{C}_{h_1}$. By applying the linearization method to each term of (5.11), we can get

$$\hat{\theta}_{2,Reg} \cong \frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{\delta_{1i}\delta_{2i}}{\pi_{1i}\pi_{2i}} y_{2i} - D_{1i,Reg}\delta_{1i}\left(\frac{\delta_{2i}}{\pi_{2i}} - 1\right)\mathbf{h}_{1i} - D_{2,Reg}\left(\frac{\delta_{1i}}{\pi_{1i}} - 1\right) \right\},$$

where $D_{1i,Reg} = D_{1,y} + C_{h_1}^{*}\pi_{1i}^{-1} - C_{h_1}^{*}(D_{1,h_1} - B_{1,h_1})$ and $D_{2,Reg} = D_{2,y} - C_{h_1}^{*}(D_{2,h_1} - B_{2,h_1})$ with

$$(B_{1,h_1}, B_{2,h_1}) = \sum_{i=1}^{N}(1-\pi_{1i})\mathbf{h}_{1i}(0, \mathbf{x}_i', y_{1i}, 1) \begin{pmatrix} \sum_{i=1}^{N} \pi_{1i}(1-\pi_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i=1}^{N}(1-\pi_{1i})(0, \mathbf{x}_i', y_{1i}, 1) \end{pmatrix}^{-1},$$

$C_{h_1}^{*} = p\lim \hat{C}_{h_1}$,

$$\begin{aligned} (D_{1,h_1}, D_{2,h_1}) &= \sum_{i=1}^{N} \mathbf{h}_{1i}\left\{(1-\pi_{1i})(0, \mathbf{x}_i', y_{1i}, 1) + (1-\pi_{2i})(1, \mathbf{x}_i', y_{2i}, 0)\right\} \\ &\times \begin{pmatrix} \sum_{i=1}^{N} \pi_{1i}(1-\pi_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i=1}^{N}(1-\pi_{1i})(0, \mathbf{x}_i', y_{1i}, 1) \end{pmatrix}^{-1}. \end{aligned}$$

Thus, the asymptotic variance is

$$\begin{aligned} V\left(\hat{\theta}_{2,Reg}\right) &\cong \frac{1}{N^2} V\left[\sum_{i=1}^{N} \frac{\delta_{1i}\delta_{2i}}{\pi_{1i}\pi_{2i}} y_{2i} - \left\{\sum_{i=1}^{N} D_{1i,Reg}\delta_{1i}\left(\frac{\delta_{2i}}{\pi_{2i}} - 1\right)\mathbf{h}_{1i}\right\} - D_{2,Reg}\sum_{i=1}^{N}\left(\frac{\delta_{1i}}{\pi_{1i}} - 1\right)\right] \\ &= \frac{1}{N^2} V\left[\sum_{i=1}^{N} \frac{\delta_{1i}}{\pi_{1i}}(y_{2i} - D_{2,Reg})\right] + \frac{1}{N^2} E\left\{\sum_{i\in A_1} \frac{1-\pi_{2i}}{\pi_{1i}^2\pi_{2i}}(y_{2i} - D_{1i,Reg}\pi_{1i}\mathbf{h}_{1i})^2\right\}. \end{aligned}$$

A consistent estimator fo variance of the two-phase regression estimator is given by

$$\hat{V}(\hat{\theta}_{2,Reg}) = \frac{1}{N^2}\sum_{i\in A_2} \frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2\hat{\pi}_{2i}}(y_{2i} - \hat{D}_{2,Reg})^2 + \frac{1}{N^2}\sum_{i\in A_2} \frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{1i}^2\hat{\pi}_{2i}^2}(y_{2i} - \hat{D}_{1i,Reg}\hat{\pi}_{1i}\mathbf{h}_{1i})^2,$$

where $\hat{D}_{1i,Reg} = \hat{D}_{1,y} + \hat{C}_{h_1}\hat{\pi}_{1i}^{-1} - \hat{C}_{h_1}(\hat{D}_{1,h_1} - \hat{B}_{1,h_1})$ and $\hat{D}_{2,Reg} = \hat{D}_{2,y} - \hat{C}_{h_1}(\hat{D}_{2,h_1} - \hat{B}_{2,h_1})$, with

$$(\hat{B}_{1,h_1}, \hat{B}_{2,h_1}) = \sum_{i\in A_1} \frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}}\mathbf{h}_{1i}(0, \mathbf{x}_i', y_{1i}, 1) \begin{pmatrix} \sum_{i\in A_2} \hat{\pi}_{2i}^{-1}(1-\hat{\pi}_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i\in A_1} \hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1) \end{pmatrix}^{-1}, \tag{5.13}$$

$$(\hat{D}_{1,h_1}, \hat{D}_{2,h_1}) = \sum_{i\in A_2} \frac{\mathbf{h}_{1i}}{\hat{\pi}_{1i}\hat{\pi}_{2i}} \left\{(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1) + (1-\hat{\pi}_{2i})(1, \mathbf{x}_i', y_{2i}, 0)\right\} \times \begin{pmatrix} \sum_{i\in A_2} \hat{\pi}_{2i}^{-1}(1-\hat{\pi}_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i\in A_1} \hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1) \end{pmatrix}^{-1} \tag{5.14}$$

and $(\hat{D}_{1,y}, \hat{D}_{2,y})$ is defined in (5.10).

Remark 5.3.1 *Instead of $\hat{C}_{h_1}$ in (5.12), the optimal choice of $\hat{C}_{h_1}$ that minimizes the variance among the class of regression estimators with a form specified by (5.11) is*

$$\hat{C}_{h_1,opt} = \hat{C}\left(\hat{\theta}_2, \hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}\right)\left\{\hat{V}(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1})\right\}^{-1}, \tag{5.15}$$

which reduces to

$$\hat{C}_{h_1,opt} = \left\{\frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}(\frac{y_{2i}}{\hat{\pi}_{1i}} - \hat{D}_{1,y}\mathbf{h}_{1i})\hat{\boldsymbol{\eta}}_i' - \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}(\frac{y_{2i}}{\hat{\pi}_{2i}} - \hat{D}_{2,y})\hat{\boldsymbol{\tau}}'\right\} \times \left(\frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}\hat{\boldsymbol{\eta}}_i^{\otimes 2} + \frac{1}{N^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}\hat{\boldsymbol{\tau}}^{\otimes 2}\right)^{-1}, \tag{5.16}$$

where $\hat{\boldsymbol{\eta}}_i = \mathbf{h}_{1i}'\hat{\pi}_{1i}^{-1} - (\hat{D}_{1,h_1} - \hat{B}_{1,h_1})\mathbf{h}_{1i}'$, $\hat{\boldsymbol{\tau}} = \hat{D}_{2,h_1} - \hat{B}_{2,h_1}$ with $(\hat{B}_{1,h_1}, \hat{B}_{2,h_1})$ and $(\hat{D}_{1,h_1}, \hat{D}_{2,h_1})$ defined in (5.13) and (5.14).

In Section A of Appendix D, we further discuss how to make use of the population level auxiliary information.

5.4 Extension to non-nested two-phase sampling

In this section, we consider a non-nested two-phase sampling setup where the two samples, A_1 and A_2, are independently selected. In that case, the two samples are assumed to be obtained independently and the classical capture-recapture (CR) sampling setup can be applied. We extend the idea of CR experiments to estimate the selection probabilities of the volunteer sample. Capture-recapture (CR) sampling is very popular in estimating the population size of wildlife animals. Amstrup et al. (2005) provided a comprehensive summaries of the existing methods for CR analysis. Pollock et al. (1984), Huggins, (1989, 1991), Alho (1990) incorporated covariates information into a CR experiment and used the conditional likelihood approach to do inference. Huggins and Hwang (2011) provided a review of the conditional likelihood approach in CR experiments.

To apply the conditional likelihood approach, we assume that the measurement for y is the same. That is, $y_1 = y_2 = y$. Thus, $(\mathbf{x}_i', y_i)$ are observed in A_1 and A_2 and the two sampling indicators, δ_{1i} and δ_{2i}, are assumed to be independently generated from Bernoulli distributions with probabilities

$$\pi_{1i}(\phi) = \Pr(\delta_{1i} = 1 | \mathbf{x}_i, y_i) = \frac{\exp(\phi_0 + \phi_1' \mathbf{x}_i + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1' \mathbf{x}_i + \phi_2 y_i)}$$

and

$$\pi_{2i}(\phi^*) = \Pr(\delta_{2i} = 1 | \mathbf{x}_i, y_i) = \frac{\exp(\phi_0^* + \phi_1^{*'} \mathbf{x}_i + \phi_2^* y_i)}{1 + \exp(\phi_0^* + \phi_1^{*'} \mathbf{x}_i + \phi_2^* y_i)},$$

respectively, where $\phi = (\phi_0, \phi_1', \phi_2)'$ and $\phi^* = (\phi_0^*, \phi_1^{*'}, \phi_2^*)'$. Write $\Phi = (\phi', \phi^{*'})'$. An efficient estimator of Φ can be obtained by maximizing the conditional likelihood

$$L_C(\Phi) = \prod_{i \in A_1/A_2} \frac{\pi_{1i}(\phi)\left\{1 - \pi_{2i}(\phi^*)\right\}}{p_i(\phi, \phi^*)} \prod_{i \in A_1 \cap A_2} \frac{\pi_{1i}(\phi)\pi_{2i}(\phi^*)}{p_i(\phi, \phi^*)} \prod_{i \in A_2/A_1} \frac{\left\{1 - \pi_{1i}(\phi)\right\}\pi_{2i}(\phi^*)}{p_i(\phi, \phi^*)},$$

where $p_i(\phi, \phi^*) = 1 - \left\{1 - \pi_{1i}(\phi)\right\}\left\{1 - \pi_{2i}(\phi^*)\right\}$. The conditional likelihood is obtained by considering the conditional distribution of $(\delta_{1i}, \delta_{i2})$ given that unit i is selected in either of the two samples. The log-likelihood of the conditional distribution is

$$l_C(\Phi) = \sum_{i \in A_1} \log(\pi_{1i}) + \sum_{i \in A_2} \log(\pi_{2i}) + \sum_{i \in A_1/A_2} \log(1 - \pi_{2i}) + \sum_{i \in A_2/A_1} \log(1 - \pi_{1i}) - \sum_{i \in A_1 \cup A_2} \log(p_i).$$

The conditional maximum likelihood estimator (CMLE) that maximizes the conditional likelihood can be obtained by solving $S_C(\Phi) = 0$ where $S_C(\Phi) = \partial l_C(\Phi)/\partial \Phi = (S_{C1}'(\Phi), S_{C2}'(\Phi))'$ with

$$S_{C1}(\Phi) \triangleq \sum_{i \in A_1} (1, \mathbf{x}_i', y_i)' - \sum_{i \in A_1 \cup A_2} \frac{\pi_{1i}(\phi)}{p_i(\phi, \phi^*)} (1, \mathbf{x}_i', y_i)'$$

and

$$S_{C2}(\Phi) \triangleq \sum_{i \in A_2} (1, \mathbf{x}_i', y_i)' - \sum_{i \in A_1 \cup A_2} \frac{\pi_{2i}(\phi)}{p_i(\phi, \phi^*)} (1, \mathbf{x}_i', y_i)'.$$

Once the CMLE of Φ, denoted by $\hat{\Phi}$, is obtained, we can construct the following propensity score estimator of $\theta = E(Y)$ based on A_1 by

$$\hat{\theta} = \frac{\sum_{i \in A_1} \pi_{1i}^{-1}(\hat{\phi}) y_i}{\sum_{i \in A_1} \pi_{1i}^{-1}(\hat{\phi})}. \tag{5.17}$$

To discuss the asymptotic properties of $\hat{\theta}$ in (5.17), note that the proposed estimators $(\hat{\theta}, \hat{\Phi}')'$ can be written as a solution to

$$U_p(\theta, \Phi) = 0, \quad S_C(\Phi) = 0,$$

where

$$U_p(\theta, \Phi) \triangleq \frac{1}{N} \sum_{i \in A_1} \frac{1}{\pi_{1i}(\phi)} (y_i - \theta).$$

Denote $(\theta^*, \Phi^{*'})'$ as the probability limit of $(\hat{\theta}, \hat{\Phi}')'$, then use Taylor linearization, as presented in Section 3, to get

$$\begin{aligned} \hat{\theta} - \theta^* &\cong U_p(\theta^*, \Phi^*) - E\left\{ \frac{\partial U_p(\theta^*, \Phi^*)}{\partial \Phi} \right\} \left[E\left\{ \frac{\partial S_C(\Phi^*)}{\partial \Phi} \right\} \right]^{-1} S_C(\Phi^*) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\delta_{1i}}{\pi_{1i}} (y_i - \theta^*) - B_1(\delta_{1i} - \frac{\delta_i^*}{p_i} \pi_{1i}) \mathbf{h}_i - B_2(\delta_{2i} - \frac{\delta_i^*}{p_i} \pi_{2i}) \mathbf{h}_i \right\}, \end{aligned} \tag{5.18}$$

where

$$(B_1, B_2) = N \times E\left\{ \frac{\partial U_p(\theta^*, \Phi^*)}{\partial \Phi} \right\} \left[E\left\{ \frac{\partial S_C(\Phi^*)}{\partial \Phi} \right\} \right]^{-1},$$

$$E\left\{ \frac{\partial U_p(\theta^*, \Phi^*)}{\partial \Phi} \right\} = -\frac{1}{N} \sum_{i=1}^{N} (1 - \pi_{1i})(y_i - \theta^*)(\mathbf{h}_i', \mathbf{0}_{1 \times r}),$$

with $r = 2 + p$, and p is the dimension of $\mathbf{x}$.

$$E\left\{ \frac{\partial S_C(\Phi^*)}{\partial \Phi} \right\} = -\sum_{i=1}^{N} p_i^{-1} \pi_{1i} \pi_{2i} \begin{pmatrix} (1 - \pi_{1i}) \mathbf{h}_i \mathbf{h}_i' & -(1 - \pi_{1i})(1 - \pi_{2i}) \mathbf{h}_i \mathbf{h}_i' \\ -(1 - \pi_{1i})(1 - \pi_{2i}) \mathbf{h}_i \mathbf{h}_i' & (1 - \pi_{2i}) \mathbf{h}_i \mathbf{h}_i' \end{pmatrix},$$

and $\mathbf{h}_i = (1, \mathbf{x}_i', y_i)'$. Hence, we have

$$\begin{aligned} V(\hat{\theta}) &\cong V\left[\frac{1}{N} \sum_{i=1}^{N} \frac{\delta_{1i}}{\pi_{1i}} \left\{ y_i - \theta^* - B_1 p_i^{-1} \pi_{1i} \pi_{2i} \mathbf{h}_i + B_2 p_i^{-1} \pi_{1i} \pi_{2i} (1 - \pi_{2i}) \mathbf{h}_i \right\} \right] \\ &+ E\left[\frac{1}{N^2} \sum_{i=1}^{N} \pi_{2i}(1 - \pi_{2i}) \left\{ B_1 p_i^{-1} (1 - \delta_{1i}) \pi_{1i} \mathbf{h}_i - B_2 p_i^{-1} (\pi_{1i} - \pi_{1i} \pi_{2i} + \delta_{1i} \pi_{2i}) \mathbf{h}_i \right\}^2 \right]. \end{aligned}$$

So, the consistent estimator of $V(\hat{\theta})$ can be written as

$$\begin{aligned}\hat{V}(\hat{\theta}) &= \frac{1}{\hat{N}^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}\left\{y_i-\hat{\theta}-\hat{B}_1\hat{p}_i^{-1}\hat{\pi}_{1i}\hat{\pi}_{2i}\mathbf{h}_i+\hat{B}_2\hat{p}_i^{-1}\hat{\pi}_{1i}\hat{\pi}_{2i}(1-\hat{\pi}_{2i})\mathbf{h}_i\right\}^2 \\ &+ \frac{1}{\hat{N}^2}\sum_{i\in A_1\cup A_2}\frac{\hat{\pi}_{2i}(1-\hat{\pi}_{2i})}{\hat{p}_i}\left\{\hat{B}_1\hat{p}_i^{-1}(1-\delta_{1i})\hat{\pi}_{1i}\mathbf{h}_i-\hat{B}_2\hat{p}_i^{-1}(\hat{\pi}_{1i}-\hat{\pi}_{1i}\hat{\pi}_{2i}+\delta_{1i}\hat{\pi}_{2i})\mathbf{h}_i\right\}^2,\end{aligned} \tag{5.19}$$

where $\hat{N}=\sum_{i\in A_1\cup A_2}\hat{p}_i^{-1}$,

$$(\hat{B}_1,\hat{B}_2)=\hat{N}\times\hat{E}\left\{\frac{\partial U_p(\theta^*,\Phi^*)}{\partial\Phi}\right\}\left[\hat{E}\left\{\frac{\partial S_C(\Phi^*)}{\partial\Phi}\right\}\right]^{-1},$$

$$\hat{E}\left\{\frac{\partial U_p(\theta^*,\Phi^*)}{\partial\Phi}\right\}=-\frac{1}{\hat{N}}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}}(y_i-\hat{\theta})(\mathbf{h}_i',\mathbf{0}_{1\times r}),$$

and

$$\hat{E}\left\{\frac{\partial S_C(\Phi^*)}{\partial\Phi}\right\}=-\sum_{i\in A_1\cup A_2}\hat{p}_i^{-2}\hat{\pi}_{1i}\hat{\pi}_{2i}\begin{pmatrix}(1-\hat{\pi}_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{1i})\mathbf{h}_i\mathbf{h}_i' & (1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i'\end{pmatrix}.$$

We now discuss incorporating auxiliary information in the non-nested two-phase sampling. Incorporating the auxiliary information into the propensity weights for non-nested two-phase sampling has been an area of considerable interest. For example, Zieschang (1990), Renssen and Nieuwenbroek (1997) considered using the generalized regression (GREG) estimator to incorporate both sample and population based auxiliary information under complex sampling design with known inclusion probabilities. Wu (2004) used empirical likelihood (EL) method to incorporate the information and proved asymptotic equivalence of the EL estimator and the GREG estimator. The sample information can be incorporated through the following two-phase regression estimator:

$$\hat{\theta}_{Reg}=\hat{\theta}-\hat{B}_{Reg}(\hat{\mathbf{h}}_1^*-\hat{\mathbf{h}}_2^*), \tag{5.20}$$

where

$$\hat{B}_{Reg}=\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(y_i-\hat{\theta})(\mathbf{h}_i^*-\hat{\mathbf{h}}_1^*)'\left\{\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(\mathbf{h}_i^*-\hat{\mathbf{h}}_1^*)^{\otimes 2}+\sum_{i\in A_2}\hat{\pi}_{2i}^{-1}(\mathbf{h}_i^*-\hat{\mathbf{h}}_2^*)^{\otimes 2}\right\}^{-1},$$

$$\hat{\mathbf{h}}_1^*=\frac{\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}\mathbf{h}_i^*}{\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}},\quad \hat{\mathbf{h}}_2^*=\frac{\sum_{i\in A_2}\hat{\pi}_{2i}^{-1}\mathbf{h}_i^*}{\sum_{i\in A_2}\hat{\pi}_{2i}^{-1}}.$$

with $\mathbf{h}_i^* = (\mathbf{x}_i', y_i)'$. Furthermore, we can incorporate the population information $\bar{\mathbf{X}}_N$ by using the regression estimator:

$$\hat{\theta}^*_{Reg} = \hat{\theta} - \hat{B}_{1,Reg}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*) - \hat{B}_{2,Reg}(\hat{\theta}_{x,1} - \bar{\mathbf{X}}_N), \tag{5.21}$$

where $\hat{\theta}_{x,1} = \sum_{i\in A_1} \hat{\pi}_{1i}^{-1}\mathbf{x}_i / \sum_{i\in A_1} \hat{\pi}_{1i}^{-1}$. The asymptotic properties of $\hat{\theta}_{Reg}$ and $\hat{\theta}^*_{Reg}$ can be derived accordingly with similar arguments. Details are presented in Section B of Appendix D.

5.5 Simulation Study

To test our theory, we performed two simulation studies. In the first simulation, a nested two-phase sampling case was considered and the second simulation deals with non-nested two-phase sampling.

5.5.1 Simulation One

In the first simulation study, we first generated the following finite population of size $N = 10,000$ from the following joint distribution

$$Y_{1i} = 3 + 0.2(X_i - 2) + e_{1i}, \quad Y_{2i} = 3 + 0.2(X_i - 2) + e_{2i},$$

where $X_i \sim N(2, 1)$, and

$$\begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim^{iid} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right], \quad i = 1, 2, \ldots, N.$$

From the finite population, we repeatedly generated two-phase samples with approximate sample size $n_1 = 940$ and $n_2 = 760$ for the phase one and phase two sample, respectively. The sampling indicators δ_{1i} for phase one are generated from Bernoulli(π_{1i}) where logit(π_{1i}) $= -5 + 0.5x_i + 0.5y_{1i}$. The sampling indicators δ_{2i} for phase two are generated from Bernoulli(π_{2i}), where logit(π_{2i}) $= -1 + 0.5x_i + 0.5y_{2i}$, among $\delta_{1i} = 1$. Thus, the simulation setup assumes that the propensity models (5.1) and (5.2) hold. We used $B = 2,000$ Monte Carlo sample size.

From each sample, we computed the following four estimators for $\theta_1 = E(Y_1)$:

1. Naive: Calibration estimator which assumes ignorable missing mechanism;

2. PS: Proposed propensity score estimator, as defined in (5.4);

3. REG: Proposed regression estimator, as defined in (A.1) of Section A;

4. OPT: Proposed optimal estimator, as defined in (A.2) of Section A.

We also computed the following five estimators for $\theta_2 = E(Y_2)$:

1. Naive: Calibration estimator which assumes ignorable missing mechanism;

2. PS: Proposed propensity score estimator, as defined in (5.5);

3. REG: Proposed regression estimator, as defined in (5.11);

4. OPT1: Proposed sample-based optimal estimator, $\hat{\theta}_{2,opt} = \hat{\theta}_2 - \hat{C}_{h_1,opt}(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1})$, where $\hat{\theta}_2$ is defined in (5.5), $\hat{C}_{h_1,opt}$ is defined in (5.15), $\hat{\mathbf{h}}_{1,1}$ and $\hat{\mathbf{h}}_{2,1}$ are defined in (5.11);

5. OPT2: Proposed optimal estimator that incoporates both sample and population level auxiliary information, defined in (A.3) of Section A of Appendix D.

The simulation results for point estimation are given in Table 5.1, which reveals negligible biases for the proposed estimators but significant biases for the naive estimator, which assumes ignorabe missing mechanism. In addition, the optimal estimator achieves the smallest variance which is consistent with the theory. In addition to the point estimation, we also computed variance estimators and computed their relative biases based on our asymptotic theorems in Section 3 and Section A of Appendix D, which are all negligible (less than 10% in absolute values).

5.5.2 Simulation Two

In the second simulation study, a finite population of size $N = 10,000$ was generated from

$$Y_i = 3 + 0.2(X_i - 2) + e_i,$$

where $X_i \sim N(2, 1)$ and $e_i \sim N(0, 1)$. Two independent Bernoulli trails δ_{1i} and δ_{2i} were generated with probability π_{1i} and π_{2i}, respectively, where

$$\pi_{1i} = \Pr(\delta_{1i} = 1 | X_i, Y_i) = \frac{\exp(\phi_0 + \phi_1 X_i + \phi_2 Y_i)}{1 + \exp(\phi_0 + \phi_1 X_i + \phi_2 Y_i)}$$

and

$$\pi_{2i} = \Pr(\delta_{2i} = 1|X_i, Y_i) = \frac{\exp(\phi_0^* + \phi_1^* X_i + \phi_2^* Y_i)}{1 + \exp(\phi_0^* + \phi_1^* X_i + \phi_2^* Y_i)},$$

with $(\phi_0, \phi_1, \phi_2, \phi_0^*, \phi_1^*, \phi_2^*) = (-5, 0.5, 0.5, -4, 0.4, 0.4)$. The approximate sample sizes for A_1 and A_2 are $n_1 = 900$ and $n_2 = 1,200$. We used $B = 2,000$ Monte Carlo samples in the simulation.

From each sample, we computed five estimators of $\theta = E(Y)$.

1. Naive estimator (Naive), $\sum_{i \in A_1} y_i / n_1$;

2. Proposed propensity score estimator (PS), as defined in (5.17);

3. Proposed regression estimator (REG), as defined in (5.20);

4. Optimal estimator (OPT1) that incorporates sample auxiliary information, as defined in (B.4) of Section B in Appendix D;

5. Optimal estimator (OPT2) that incorporates both sample and population auxiliary information, as defined in (B.9) of Section B in Appendix D.

The results for point estimation are in Table 5.2. According to the results, our proposed estimators all have small bias and variance. The naive estimator that ignores design and nonresponse weights has huge biases. The sample-based regression estimator (REG) has smaller variances. The sample-based optimal estimator (OPT1) is more efficient than the original (PS) estimator and regression estimator (REG). The population and sample-based optimal estimator (OPT2) achieves the smallest variances, which is consistent with our asymptotic results in Section 4. Besides point estimation, we also examined the performances of variance estimators based on formulas (5.19) in Section 4, (B.3) and (B.8) in Section B of Appendix D. The relative bias for PS, REG, OPT1 and OPT2 are -0.036, -0.019, -0.019 and -0.062, which verifies the validity of our variance estimators in Section 4.

5.6 Empirical Study

The proposed two-phase propensity score estimator is applied to the data obtained from the 2012 Iowa Caucus survey (ICS). The Iowa political party caucuses are a significant component of the presidential candidate selection process. In 2011, two caucus polls were conducted to be implemented prior to the January 2012 Iowa Republican Caucus. In the first poll, approximately 1200 registered Republicans and Independents (No Party) were interviewed in November of 2011. The second poll is a follow-up poll conducted in December of 2012 with the November respondents to identify changes in their voting preferences.

The sampling frame for the November poll was constructed from the Iowa voter registry provided by the Iowa Secretary of State. The telephone numbers on the list were reported by voters at the time of their registration, and therefore included both landlines and cell phone numbers. A stratified systematic sampling design was used to select the initial sample. Five variables were used to create strata or sorting variables to ensure spread across the range of variation in age, voter activity, geography, gender, and party affiliation. One indicator variable was created to differ voters 35 years or above from younger voters, and a second indicator variable defined whether a voter has attended one or more of the last five primaries. Three additional variables used in designing the sample were congressional district, registered party, and gender.

Strata were defined by party affiliation, congressional district, the age indicator, and the prior primary attendance indicator. Within parties, sample size allocation incorporated an oversampling of primary attendees, in order to maximize the chances of reaching likely Caucus attendees. Sample allocation across the remaining strata was defined in proportion to the number of voters in each stratum. The stratified design was implemented using a systematic probability proportional to size selection scheme. The size measure was based on the relative proportion of voters in each stratum. For each party list, the systematic selection scheme was applied to a list of voters sorted by congressional district, age indicator, previous primary attendance indicator, and gender.

A sample of 9,000 voters was selected for the November poll, consisting of 6,000 Republicans and 3,000 Independents. Telephone numbers were unavailable for 836 of the sampled voters. The remaining 8,164 sample households were contacted. Excluding 190 non-eligible numbers, 1,256 registered voters were finally interviewed from the November poll, which leads to a 15.8% response rate. The November survey of registered Republicans and Independents contained questions related to anticipated caucus attendance, candidates of choice, and opinions on candidate characteristics, as well as demographic and background items. In the December poll, 1,256 respondents from the November poll were contacted again for a follow-up survey and 940 interviews were completed, leading to 74.9 % response rate. Figure 1 summarizes the two-phase sampling structure of the 2012 Iowa caucus survey.

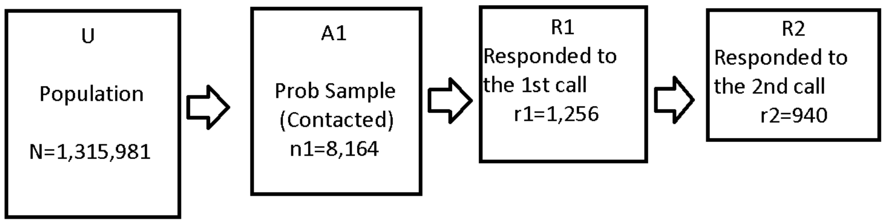

Figure 5.1 Sample structure of 2012 Iowa Caucus Survey

To apply our proposed method to the ICS data, let Y be the reported value of the "First Choice" candidate. After preliminary analyses, we decided to use $X = (Party, Age)$ as the auxiliary variable in the propensity model. The auxiliary variable has a known total at the population level and is also related to the survey participation rate. The population size is $N = 1,315,981$. Denote DY_1 and DY_2 as the dummy variables of "First Choice" based on the first sample A_1 and the second sample A_2 and let DX be the dummy variables based on X. then the parameters of interest is

$$\theta_1 = \frac{\sum_{i\in U} Z_i DY_{1i}}{\sum_{i\in U} Z_i}$$

and

$$\theta_2 = \frac{\sum_{i\in U} Z_i DY_{2i}}{\sum_{i\in U} Z_i},$$

where Z_i is the indicator of "Caucus Attendance" for unit i. That is, $Z_i = 1$ if "Caucus Attendance = Definitely attend" or "Caucus Attendance = Likely to attend". The outcome of the Iowa Caucus on January 3, 2012 is

$$\theta_0 = (24.5\%, 10.3\%, 21.4\%, 43.7\%) \tag{5.22}$$

for "First Choice" candidate: Romney, Perry, Paul, Others. Note that our parameters θ_1 and θ_2 are not necessarily equal to θ_0, although they may be close for certain candidates.

The propensity model used for proposed estimator is

$$\pi_{1i}(\phi) = \frac{\exp(\phi_0 + \phi_1' DX_i + \phi_2' DY_{1i})}{1 + \exp(\phi_0 + \phi_1' DX_i + \phi_2' DY_{1i})} \tag{5.23}$$

and

$$\pi_{2i}(\phi^*) = \frac{\exp(\phi_0^* + \phi_1' DX_i + \phi_2' DY_{2i})}{1 + \exp(\phi_0^* + \phi_1' DX_i + \phi_2' DY_{2i})}, \tag{5.24}$$

where

$$DX_i = (DX_{1i}, DX_{2i})', \quad DY_{1i} = (DY_{11i}, DY_{12i}, DY_{13i}, DY_{14i}, DY_{15i})'$$

and

$$DY_{2i} = (DY_{21i}, DY_{22i}, DY_{23i}, DY_{24i}, DY_{25i})'.$$

Using the proposed methods in Section 3, we obtain parameter estimates for the selection model. The estimated parameters are given in Table 5.3. Table 5.3 shows that variables DX_1, DX_2 and DY_{11} have significant effects on the selection mechanisms, which supports our model for nonignorable sample selection.

We consider three estimators for estimating θ_t for $t = 1, 2$: (i) Naive estimator (Naive) based on the respondents, computed by $\hat{\theta}_{tN} = \sum_{i \in A_t} Z_i DY_{ti} / (\sum_{i \in A_t} Z_i)$, ($ii$) Ignorable-response estimator (Ignorable), computed by $\hat{\theta}_{tIE} = \sum_{i \in A_t} \omega_{ti} Z_i DY_{ti} / (\sum_{i \in A_t} \omega_{ti} Z_i)$, where ω_{ti} is the propensity score obtained by assuming ignorable adjustment weight which is obtained by setting $\phi_2 = 0$ in the sample selection model, and (iii) the proposed propensity score estimator using non-ignorable sample selection models in (5.23) and (5.24). The proposed propensity score estimators are computed by (5.4) and (5.5).

The results for point estimation are given in Table 5.4. The proposed estimates are closer to the Iowa Caucus results in (5.22) than the other estimates for Romney and Perry. Furthermore, the proposed method enables us to compute the estimated standard errors of the point estimates using the theory discussed in Section 3.

5.7 Concluding Remarks

Estimator from voluntary samples can suffer from selection bias. Propensity score weighting using demographic variables can reduce selection bias, bias may remain important if survey participation depends on the study variable itself. We make assumptions about the selection mechanism that explicitly includes the study variable in the selection model. To estimate the model parameters, we propose obtaining a second survey from the original voluntary sample. If the second survey has similar questions as the first one, we assume that the regression coefficients for the explanatory variables in the propensity model are the same as for the original sample. The propensity model is then identified and the model parameters can be estimated using generalized method of moments. When the two samples are not nested and are obtained independently, the theory of capture-recapture sampling can be used to estimate the parameters.

The proposed method provides a useful tool for analyzing voluntary samples, and in particular, web-based panel surveys. In a panel survey, the same sample can be contacted several times and the proposed two-phase estimation approach can be extended to multi-phase estimation. This is a topic of future study.

Table 5.1 Simulation results of the point estimators for θ_1 and θ_2 in Simulation One.

Parameter	Method	Bias($\times 10^2$)	SE($\times 10^2$)	RMSE($\times 10^2$)
θ_1	Naive	43.7	4.00	43.9
	PS	-0.518	13.3	13.3
	REG	0.073	16.2	16.2
	OPT	-0.020	12.8	12.8
θ_2	Naive	44.0	4.28	44.2
	PS	-0.630	12.2	12.2
	REG	-0.370	11.9	11.9
	OPT1	0.572	11.5	11.5
	OPT2	0.829	11.2	11.2

Table 5.2 Simulation results of the point estimators for θ in Simulation Two.

Method	Bias($\times 10^2$)	SE($\times 10^2$)	RMSE($\times 10^2$)
Naive	52.6	2.92	52.6
PS	-0.351	8.96	8.97
REG	-0.171	8.40	8.40
OPT1	0.968	8.22	8.28
OPT2	1.409	7.81	7.93

Table 5.3 Estimated coefficients in the propensity model

Coefficient	Age	Party	Romley	Perry	Paul	Others
Est	0.588	0.782	0.991	0.454	0.866	1.307
S.E.	0.266	0.251	0.454	0.663	0.841	0.985
t.value	2.211	3.116	2.183	0.685	1.030	1.327

Table 5.4 Estimated parameters (s.e.) for 2012 Iowa Caucus Survey Results

Survey	Method	Romney	Perry	Paul	Others
Nov.	Naive	0.340	0.108	0.130	0.422
	Ignorable	0.316	0.103	0.146	0.435
	Proposed	0.303	0.106	0.093	0.499
		(0.062)	(0.039)	(0.107)	(0.046)
Dec.	Naive	0.281	0.140	0.131	0.448
	Ignorable	0.270	0.144	0.148	0.437
	Proposed	0.244	0.134	0.112	0.509
		(0.043)	(0.026)	(0.046)	(0.036)

CHAPTER 6. FUTURE RESEARCH TOPICS

Here is a brief description of the research topics that I have been working but was not able to finish in time. These topics will be pursued in the future.

6.1 Jackknife empirical likelihood for inference with imputed data

Missing data occurs very frequently in social science, survey sampling and other fields. Simply ignoring the missing values may lead to biased inference. Little and Rubin (2002) provided a comprehensive review on the missing data problems. There are two main approaches for inference under missing data. The first one is propensity score weighting approach, which requires correctly specifying the response mechanism model. The second method is imputation, which assumes correct outcome regression model. Propensity score method has been studied in Kim and Kim (2007) and Kim and Riddles (2012), among others. For imputation approach, Kim and Rao (2009) provided unified linearization approach. Their approach leads to wald-type confidence region. Cheng (1994) proposed nonparametric imputaion based on kernel smoothing for ignorable data. Recently, Kim and Yu (2011) proposed using the kernel smoothing method to deal with non-ignorable missing data problem, however, construction of confidence interval has not been well developed yet.

Wang and Rao (2002), Wang and Chen (2009) used empirical likelihood (EL) method to construct likelihood ratio-based confidence interval for the mean functionals under missing at random (MAR) assumption in the sense of Rubin (1976). However, likelihood ratios in their papers converge to a scaled Chi-squred distribution instead of the standard Chi-squred distribution, hence, we need to estimate the scale factor for the inference, which may be cumbersome. The jackknife empirical likelihood (JEL) proposed by Jing et al. (2009) combines two powerful nonparametric tools, EL and Jackknife, for inference when we use one sample and two sample U-statistics . We propose using jackknife empirical likelihood (JEL) method for inference with deterministic imputation under MAR assumption. Under the nonignorable missing mechanism, we can extend Kim and Yu (2011)'s work and still use JEL ratio-based inferences. For the JEL method, the likelihood ratio converge to standard Chi-squred distribution. The EL-based confidence interval has several advantages over normal approximation (NA)-based interval. First of all, EL-based intervals do not have a predetermined shape, but NA-based intervals have symmetric intervals. Secondly, it respect the range of the parameter and transformation respecting. Thirdly, it may have better coverage rates than NA-based confidence intervals.

To explain the setup, consider a independently identically distributed copies $(X_i, Y_i, r_i), (i = 1, 2, \ldots, n)$ from an infinite population. The study variable Y_i is subject to missingness and X_i is always observed. The response indicator function r_i equals to one if Y_i is observed and zero otherwise. The response mechanism can be either ignorable or nonignorable. For simplicity, suppose the parameter of interest is $\theta_0 = E(Y)$. A consistent estimator of θ_0 can be written as

$$\hat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i Y_i + (1 - r_i)\hat{E}(Y_i|X_i, r_i = 0)\right\}, \tag{6.1}$$

where $\hat{E}(Y_i|X_i, r_i = 0)$ is a consistent estimator of $E(Y_i|X_i, r_i = 0)$. We can use either parametric model, such as regression imputation or nonparametric model, such as kernel smoothing method for the estimation. Nextly, I will describe the basic idea of JEL proposed by Jing et al. (2009). Let $Z_1, \ldots, Z_n$ be independent (may not be identically distributed) r.v.'s. Let

$$T_n = T(Z_1, \ldots, Z_n)$$

be a consistent estimator of the parameter θ. Define the jackknife pseduo-values by

$$\hat{V}_i = nT_n - (n-1)T_{n-1}^{(-i)}, \tag{6.2}$$

where $T_{n-1}^{(-i)} = T(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. JEL estimator $\hat{\theta}_{JEL}$ proposed by Jing et al. (2009) can be obtained by maximizing

$$l_e = \sum_{i=1}^{n}\log(p_i), \tag{6.3}$$

subject to

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \hat{V}_i = \theta. \tag{6.4}$$

Jing et al. (2009) proved for one sample or two sample U-statistics T_n, under certain regularity conditions, we have

$$2\left\{l_e(\hat{\theta}_{JEL}) - l_e(\theta_0)\right\} \to^d \chi_1^2. \tag{6.5}$$

We propose using the following T_n in (6.2) for inference with imputed data

$$T_n = T(Z_1, \ldots, Z_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i Y_i + (1 - r_i)\hat{E}(Y_i|X_i, r_i = 0)\right\}, \tag{6.6}$$

where $Z_i = (X_i, Y_i, r_i)$, $\hat{E}(Y_i|X_i, r_i = 0)$ can be parametric regression imputation or nonparametric regression imputation. We know that T_n is a consistent estimator of the parameter θ. Under mild conditions, it can be shown that (6.5) still holds for this case. Therefore, likelihood ratio type confidence interval can be constructed accordingly. In addition, if we know the population mean of auxiliary variable X, which denotes as μ_x, then the efficiency of our proposed estimator can be improved by incorporating constraint $\sum_{i=1}^{n} p_i X_i = \mu_x$. The JEL method can also be extended to fractional imputation, which is under investigation.

6.2 Nonparametric propensity score estimation

Assume the same setups as section 6.1. The parameter of interest is still $\theta_0 = E(Y)$. Without any missing values, a consistent estimator $\hat{\theta}$ of θ_0 can be written as

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{6.7}$$

If we have missing values on Y, then we can estimate propensity nonparametrically as below

$$\hat{\pi}_h(x) = \frac{\sum_{i=1}^{n} r_i K_h(x, X_i)}{\sum_{i=1}^{n} K_h(x, X_i)}, \tag{6.8}$$

where K_h is the kernel function which satisfies certain conditions, and h is the bandwidth. Using nonparametric estimation of propensity scores have been considered by Hirano et al. (2003) and Cattaneo (2010). Chen et al. (2010), Chen and Tang (2011) considered nonparametric estimation of propensity cores in dual system problems. Specifically, they proved the efficiency gain by seperating the discrete and continuous variables' kernels. Xue (2009) used nonparametric estimation of propensity scores in empirical likelihood method. After estimating π_i, the nonparametric propensity score estimator $\hat{\theta}_{NPS}$ can be written as

$$\hat{\theta}_{NPS} = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i}{\hat{\pi}_h(X_i)} Y_i. \tag{6.9}$$

Assume the following regularity conditions:

(C1) $f(x)$, $\pi(x)$ have bounded partial derivatives with respect to x up to order 2 almost surely.

(C2) The kernel function $K(s)$ is a probability density function such that

1. It is bounded and has compact support.
2. It is symmetric with $\sigma_k^2 = \int s^2 K(s)ds < \infty$.
3. $K(s) \geq c$ for some $c > 0$ in some closed interval centered at zero.

(C3) $nh^2 \to \infty$ and $nh^4 \to 0$.

(C4) $E(Y^2)$ is finite and the density of X decays exponentially fast.

(C5) $1 > \pi(x) > d > 0$ almost surely.

We have the following theorem.

Theorem 6.2.1 *Under the regularity conditions (C1)-(C5), we have*

$$\hat{\theta}_{NPS} = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i Y_i}{\pi(X_i)} + \frac{1}{n}\sum_{i=1}^{n} \left\{1 - \frac{r_i}{\pi(X_i)}\right\} m(X_i) + o_p(n^{-1/2}), \tag{6.10}$$

where $m(X_i) = E(Y_i|X_i)$. Futhermore, we have

$$\sqrt{n}(\hat{\theta}_{NPS} - \theta_0) \to^d N(0, V_{NPS}), \tag{6.11}$$

with $V_{NPS} = V\left[r_i\pi^{-1}(X_i)Y_i + \left\{1 - r_i\pi^{-1}(X_i)\right\} m(X_i)\right]$.

Note that the assumption (C3) can be relaxed to $nh \to \infty$ and $nh^4 \to 0$ if we use

$$\hat{\pi}_h(X_i) = \frac{\sum_{j\neq i} r_j K_h(X_i, X_j)}{\sum_{j\neq i} K_h(X_i, X_j)}.$$

Note that the variance V_{NPS} in (6.11) is called the "semi-parametric lower bound", which has been discussed in Robins et al. (1994) and Chen et al. (2008). A consistent estimator of V_{NPS} can be written as

$$\hat{V}_{NPS} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\eta}_i - \bar{\hat{\eta}}_n)^2,$$

with $\hat{\eta}_i = r_i \hat{\pi}_h^{-1}(X_i)Y_i + \left\{1 - r_i \hat{\pi}_h^{-1}(X_i)\right\}\hat{m}(X_i)$, and

$$\hat{m}(X_i) = \frac{\sum_{j=1}^{n} r_j K_h(X_i, X_j)Y_j}{\sum_{j=1}^{n} r_j K_h(X_i, X_j)}.$$

Alternatively, resampling methods, such as jackknife and bootstrap methods can be used to estimate V_{NPS}. In addition, we can use jackknife empirical likelihood (JEL) method for the inference. Instead of using kernel smoothing method, other nonparametric methods, such as local polynomial regression, splines can also be used, and similar results can be obtained. Theorem 6.2.1 is similar as theory 2.1 in Cheng (1994). Similar results can be obtained under nonignorable missing mechanism by using the setup of Kim and Yu (2011).

6.3 Inference with parametric fractional imputation

We consider the setup of parametric fractional imputation (PFI) proposed by Kim (2011). One of the key result of the PFI method is that the MLE can be obtained by maximizing

$$Q^*(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{M}\omega_{ij}^*(\theta)\log f\left(y_{i,obs}, y_{i,mis}^{*(j)};\theta\right), \tag{6.12}$$

where

$$\omega_{ij}^*(\theta) = \frac{f(y_{i,obs}, y_{i,mis}^{*(j)};\theta)/q(y_{i,obs}, y_{i,mis}^{*(j)};\theta)}{\sum_{k=1}^{M} f(y_{i,obs}, y_{i,mis}^{*(k)};\theta)/q(y_{i,obs}, y_{i,mis}^{*(k)};\theta)}.$$

Writing $Q^*(\theta)$ in (6.12) as

$$Q^*(\theta;\eta) = \sum_{i=1}^{n}\sum_{j=1}^{M}\omega_{ij}^*(\eta)\log f(y_{i,obs}, y_{i,mis}^{*(j)};\theta). \tag{6.13}$$

The MLE can be obtained by the EM-type algorithm

$$\hat{\theta}^{(t+1)} \leftarrow \arg\max_{\theta} Q^*(\theta;\hat{\theta}^{(t)}).$$

Instead of EM algorithm, Newton-type algorithm can also be used. We may develop some theories for Newton-Raphson method that computes the maximum of (6.12). Or, some mathematical programming techniques (such as Geometric programming) can be used to find the maximum of (6.12).

One advantage of the PFI method is to replace the integration over missing values by an weighted summation with imputed values. For example, $\hat{\eta} = \sum_{i=1}^{n} E\{g(Y_i)|y_{i,obs}\}$ can be computed by $\hat{\eta}_{FI} = \sum_{i=1}^{n}\sum_{j=1}^{M} \omega_{ij}^{*} g(y_{i,obs}, y_{i,mis}^{*(j)})$, where $\omega_{ij}^{*} = \omega_{ij}^{*}(\hat{\theta})$ and $\hat{\theta}$ is the MLE obtained from (6.12). In particular, we are interested in computing the observed log-likelihood

$$l_{obs} = \sum_{i=1}^{n} \log\{f_{obs,i}(y_{i,obs};\theta)\} = \sum_{i=1}^{n} \log\left\{\int f(y_i;\theta)dy_{mis,i}\right\}.$$

If we use the idea of fractional imputation, we can express

$$f_{obs,i}(y_{i,obs};\theta) = \frac{\sum_{j=1}^{M} f(y_{i,obs}, y_{i,mis}^{*(j)})/q(y_{i,obs}, y_{i,mis}^{*(j)})}{\sum_{j=1}^{M} 1/q(y_{i,obs}, y_{i,mis}^{*(j)})} = \frac{1}{\sum_{j=1}^{M} \omega_{ij}^{*}(\theta)/f(y_{i,obs}, f_{i,mis}^{*(j)};\theta)}$$

and

$$l_{obs}^{*} = -\sum_{i=1}^{n} \log\left\{\sum_{i=1}^{M} \omega_{ij}^{*}(\theta)/f(y_{i,obs}, y_{i,mis}^{*(j)};\theta)\right\} \tag{6.14}$$

as an approximation of $l_{obs}(\theta)$.

If we are interested in making inference about θ, we can build a likelihood ratio (LR) statistics from $l_{obs}(\theta)$, or from $l_{obs}^{*}(\theta)$. That is, under some regularity conditions, we can show that

$$-2\left\{l_{obs}^{*}(\theta_0) - l_{obs}^{*}(\hat{\theta})\right\} \sim \chi_p^2. \tag{6.15}$$

Also, the model selection criteria, such as AIC or BIC, can be developed from the FI likelihood. To show (6.15), note that we can use the second-order Taylor expansion to obtain

$$l^{*}(\theta_0) \cong l_{obs}^{*}(\hat{\theta}) + \frac{\partial l_{obs}^{*}(\hat{\theta})}{\partial\theta}(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})'\left\{\frac{\partial^2 l_{obs}^{*}(\hat{\theta})}{\partial\theta\partial\theta'}\right\}(\theta_0 - \hat{\theta}).$$

Note that, by the definition of $\hat{\theta}$, we have $\partial l_{obs}^{*}(\hat{\theta})/\partial\theta' = 0$. Also, after some algebra, it can be shown that

$$-\frac{\partial^2 l_{obs}^{*}(\theta)}{\partial\theta\partial\theta'} = -\sum_{i=1}^{n}\sum_{j=1}^{M} \omega_{ij}^{*}(\theta)\dot{S}(\theta;y_{ij}^{*}) - \sum_{i=1}^{n}\sum_{j=1}^{M} \omega_{ij}^{*}(\theta)\left\{S(\theta;y_{ij}^{*}) - \bar{S}_i(\theta)\right\}^{\otimes 2}, \tag{6.16}$$

where $S(\theta; y) = \partial \log f(y;\theta)/\partial\theta$, $\dot{S}(\theta; y) = \partial S(\theta; y)/\partial\theta$ and $\bar{S}_i(\theta) = \sum_{j=1}^{M} \omega_{ij}^* S(\theta; y_{ij}^*)$. Note that, for $M \to \infty$, the right side of (6.16) converges to

$$-\sum_{i=1}^{n} E\left\{\dot{S}(\theta; y_i)|y_{i,obs}\right\} - \sum_{i=1}^{n} V\left\{S(\theta; y_i)|y_{i,obs}\right\}$$

which is equal to the observed information matrix discussed in Louis (1982). Thus, we have

$$-\frac{\partial^2 l_{obs}^*(\hat{\theta})}{\partial\theta\partial\theta'} \to^p \mathcal{I}_{obs}(\theta_0) = \left[V(\hat{\theta})\right]^{-1}$$

and result (6.15) follows.

We can also use a Newton method for computing the MLE from the equality in (6.16). That is, the MLE is computed by

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \left\{I_{obs}^*(\hat{\theta}^{(t)})\right\}^{-1} \bar{S}^*(\hat{\theta}^{(t)})$$

where

$$I_{obs}^*(\theta) = -\sum_{i=1}^{n}\sum_{j=1}^{M} \omega_{ij}^*(\theta)\dot{S}(\theta; y_{ij}^*) - \sum_{i=1}^{n}\sum_{j=1}^{M} \omega_{ij}^*(\theta)\left\{S(\theta; y_{ij}^*) - \bar{S}_i(\theta)\right\}^{\otimes 2}$$

and

$$\bar{S}^*(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{M} \omega_{ij}^*(\theta) S(\theta; y_{ij}^*).$$

APPENDIX A. PROOFS FOR CHAPTER 2

A: Proof of Theorem 2.2.1

To prove (2.9), we need the following lemma.

Lemma A.1 *Assume that*

(A.1) $\hat{Q}(\gamma)$ *converges to* $Q(\gamma)$ *uniformly on the compact set* B *containing* γ_0.

(A.2) $\hat{Q}(\gamma)=0$ *has a unique solution* $\hat{\gamma}$ *and* $Q(\gamma)=0$ *also has a unique solution* γ_0.

(A.3) $\partial Q(\gamma)/\partial\gamma$ *is continuous almost everywhere.*

Then, we have

$$p\lim_{n\to\infty}\hat{\gamma}=\gamma_0$$

Lemma A.1 is similar to Corollary II.2 of Andersen and Gill (1982) and its proof is skipped here.

To prove Theorem 2.2.1, by (C4), we can write

$$\hat{Q}_1(\gamma)\to^p Q_1(\gamma)=E\left\{m\left(\lambda^T U(Z;\theta)\right)U(Z;\theta)\right\}/E\left\{m\left(\lambda^T U(Z;\theta)\right)\right\}$$

and note that, by $m(0)=1$, we have $Q_1(\gamma_0)=0$ where $\gamma_0=(\theta_0,0)$.

Also, since $\hat{Q}_2(\gamma)=n^{-1}\sum_{i=1}^n\omega_i^{-1}(d\omega_i/d\theta)$, we have

$$\begin{aligned}\hat{Q}_2(\gamma) &= n^{-1}\sum_{i=1}^n\frac{1}{\omega_i}\left[(\frac{d\lambda}{d\theta})^T\frac{\partial\omega_i}{\partial\lambda}+\frac{\partial\omega_i}{\partial\theta}\right]\\ &= n^{-1}\sum_{i=1}^n(\frac{d\lambda}{d\theta})^T\left\{\frac{m'(\lambda^TU_i(\theta))U_i(\theta)}{m(\lambda^TU_i(\theta))}-\frac{\sum_{j=1}^n m'(\lambda^TU_i(\theta))U_j(\theta)}{\sum_{j=1}^n m(\lambda^TU_j(\theta))}\right\}\\ &+ n^{-1}\sum_{i=1}^n\left\{\frac{m'(\lambda^TU_i(\theta))\dot{U}(Z_i;\theta)^T\lambda}{m(\lambda^TU_i(\theta))}-\frac{\sum_{j=1}^n m'(\lambda^TU_j(\theta))\dot{U}_j(\theta)^T\lambda}{\sum_{j=1}^n m(\lambda^TU_j(\theta))}\right\}\end{aligned} \tag{A.1}$$

where $U_i(\theta) = U(z_i;\theta)$ and $\dot{U}(\theta) = \partial U(z_i;\theta)/\partial\theta$. From (2.7), we have

$$\sum_{i=1}^{n} m'\left\{\lambda^T U_i(\theta)\right\} U_i(\theta) \left\{U_i(\theta)^T \left(\frac{\partial\lambda}{\partial\theta}\right) + \lambda^T \dot{U}_i(\theta)\right\} + \sum_{i=1}^{n} m\left\{\lambda^T U_i(\theta)\right\} \dot{U}_i(\theta) = 0$$

and so

$$\begin{aligned} \frac{\partial\lambda}{\partial\theta} &= -\left\{\sum_{i=1}^{n} m'\left\{\lambda^T U_i(\theta)\right\} U_i(\theta)^{\otimes 2}\right\}^{-1} \\ &\times \left\{\sum_{i=1}^{n} m'\left\{\lambda^T U_i(\theta)\right\} U_i(\theta)\lambda^T \dot{U}_i(\theta) + \sum_{i=1}^{n} m\left\{\lambda^T U_i(\theta)\right\} \dot{U}_i(\theta)\right\}, \end{aligned}$$

where $B^{\otimes 2} = BB^T$. For $\lambda = 0$, we can write

$$\frac{\partial\lambda}{\partial\theta} = -\left\{\sum_{i=1}^{n} U_i(\theta)^{\otimes 2}\right\}^{-1} \sum_{i=1}^{n} \dot{U}_i(\theta)$$

which is bounded in probability. Thus, we can use (C4) to get

$$\hat{Q}_2(\gamma) \to^p Q_2(\gamma)$$

where

$$Q_2(\gamma) = \left[E\left\{\frac{m'(\lambda^T U)U}{m(\lambda^T U)}\right\} - \frac{E\left\{m'(\lambda^T U)U\right\}}{E\left\{m(\lambda^T U)\right\}}\right] S(\theta) + E\left[\frac{m'(\lambda^T U)\lambda\dot{U}}{E\left\{m(\lambda^T U)\right\}}\right] - \frac{E\left\{m'(\lambda^T U)\lambda\dot{U}\right\}}{E\left\{m(\lambda^T U)\right\}}$$

where $S(\theta) = p\lim d\lambda/d\theta$. Since $m(0) = m'(0) = 1$, we have $Q_2(\theta, 0) = 0$ for any θ. Thus, $\gamma_0 = (\theta_0, 0)$ is a unique solution to $Q_1(\gamma) = 0$ and $Q_2(\gamma) = 0$. Therefore, by Lemma A.1, we prove (2.9).

B: Proof of Theorem 2.2.2

Since $\hat{Q}_1(\theta,\lambda) = n^{-1}\sum_{i=1}^{n} \omega\left\{\lambda' U(z_i;\theta)\right\} U(z_i;\theta)$, we have

$$\hat{Q}_1(\theta_0, 0) = \frac{1}{n}\sum_{i=1}^{n} U(Z_i;\theta_0), \quad \frac{\partial\hat{Q}_1(\theta_0,0)}{\partial\theta} = \frac{1}{n}\sum_{i=1}^{n} \dot{U}(Z_i;\theta_0)$$

and

$$\frac{\partial\hat{Q}_1(\theta_0,0)}{\partial\lambda} = \frac{1}{n}\sum_{i=1}^{n}(U(Z_i;\theta_0) - \bar{U}_n)(U(Z_i;\theta_0) - \bar{U}_n)^T$$

where $\dot{U}(Z_i;\theta) = \partial U(Z_i;\theta)/\partial\theta$. Also, by (A.1) and using $m(0) = m'(0) = 1$, it can be shown that

$$\hat{Q}_2(\theta_0, 0) = 0, \quad \frac{\partial\hat{Q}_2(\theta_0,0)}{\partial\theta} = 0, \quad \frac{\partial\hat{Q}_2(\theta_0,0)}{\partial\lambda} = \frac{1}{n}\sum_{i=1}^{n} \dot{U}(Z_i;\theta_0).$$

By (2.9), we can apply the standard arguments using Taylor expansion to get

$$0 = \hat{Q}_1(\hat{\theta}, \hat{\lambda}) = \hat{Q}_1(\theta_0, 0) + \frac{\partial \hat{Q}_1(\theta_0, 0)}{\partial \theta^T}(\hat{\theta} - \theta_0) + \frac{\partial \hat{Q}_1(\theta_0, 0)}{\partial \lambda^T}(\hat{\lambda} - 0) + o_p(\delta_n)$$

and

$$0 = \hat{Q}_2(\hat{\theta}, \hat{\lambda}) = \hat{Q}_2(\theta_0, 0) + \frac{\partial \hat{Q}_2(\theta_0, 0)}{\partial \theta^T}(\hat{\theta} - \theta_0) + \frac{\partial \hat{Q}_2(\theta_0, 0)}{\partial \lambda^T}(\hat{\lambda} - 0) + o_p(\delta_n)$$

where $\delta_n = ||\hat{\theta} - \theta_0|| + ||\hat{\lambda}||$. Thus, we have

$$\begin{pmatrix} \hat{\lambda} - 0 \\ \hat{\theta} - \theta_0 \end{pmatrix} = -\mathbf{S}_n^{-1} \begin{pmatrix} \hat{Q}_1(\theta_0, 0) + o_p(\delta_n) \\ \hat{Q}_2(\theta_0, 0) + o_p(\delta_n) \end{pmatrix}.$$

where

$$\begin{aligned} \mathbf{S}_n &= \begin{pmatrix} \partial \hat{Q}_1(\theta_0, 0)/\partial \lambda & \partial \hat{Q}_1(\theta_0, 0)/\partial \theta \\ \partial \hat{Q}_2(\theta_0, 0)/\partial \lambda & \partial \hat{Q}_2(\theta_0, 0)/\partial \theta \end{pmatrix} \\ &= \begin{pmatrix} n^{-1}\sum_{i=1}^n (U(Z_i;\theta_0) - \bar{U}_n)(U(Z_i;\theta_0) - \bar{U}_n)^T & n^{-1}\sum_{i=1}^n \dot{U}(Z_i;\theta_0) \\ n^{-1}\sum_{i=1}^n \dot{U}(Z_i;\theta_0)^T & 0 \end{pmatrix}. \end{aligned}$$

Because of the existence of moments, we can obtain

$$\mathbf{S_n} \to^p \begin{pmatrix} Var(U(Z;\theta_0)) & E[\dot{U}(Z;\theta_0)] \\ E[\dot{U}(Z;\theta_0)]^T & 0 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{pmatrix}. \tag{B.1}$$

Since $\hat{Q}_1(\theta_0, 0) = n^{-1}\sum_{i=1}^n U(Z_i;\theta_0) = O_p(n^{-\frac{1}{2}})$ and $\hat{Q}_2(\theta_0, 0) = 0$, we have $\delta_n = O_p(n^{-\frac{1}{2}})$ and

$$\begin{pmatrix} \hat{\lambda} - 0 \\ \hat{\theta} - \theta_0 \end{pmatrix} = -\begin{pmatrix} E(UU^T) & E(\dot{U}) \\ E(\dot{U})^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \hat{Q}_{1n}(\theta_0, 0) \\ \hat{Q}_{2n}(\theta_0, 0) \end{pmatrix} + o_p(\delta_n). \tag{B.2}$$

Thus,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -S_{22.1}^{-1} S_{21} S_{11}^{-1} \sqrt{n} \hat{Q}_1(\theta_0, 0) + o_p(1) \to^d N(0, V_1). \tag{B.3}$$

where

$$V_1 = S_{22.1}^{-1} = \{E(\frac{\partial U}{\partial \theta})^T (EUU^T)^{-1} E(\frac{\partial U}{\partial \theta})\}^{-1}.$$

Similarly, we have

$$\sqrt{n}(\hat{\lambda} - 0) = S_{11}^{-1}(S_{12} S_{22.1}^{-1} S_{21} S_{11}^{-1} - I)\sqrt{n}\hat{Q}_1(\theta_0, 0) + o_p(1) \to^d N(0, V_2), \tag{B.4}$$

where

$$V_2 = [E(UU^T)]^{-1}\{I - E(\frac{\partial U}{\partial \theta})V_1 E(\frac{\partial U}{\partial \theta})^T [E(UU^T)]^{-1}\}.$$

Also, ignoring the smaller order terms,

$$Cov\left\{\sqrt{n}(\hat{\theta}-\theta_0), \sqrt{n}(\hat{\lambda}-0)\right\} \quad = \quad S_{22.1}^{-1} S_{21} S_{11}^{-1} S_{11} \left\{S_{11}^{-1}(S_{12} S_{22.1}^{-1} S_{21} S_{11}^{-1} - I)\right\}^T = 0.$$

Therefore, (2.10) follows.

C: Proof of Theorem 2.2.3

The generalized log-empirical likelihood ratio test statistic is

$$W(\theta_0) = 2\left[\sum_{i=1}^{n} \log\{w_i(\hat{\theta}, \hat{\lambda})\} - \sum_{i=1}^{n} \log\{w_i(\theta_0, \lambda_0)\}\right].$$

where $\lambda_0 = \hat{\lambda}(\theta_0)$ is the unique solution to

$$\hat{Q}_1(\theta_0, \lambda) \equiv \sum_{i=1}^{n} \omega\left\{\lambda^T U(z_i;\theta_0)\right\} U(z_i;\theta_0) = 0.$$

Because $\hat{Q}_1(\theta_0, \lambda)$ converges uniformly in probability to

$$Q_1(\theta_0, \lambda) = E\left[m\left\{\lambda^T U(Z;\theta_0)\right\} U(Z;\theta_0)\right] / E\left[m\left\{\lambda^T U(Z;\theta_0)\right\}\right]$$

and $\lambda = 0$ is the unique solution to $Q_1(\theta_0, \lambda) = 0$, we can apply Lemma A.1 to get

$$p \lim_{n\to\infty} \lambda_0 \to 0. \tag{C.1}$$

Thus, we can apply the standard argument using Taylor expansion to get

$$\begin{aligned} 0 &= \hat{Q}_1(\theta_0, \lambda_0) \\ &= \hat{Q}_1(\theta_0, 0) + [\frac{\partial \hat{Q}_1(\theta_0, 0)}{\partial \lambda}]\lambda_0 + o_p(\frac{1}{\sqrt{n}}) \end{aligned}$$

which implies that

$$\lambda_0 = -S_{11}^{-1}\hat{Q}_1(\theta_0, 0) + o_p(1), \tag{C.2}$$

where S_{11} is defined in (B.1).

Now, applying two-dimensional Taylor expansion to $l(\hat{\theta}, \hat{\lambda})$ in (2.8) around $(\theta, \lambda) = (\theta_0, 0)$, we have

$$\begin{aligned} l(\hat{\theta}, \hat{\lambda}) &= l(\theta_0, 0) + \frac{\partial l(\theta_0, 0)}{\partial \theta}(\hat{\theta} - \theta_0) + \frac{\partial l(\theta_0, 0)}{\partial \lambda}(\hat{\lambda} - 0) \\ &+ \frac{1}{2}\left\{(\hat{\theta} - \theta_0)^T \frac{\partial^2 l(\theta_0, 0)}{\partial \theta \partial \theta^T}(\hat{\theta} - \theta_0) + 2(\hat{\theta} - \theta_0)^T \frac{\partial^2 l(\theta_0, 0)}{\partial \theta \partial \lambda^T}(\hat{\lambda} - 0)\right. \\ &+ \left.(\hat{\lambda} - 0)^T \frac{\partial^2 l(\theta_0, 0)}{\partial \lambda \partial \lambda^T}(\hat{\lambda} - 0)\right\} + o_p(1). \end{aligned}$$

After some algebra, it can be shown that

$$\frac{\partial l(\theta_0, 0)}{\partial \theta} = 0, \quad \frac{\partial l(\theta_0, 0)}{\partial \lambda} = 0.$$

$$\frac{\partial^2 l(\theta_0, 0)}{\partial \theta \partial \theta^T} = 0, \quad \frac{\partial^2 l(\theta_0, 0)}{\partial \theta \partial \lambda^T} = 0.$$

$$\frac{\partial^2 l(\theta_0, 0)}{\partial \lambda \partial \lambda^T} = -\sum_{i=1}^{n} U_i U_i^T + \frac{1}{n}[\sum_{j=1}^{n} U_j][\sum_{j=1}^{n} U_j]^T.$$

Hence, we have

$$l(\hat{\theta}, \hat{\lambda}) = \sum_{i=1}^{n} log(\frac{1}{n}) + \frac{1}{2}(\hat{\lambda} - 0)^T \left\{\frac{\partial^2 l(\theta_0, 0)}{\partial \lambda \partial \lambda^T}\right\}(\hat{\lambda} - 0) + o_p(1). \tag{C.3}$$

Using

$$-\frac{1}{n}\frac{\partial^2 l(\theta_0, 0)}{\partial \lambda \partial \lambda^T} = \frac{1}{n}\sum_{i=1}^{n}(U_i - \bar{U}_n)(U_i - \bar{U}_n)^T \to^p Var(U) = E[UU^T] \tag{C.4}$$

and by (B.4), we have

$$l(\hat{\theta}, \hat{\lambda}) = \sum_{i=1}^{n} log(\frac{1}{n}) - \frac{1}{2}[\sqrt{n}\hat{Q}_1(\theta_0, 0)]^T[S_{11}^{-1} - S_{11}^{-1}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1}][\sqrt{n}\hat{Q}_1(\theta_0, 0)] + o_p(1). \tag{C.5}$$

Similarly, by one dimensional Taylor expansion, we can obtain

$$l(\theta_0, \lambda_0) = l(\theta_0, 0) + \frac{\partial l(\theta_0, 0)}{\partial \lambda}(\lambda_0 - 0) + \frac{1}{2}(\lambda_0 - 0)^T \frac{\partial^2 l(\theta_0, 0)}{\partial \lambda \partial \lambda^T}(\lambda_0 - 0) + o_p(1). \tag{C.6}$$

Because $\partial l(\theta_0, 0)/\partial \lambda = 0$, and by (C.2) and (C.4), we have

$$l(\theta_0, \lambda_0) = \sum_{i=1}^{n} log(\frac{1}{n}) - \frac{1}{2}[\sqrt{n}\hat{Q}_1(\theta_0, 0)]^T S_{11}^{-1}[\sqrt{n}\hat{Q}_1(\theta_0, 0)] + o_p(1). \tag{C.7}$$

Thus, by (C.5) and (C.7), we have

$$\begin{aligned} 2[l(\hat{\theta}, \hat{\lambda}) - l(\theta_0, \lambda_0)] &= [\sqrt{n}\hat{Q}_1(\theta_0, 0)]^T[S_{11}^{-1}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1}]\sqrt{n}\hat{Q}_1(\theta_0, 0) + o_p(1) \\ &= [(S_{11}^{-1/2})\sqrt{n}\hat{Q}_1(\theta_0, 0)]^T[(S_{11})^{-1/2}S_{12}S_{22.1}^{-1}S_{21}(S_{11})^{-1/2}] \\ &\times [(S_{11})^{-1/2}\sqrt{n}\hat{Q}_1(\theta_0, 0)] + o_p(1). \end{aligned}$$

Note that $(S_{11})^{-1/2}\sqrt{n}\hat{Q}_1(\theta_0, 0)$ converges to a standard multivariate normal distribution and that $(S_{11})^{-1/2}S_{12}S_{22.1}^{-1}S_{21}(S_{11})^{-1/2}$ is symmetric and idempotent, with trace equal to p. Therefore, the generalized empirical likelihood ratio statistic $W(\theta_0)$ converges to χ^2_p.

D: Proof of Corollary 2.2.1

For fixed θ_1^0, define $U_2\,(z;\theta_2) = U(z;\theta_1^0,\theta_2)$ and also define θ_2^0 to be the unique solution to $E\,\{U_2(Z;\theta_2)\} = 0$. Also, define $l_2(\theta_2,\lambda) = l(\theta_1^0,\theta_2,\lambda)$, where $l(\theta_1,\theta_2,\lambda)$ is the semiparametric log-likelihood function defined in (2.8). Let $\hat{\theta}_2^0 = \hat{\theta}_2(\theta_1^0)$ be the solution that maximizes $l_2(\hat{\theta}_2,\lambda)$ where λ satisfies

$$\sum_{i=1}^{n}\omega_i(\theta_1^0,\theta_2,\lambda)U_2(z_i;\theta_2) = 0.$$

Let

$$\tilde{Q}_1(\theta_2,\lambda) = \sum_{i=1}^{n}\omega_i(\theta_2,\lambda)U_i, \quad \tilde{Q}_2(\theta_2,\lambda) = \frac{1}{n}\frac{dl_2(\theta_2,\lambda)}{d\theta_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega_i}[\frac{\partial\omega_i}{\partial\lambda}\frac{d\lambda}{d\theta_2}+\frac{\partial\omega_i}{\partial\theta_2}].$$

Under this setup, we can use

$$0 = \tilde{Q}_1(\hat{\theta}_2^0,\hat{\lambda}) = \tilde{Q}_1(\theta_2^0,0) + \frac{\partial\tilde{Q}_1(\theta_2^0,0)}{\partial\theta_2^T}(\hat{\theta}_2^0-\theta_2^0) + \frac{\partial\tilde{Q}_1(\theta_2^0,0)}{\partial\lambda^T}(\hat{\lambda}-0) + o_p(\delta_n)$$

and

$$0 = \tilde{Q}_2(\hat{\theta}_2^0,\hat{\lambda}) = \tilde{Q}_2(\theta_2^0,0) + \frac{\partial\tilde{Q}_2(\theta_2^0,0)}{\partial\theta_2^T}(\hat{\theta}_2^0-\theta_2^0) + \frac{\partial\tilde{Q}_2(\theta_2^0,0)}{\partial\lambda^T}(\hat{\lambda}-0) + o_p(\delta_n)$$

where $\delta_n = ||\hat{\theta}_2^0-\theta_2^0|| + ||\hat{\lambda}||$. Thus, we have

$$\begin{pmatrix}\hat{\lambda}-0\\ \hat{\theta}_2^0-\theta_2^0\end{pmatrix} = -\tilde{\mathbf{S}}_n^{-1}\begin{pmatrix}\tilde{Q}_1(\theta_2^0,0)+o_p(\delta_n)\\ \tilde{Q}_2(\theta_2^0,0)+o_p(\delta_n)\end{pmatrix}.$$

where

$$\begin{aligned}\tilde{\mathbf{S}}_n &= \begin{pmatrix}\partial\tilde{Q}_1(\theta_2^0,0)/\partial\lambda & \partial\tilde{Q}_1(\theta_2^0,0)/\partial\theta_2\\ \partial\tilde{Q}_2(\theta_2^0,0)/\partial\lambda & \partial\tilde{Q}_2(\theta_2^0,0)/\partial\theta_2\end{pmatrix}\\ &= \begin{pmatrix}n^{-1}\sum_{i=1}^{n}(U(Z_i;\theta_0)-\bar{U}_n)(U(Z_i;\theta_0)-\bar{U}_n)^T & n^{-1}\sum_{i=1}^{n}\partial U_i/\partial\theta_2\\ n^{-1}\sum_{i=1}^{n}(\partial U_i/\partial\theta_2)^T & 0\end{pmatrix}.\end{aligned}$$

Because of the existence of moments, we can get

$$\tilde{\mathbf{S}}_{\mathbf{n}} \to^p \begin{pmatrix} Var(U(Z;\theta_0)) & E[\partial U_i/\partial\theta_2] \\ E[\partial U_i/\partial\theta_2]^T & 0 \end{pmatrix} = \begin{pmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ \tilde{S}_{21} & 0 \end{pmatrix}.$$

Thus, we can apply the same argument for deriving (C.3) to get

$$l(\theta_1^0, \hat{\theta}_2^0) = \sum_{i=1}^n log(\frac{1}{n}) - \frac{1}{2}[\sqrt{n}\hat{Q}_1(\theta_0,0)]^T(\tilde{S}_{11}^{-1} - \tilde{S}_{11}^{-1}\tilde{S}_{12}\tilde{S}_{22.1}^{-1}\tilde{S}_{21}\tilde{S}_{11}^{-1})\sqrt{n}\hat{Q}_1(\theta_0,0) + o_p(1).$$

Now, using (C.5), we have

$$\begin{aligned} 2W_2 &= 2[l(\hat{\theta}_1,\hat{\theta}_2) - l(\theta_1^0,\hat{\theta}_2^0)] \\ &= [\sqrt{n}\hat{Q}_1(\theta_0,0)]^T[S_{11}^{-1}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1} - \tilde{S}_{11}^{-1}\tilde{S}_{12}\tilde{S}_{22.1}^{-1}\tilde{S}_{21}\tilde{S}_{11}^{-1}][\sqrt{n}\hat{Q}_1(\theta_0,0)] + o_p(1) \\ &= [S_{11}^{-1/2}\sqrt{n}\hat{Q}_1(\theta_0,0)]^T S_{11}^{-1/2}[S_{12}S_{22.1}^{-1}S_{21} - \tilde{S}_{12}\tilde{S}_{22.1}^{-1}\tilde{S}_{21}]S_{11}^{-1/2}[S_{11}^{-1/2}\sqrt{n}\hat{Q}_1(\theta_0,0)] + o_p(1) \end{aligned}$$

Now,

$$\begin{aligned} S_{12}S_{22.1}^{-1}S_{21} &= (E\frac{\partial U}{\partial\theta})\big[(E\frac{\partial U}{\partial\theta})^T(EUU^T)^{-1}(E\frac{\partial U}{\partial\theta})\big]^{-1}(E\frac{\partial U}{\partial\theta})^T \\ &\geq (E\frac{\partial U}{\partial\theta_1}, E\frac{\partial U}{\partial\theta_2})\begin{pmatrix} 0 & 0 \\ 0 & \big[(E\frac{\partial U}{\partial\theta_2})^T(EUU^T)^{-1}(E\frac{\partial U}{\partial\theta_2})\big]^{-1}\end{pmatrix}(E\frac{\partial U}{\partial\theta_1}, E\frac{\partial U}{\partial\theta_2})^T \\ &= (E\frac{\partial U}{\partial\theta_2})\big[(E\frac{\partial U}{\partial\theta_2})(EUU^T)^{-1}(E\frac{\partial U}{\partial\theta_2})\big]^{-1}(E\frac{\partial U}{\partial\theta_2})^T \\ &= \tilde{S}_{12}\tilde{S}_{22.1}^{-1}\tilde{S}_{21}. \end{aligned}$$

Since $S_{11}^{-1/2}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1/2}$ is an idempotent matrix with rank p, $S_{11}^{-1/2}\tilde{S}_{12}\tilde{S}_{22.1}^{-1}\tilde{S}_{21}S_{11}^{-1/2}$ is idempotent matrix with rank $p-q$. So, we have that $S_{11}^{-1/2}S_{12}S_{22.1}^{-1}S_{21}S_{11}^{-1/2} - S_{11}^{-1/2}\tilde{S}_{12}\tilde{S}_{22.1}^{-1}\tilde{S}_{21}S_{11}^{-1/2}$ is a idempotent matrix with rank q (Rao, 1973, p. 187). Therefore, we have that $2W_2 \to^d \chi_q^2$.

E: Proof of Theorem 2.3.1

Since $\hat{Q}_2(\theta,\lambda) = n^{-1}dl_3/d\theta$, we have, similarly to (A.1),

$$\begin{aligned} \hat{Q}_2(\theta,\lambda) &= \frac{n^{-\alpha-1}}{\alpha+1}\sum_{i=1}^n \omega_i^{-\alpha-1}[(\frac{d\lambda}{d\theta})^T\frac{\partial\omega_i}{\partial\lambda} + \frac{\partial\omega_i}{\partial\theta}] \\ &= \frac{n^{-\alpha-1}}{\alpha+1}\sum_{i=1}^n \omega_i^{-\alpha}(\frac{d\lambda}{d\theta})^T\left\{\frac{m'(\lambda^T U_i(\theta))U_i(\theta)}{m(\lambda^T U_i(\theta))} - \frac{\sum_{j=1}^n m'(\lambda^T U_i(\theta))U_j(\theta)}{\sum_{j=1}^n m(\lambda^T U_j(\theta))}\right\} \\ &+ \frac{n^{-\alpha-1}}{\alpha+1}\sum_{i=1}^n \omega_i^{-\alpha}\left\{\frac{m'(\lambda^T U_i(\theta))\dot{U}(Z_i;\theta)^T\lambda}{m(\lambda^T U_i(\theta))} - \frac{\sum_{j=1}^n m'(\lambda^T U_j(\theta))\dot{U}_j(\theta)^T\lambda}{\sum_{j=1}^n m(\lambda^T U_j(\theta))}\right\}. \end{aligned}$$

Using

$$\frac{\partial \omega_i(\theta_0, 0)}{\partial \lambda} = \frac{U_i}{n} - \frac{\sum_{j=1}^n U_j}{n^2}, \quad \frac{\partial \omega_i(\theta_0, 0)}{\partial \theta} = 0 \tag{E.1}$$

$$\frac{d\lambda(\theta_0, 0)}{d\theta} = -\Big\{\sum_{i=1}^n [U(z_i;\theta_0) - \bar{U}_n][U(z_i;\theta_0) - \bar{U}_n]^T\Big\}^{-1} \sum_{i=1}^n \dot{U}(z_i;\theta_0). \tag{E.2}$$

it can be shown that $\hat{Q}_2(\theta_0, 0) = 0$. By using the similar techniques as in theorem 2.2.1, we can prove that

$$(\hat{\theta}, \hat{\lambda}) \rightarrow^p (\lambda_0, 0).$$

Now, using

$$\begin{aligned}
\frac{\partial \hat{Q}_2(\theta,\lambda)}{\partial \theta} &= \frac{n^{-\alpha-1}}{\alpha+1} \sum_{i=1}^n (-\alpha-1)\omega_i^{-\alpha-2} [(\frac{d\lambda}{d\theta})^T \frac{\partial \omega_i}{\partial \lambda} + \frac{\partial \omega_i}{\partial \theta}](\frac{\partial \omega_i}{\partial \theta})^T \\
&+ \frac{n^{-\alpha-1}}{\alpha+1} \sum_{i=1}^n \omega_i^{-\alpha-1} [(\frac{d\lambda}{d\theta})^T \frac{\partial^2 \omega_i}{\partial \lambda \partial \theta^T} + \frac{\partial \omega_i}{\partial \lambda} \frac{\partial (d\lambda/d\theta)}{\partial \theta} + \frac{\partial^2 \omega_i}{\partial \theta \partial \theta^T}]
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \hat{Q}_2(\theta,\lambda)}{\partial \lambda} &= \frac{n^{-\alpha-1}}{\alpha+1} \sum_{i=1}^n (-\alpha-1)\omega_i^{-\alpha-2} [(\frac{d\lambda}{d\theta})^T \frac{\partial \omega_i}{\partial \lambda} + \frac{\partial \omega_i}{\partial \theta}](\frac{\partial \omega_i}{\partial \lambda})^T \\
&+ \frac{n^{-\alpha-1}}{\alpha+1} \sum_{i=1}^n \omega_i^{-\alpha-1} [(\frac{d\lambda}{d\theta})^T \frac{\partial^2 \omega_i}{\partial \lambda \partial \lambda^T} + \frac{\partial \omega_i}{\partial \lambda} \frac{\partial (d\lambda/d\theta)}{\partial \lambda} + \frac{\partial^2 \omega_i}{\partial \theta \partial \lambda}],
\end{aligned}$$

we have $\partial \hat{Q}_2(\theta,\lambda)/\partial\theta = 0$ and $\partial \hat{Q}_2(\theta,\lambda)/\partial\lambda = n^{-1}\sum_{i=1}^n \dot{U}(Z_i;\theta)$. Thus, (B.2) holds and the asymptotic normality of (2.17) follows similarly.

To prove (2.18), note that

$$\begin{aligned}
\frac{\partial l_3(\theta,\lambda)}{\partial \theta} &= -\frac{1}{\alpha(\alpha+1)} \sum_{i=1}^n [(-\alpha)(n\omega_i)^{-\alpha-1} n \frac{\partial \omega_i}{\partial \theta}] \\
&= \frac{n^{-\alpha}}{\alpha+1} \sum_{i=1}^n \omega_i^{-\alpha-1} \frac{\partial \omega_i}{\partial \theta},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l_3(\theta,\lambda)}{\partial \lambda} &= -\frac{1}{\alpha(\alpha+1)} \sum_{i=1}^n [(-\alpha)(n\omega_i)^{-\alpha-1} n \frac{\partial \omega_i}{\partial \lambda}] \\
&= \frac{n^{-\alpha}}{\alpha+1} \sum_{i=1}^n \omega_i^{-\alpha-1} \frac{\partial \omega_i}{\partial \lambda},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l_3(\theta,\lambda)}{\partial \theta^2} &= \frac{n^{-\alpha}}{\alpha+1} \sum_{i=1}^n [(-\alpha-1)\omega_i^{-\alpha-2} \frac{\partial \omega_i}{\partial \theta}(\frac{\partial \omega_i}{\partial \theta})^T + \omega_i^{-\alpha-1} \frac{\partial^2 \omega_i}{\partial \theta \partial \theta^T}] \\
&= -n^{-\alpha} \sum_{i=1}^n \omega_i^{-\alpha-2} \frac{\partial \omega_i}{\partial \theta}(\frac{\partial \omega_i}{\partial \theta})^T + \frac{n^{-\alpha}}{\alpha+1} \sum_{i=1}^n \omega_i^{-\alpha-1} \frac{\partial^2 \omega_i}{\partial \theta \partial \theta^T}.
\end{aligned}$$

Similarly, we can get

$$\frac{\partial^2 l_3(\theta,\lambda)}{\partial\theta\partial\lambda} = -n^{-\alpha}\sum_{i=1}^{n}\omega_i^{-\alpha-2}\frac{\partial\omega_i}{\partial\theta}(\frac{\partial\omega_i}{\partial\lambda})^T + \frac{n^{-\alpha}}{\alpha+1}\sum_{i=1}^{n}\omega_i^{-\alpha-1}\frac{\partial^2\omega_i}{\partial\theta\partial\lambda^T},$$

$$\frac{\partial^2 l_3(\theta,\lambda)}{\partial\lambda^2} = -n^{-\alpha}\sum_{i=1}^{n}\omega_i^{-\alpha-2}\frac{\partial\omega_i}{\partial\lambda}(\frac{\partial\omega_i}{\partial\lambda})^T + \frac{n^{-\alpha}}{\alpha+1}\sum_{i=1}^{n}\omega_i^{-\alpha-1}\frac{\partial^2\omega_i}{\partial\lambda\partial\lambda^T}.$$

So, using (E.1) and (E.2), and by some derivations, it can be shown that

$$\frac{\partial^2 l_3(\theta_0,0)}{\partial\theta\partial\theta^T} = 0, \quad \frac{\partial^2 l_3(\theta_0,0)}{\partial\theta\partial\lambda} = 0$$

and

$$\frac{\partial^2 l_3(\theta_0,0)}{\partial\lambda\partial\lambda^T} = -\sum_{i=1}^{n}(U_i-\bar{U}_n)(U_i-\bar{U}_n)^T.$$

Hence, similar as the proof of theorem 2.2.3, we can get (2.18).

APPENDIX B. PROOFS FOR CHAPTER 3

A: Regularity Conditions for Theorem 3.3.1

(C1) $\theta_0 \in \Theta$ is the unique solution to $E\{U(X,Y;\theta)\} = 0$, and Θ is compact; $\partial U(\theta)/\partial\theta$ is continuous at each $\theta \in \Theta$ and $E\{\sup_{\theta\in\Theta} ||g(X,Y;\theta)||^{\alpha}\}$ is finite for some $\alpha > 2$, where $g(X,Y;\theta) = (U^T(X,Y;\theta), h^T(X;\theta))^T$.

(C2) The partial derivatives $\partial h(\theta)/\partial\theta$ is continuous functions of θ in the neighborhood of θ_0 almost everywhere.

(C3) $||g(X,Y;\theta)||^3$, $||\partial g(X,Y;\theta)/\partial\theta||$, $||\partial^2 g(X,Y;\theta)/(\partial\theta\partial\theta^{'})||$, are bounded by some integrable function $G(X,Y)$.

(C4) The $p \times p$ matrix $E\{\partial U(X,Y;\theta_0)/\partial\theta\}$ has full column rank p. Also, $Var\{U(X,Y;\theta)\}$ and $E(hh^{'}/\pi)$ are positive definite in the neighborhood of θ_0.

(C5) $\pi(x,y) > d > 0$, $p(x) = E\{\pi(x,y)|x\} \neq 1$ almost surely.

B: Proof of Theorem 3.3.1

To discuss the asymptotic properties of the EL estimator, we write

$$\hat{Q}_1(\theta,\lambda) = \frac{1}{n}\sum_{r_i=1} \frac{\pi_i^{-1}U_i(\theta)}{1+\lambda'\pi_i^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^n h_i(\theta)\right\}}, \tag{B.1}$$

and

$$\hat{Q}_2(\theta,\lambda) = \frac{1}{n}\sum_{r_i=1} \frac{\pi_i^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^n h_i(\theta)\right\}}{1+\lambda'\pi_i^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^n h_i(\theta)\right\}}. \tag{B.2}$$

By using similar argument of Lemma 1 of Qin and Lawless (1994), it can be shown that $(\hat{\theta}_{h1},\hat{\lambda}) \rightarrow^p (\theta_0, 0)$.

To prove the asymptotic normality of $\hat{\theta}_{h1}$, by (B.1) and (B.2), we have

$$\hat{Q}_1(\theta_0,0)=\frac{1}{n}\sum_{r_i=1}\pi_i^{-1}U_i(\theta_0),\quad \hat{Q}_2(\theta_0,0)=\frac{1}{n}\sum_{r_i=1}\pi_i^{-1}\left\{h_i(\theta_0)-n^{-1}\sum_{i=1}^{n}h_i(\theta_0)\right\},$$

$$\frac{\partial\hat{Q}_1(\theta_0,0)}{\partial\theta}=\frac{1}{n}\sum_{r_i=1}\pi_i^{-1}\frac{\partial U_i}{\partial\theta},\quad \frac{\partial\hat{Q}_1(\theta_0,0)}{\partial\lambda}=-\frac{1}{n}\sum_{r_i=1}\frac{1}{\pi_i^2}U_i\left\{h_i(\theta_0)-n^{-1}\sum_{i=1}^{n}h_i(\theta_0)\right\}',$$

$$\frac{\partial\hat{Q}_2(\theta_0,0)}{\partial\theta}=0,\quad \frac{\partial\hat{Q}_2(\theta_0,0)}{\partial\lambda}=-\frac{1}{n}\sum_{r_i=1}\frac{1}{\pi_i^2}\left\{h_i(\theta_0)-n^{-1}\sum_{i=1}^{n}h_i(\theta_0)\right\}^{\otimes 2}.$$

According to conditions (C1)-(C5) and $(\hat{\theta}_{h1},\hat{\lambda})\to^p(\theta_0,0)$, we can apply the standard arguments using Taylor expansion to get

$$0=\hat{Q}_1(\hat{\theta}_{h1},\hat{\lambda})=\hat{Q}_1(\theta_0,0)+\frac{\partial\hat{Q}_1(\theta_0,0)}{\partial\theta'}(\hat{\theta}_{h1}-\theta_0)+\frac{\partial\hat{Q}_1(\theta_0,0)}{\partial\lambda'}(\hat{\lambda}-0)+o_p(\delta_n),$$

and

$$0=\hat{Q}_2(\hat{\theta}_{h1},\hat{\lambda})=\hat{Q}_2(\theta_0,0)+\frac{\partial\hat{Q}_2(\theta_0,0)}{\partial\theta'}(\hat{\theta}_{h1}-\theta_0)+\frac{\partial\hat{Q}_2(\theta_0,0)}{\partial\lambda'}(\hat{\lambda}-0)+o_p(\delta_n),$$

where $\delta_n=||\hat{\theta}_{h1}-\theta_0||+||\hat{\lambda}||$. Thus, we have

$$\begin{pmatrix}\hat{\theta}_{h1}-\theta_0\\ \hat{\lambda}-0\end{pmatrix}=-\mathbf{S}_n^{-1}\begin{pmatrix}\hat{Q}_1(\theta_0,0)+o_p(\delta_n)\\ \hat{Q}_2(\theta_0,0)+o_p(\delta_n)\end{pmatrix},$$

where

$$\mathbf{S}_n=\begin{pmatrix}\partial\hat{Q}_1(\theta_0,0)/\partial\theta & \partial\hat{Q}_1(\theta_0,0)/\partial\lambda\\ \partial\hat{Q}_2(\theta_0,0)/\partial\theta & \partial\hat{Q}_2(\theta_0,0)/\partial\lambda\end{pmatrix}.$$

Because of the existence of moments, we have

$$\mathbf{S_n}\to^p\begin{pmatrix}E(\partial U/\partial\theta) & -E(U\tilde{h}'/\pi)\\ 0 & -E(\tilde{h}\tilde{h}'/\pi)\end{pmatrix}=\begin{pmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{pmatrix}.$$

Since $\hat{Q}_1(\theta_0,0)=n^{-1}\sum_{r_i=1}\pi_i^{-1}U_i=O_p(n^{-\frac{1}{2}})$ and $\hat{Q}_2\left(\theta_0,0\right)=n^{-1}\sum_{r_i=1}\pi_i^{-1}(h_i-n^{-1}\sum_{i=1}^{n}h_i)=O_p(n^{-\frac{1}{2}})$, we have $\delta_n=O_p(n^{-\frac{1}{2}})$ and

$$\begin{pmatrix}\hat{\theta}_{h1}-\theta_0\\ \hat{\lambda}-0\end{pmatrix}=-\begin{pmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{pmatrix}^{-1}\begin{pmatrix}\hat{Q}_1(\theta_0,0)\\ \hat{Q}_2(\theta_0,0)\end{pmatrix}+o_p(\delta_n).$$

So, after some algebra, we have

$$\hat{\theta}_{h1} - \theta_0 \quad = \quad S_{11}^{-1}\left\{-\hat{Q}_1(\theta_0, 0) + S_{12}S_{22}^{-1}\hat{Q}_2(\theta_0, 0)\right\} + o_p(n^{-1/2}).$$

Hence, by the existence of second moments, we get

$$\sqrt{n}(\hat{\theta}_{h1} - \theta_0) \to^d N(0, V_{h1}),$$

where $V_{h1} = \tau\Omega_{h1}\tau'$ with $\tau = \{E(\partial U/\partial\theta)\}^{-1}$, $\Omega_{h1} = V\left\{r\left(U - B\tilde{h}\right)/\pi + B\tilde{h}\right\}$ and $B = E(U\tilde{h}'/\pi)\left\{E(\tilde{h}\tilde{h}'/\pi)\right\}^{-1}$.

C: Proof of Corollary 3.3.1

Note that Ω_{h1} in (3.16) satisfies

$$\begin{aligned}
\Omega_{h1} &= E\left\{rU/\pi - B(r/\pi - 1)\tilde{h}\right\}^{\otimes 2} \\
&= E\left\{rU/\pi - (r/\pi - 1)E(U|X) + (r/\pi - 1)E(U|X) - B(r/\pi - 1)\tilde{h}\right\}^{\otimes 2} \\
&= E\left\{rU/\pi - (r/\pi - 1)E(U|X)\right\}^{\otimes 2} \\
&+ E\left\{(r\pi^{-1}U - (r\pi^{-1} - 1)E(U|X))(r\pi^{-1} - 1)(E(U^T|X) - \tilde{h}^T B^T)\right\} \\
&+ E\left\{(r\pi^{-1} - 1)(E(U|X) - B\tilde{h})(r\pi^{-1}U^T - (r\pi^{-1} - 1)E(U^T|X))\right\} \\
&+ E\left\{(r/\pi - 1)E(U|X) - B(r/\pi - 1)\tilde{h}\right\}^{\otimes 2} \\
&= E\left\{rU/\pi - (r/\pi - 1)E(U|X)\right\}^{\otimes 2} + E\left\{(r/\pi - 1)E(U|X) - B(r/\pi - 1)\tilde{h}\right\}^{\otimes 2} \\
&\geq E\left\{rU/\pi - (r/\pi - 1)E(U|X)\right\}^{\otimes 2},
\end{aligned}$$

where the equality is achieved when $h \propto h^* = E(U|X)$. The asymptotic variance of $\hat{\theta}_{h1}$ achieved at $h \propto h^* = E(U|X)$ is equal to

$$V_{h^*} = \tau\left\{E\left(\frac{UU'}{\pi}\right) - E\left(\frac{1-\pi}{\pi}h^*U'\right)\right\}\tau'.$$

D: Proof of Theorem 3.4.1

To discuss the asymptotic properties of the EL estimator, we write

$$\hat{Q}_1(\theta, \lambda, \phi) = \frac{1}{n}\sum_{r_i=1}\frac{\pi_i(\phi)^{-1}U_i(\theta)}{1 + \lambda'\pi_i(\phi)^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^{n}h_i(\theta)\right\}}, \tag{D.1}$$

and

$$\hat{Q}_2(\theta,\lambda,\phi) = \frac{1}{n}\sum_{r_i=1} \frac{\pi_i(\phi)^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^n h_i(\theta)\right\}}{1+\lambda'\pi_i(\phi)^{-1}\left\{h_i(\theta) - n^{-1}\sum_{i=1}^n h_i(\theta)\right\}}. \tag{D.2}$$

Hence $(\hat{\theta}_{h2},\hat{\lambda})$ is the solution defined by equation $\hat{Q}_1(\hat{\theta}_{h2},\hat{\lambda},\hat{\phi}) = 0$ and $\hat{Q}_2(\hat{\theta}_{h2},\hat{\lambda},\hat{\phi}) = 0$, where $\hat{\phi}$ is defined in (3.18). By using a similar argument as that for Theorem 3.3.1, we can prove $(\hat{\theta}_{h2},\hat{\lambda}) \to^p (\theta_0, 0)$. Next, we want to prove the asymptotic normality of $\hat{\theta}_{h2}$. According to (D.1) and (D.2), we have

$$\hat{Q}_1(\theta_0,0,\phi_0) = \frac{1}{n}\sum_{r_i=1}\pi_i^{-1}U_i(\theta_0), \quad \hat{Q}_2(\theta_0,0,\phi_0) = \frac{1}{n}\sum_{r_i=1}\pi_i^{-1}\left\{h_i(\theta_0) - n^{-1}\sum_{i=1}^n h_i(\theta_0)\right\},$$

$$\frac{\partial \hat{Q}_1(\theta_0,0,\phi_0)}{\partial\theta} = \frac{1}{n}\sum_{r_i=1}\pi_i^{-1}\frac{\partial U_i}{\partial\theta}, \quad \frac{\partial \hat{Q}_1(\theta_0,0,\phi_0)}{\partial\lambda} = -\frac{1}{n}\sum_{r_i=1}\frac{1}{\pi_i^2}U_i\left\{h_i(\theta_0) - n^{-1}\sum_{i=1}^n h_i(\theta_0)\right\}',$$

$$\frac{\partial \hat{Q}_1(\theta_0,0,\phi_0)}{\partial\phi} = -\frac{1}{n}\sum_{r_i=1}\frac{1}{\pi_i^2}\frac{\partial\pi_i}{\partial\phi}U_iU_i', \quad \frac{\partial \hat{Q}_2(\theta_0,0,\phi_0)}{\partial\theta} = 0,$$

$$\frac{\partial \hat{Q}_2(\theta_0,0,\phi_0)}{\partial\lambda} = -\frac{1}{n}\sum_{r_i=1}\frac{1}{\pi_i^2}\left\{h_i(\theta_0) - n^{-1}\sum_{i=1}^n h_i(\theta_0)\right\}^{\otimes 2},$$

$$\frac{\partial \hat{Q}_2(\theta_0,0,\phi_0)}{\partial\phi} = -\frac{1}{n}\sum_{r_i=1}\frac{1}{\pi_i^2}\frac{\partial\pi_i}{\partial\phi}\left\{h_i(\theta_0) - n^{-1}\sum_{i=1}^n h_i(\theta_0)\right\}^{\otimes 2}.$$

By using Taylor expansion around $(\theta_0, 0, \phi_0)$, we have

$$\begin{aligned} 0 &= \hat{Q}_1(\hat{\theta}_{h2},\hat{\lambda},\hat{\phi}) = \hat{Q}_1(\theta_0,0,\phi_0) + \frac{\partial \hat{Q}_1(\theta_0,0,\phi_0)}{\partial\theta'}(\hat{\theta}_{h2}-\theta_0) + \frac{\partial \hat{Q}_1(\theta_0,0,\phi_0)}{\partial\lambda'}(\hat{\lambda}-0) \\ &+ \frac{\partial \hat{Q}_1(\theta_0,0,\phi_0)}{\partial\phi'}(\hat{\phi}-\phi_0) + o_p(\delta_n), \end{aligned}$$

and

$$\begin{aligned} 0 &= \hat{Q}_2(\hat{\theta},\hat{\lambda},\hat{\phi}) = \hat{Q}_2(\theta_0,0,\phi_0) + \frac{\partial \hat{Q}_2(\theta_0,0,\phi_0)}{\partial\theta'}(\hat{\theta}_{h2}-\theta_0) + \frac{\partial \hat{Q}_2(\theta_0,0,\phi_0)}{\partial\lambda'}(\hat{\lambda}-0) \\ &+ \frac{\partial \hat{Q}_2(\theta_0,0,\phi_0)}{\partial\phi'}(\hat{\phi}-\phi_0) + o_p(\delta_n), \end{aligned}$$

where $\delta_n = ||\hat{\theta}_{h2}-\theta_0|| + ||\hat{\lambda}|| + ||\hat{\phi}-\phi_0||$. By using a similar argument as the proof of Theorem 3.3.1 and by (3.18), after some algebra, we have

$$\hat{\theta}_{h2} - \theta_0 = -S_{11}^{-1}n^{-1}\sum_{i=1}^n\left\{\frac{r_i}{\pi_i}U_i(\theta_0) - B(\frac{r_i}{\pi_i}-1)\tilde{h}_i(\theta_0) - Cb_i(\phi_0)\right\} + o_p(n^{-1/2}),$$

where $B = S_{12}S_{22}^{-1}$, $C = E\big[\pi^{-1}(U - B\tilde{h})(\partial\pi/\partial\phi)'\big]$ and $S_{11} = E(\partial U/\partial\theta)$, $S_{12} = -E(U\tilde{h}'/\pi)$ and $S_{22} = -E(\tilde{h}\tilde{h}'/\pi)$. Hence, we have

$$\sqrt{n}(\hat{\theta}_{h2} - \theta_0) \to^d N(0, V_{h2}),$$

where $V_{h2} = S_{11}^{-1}Var\big\{r\pi^{-1}U - B(r\pi^{-1} - 1)\tilde{h} - Cb\big\}S_{11}^{-1}$.

E: Proof of Theorem 3.4.2

Because $\hat{\phi}_q \to^p \phi_0$, under some moment conditions, by using Taylor expansion, we have

$$\hat{\phi}_q - \phi_0 = \frac{1}{n}\sum_{i=1}^{n} b_i(\phi_0) + o_p(n^{-1/2}),$$

where $b_i(\phi_0) = S^{-1}(r_i\pi_i^{-1} - 1)q_i$, and $S = E\left\{\pi^{-1}q(\partial\pi/\partial\phi)'\right\}$. Using the result of Theorem 3.4.1, we can get (3.23). So, under the existence of moments, we have

$$\sqrt{n}(\hat{\theta}_{h2} - \theta_0) \to^d N(0, V_{h2}),$$

where $V_{h2} = \tau\Omega_{h2}\tau'$ and $\Omega_{h2} = V\left\{r\pi^{-1}U - (r\pi^{-1} - 1)(B\tilde{h} + CS^{-1}q)\right\}$. Using a similar argument as the proof of Corollary 1, we have

$$\begin{aligned}
\Omega_{h2} &= V\left\{r\pi^{-1}U - (r\pi^{-1} - 1)(B\tilde{h} + CS^{-1}q)\right\} \\
&= E\left\{r\pi^{-1}U - (r\pi^{-1} - 1)(B\tilde{h} + CS^{-1}q)\right\}^{\otimes 2} \\
&= E\left\{r\pi^{-1}U - (r\pi^{-1} - 1)h^* + (r\pi^{-1} - 1)h^* - (r\pi^{-1} - 1)(B\tilde{h} + CS^{-1}q)\right\}^{\otimes 2} \\
&= E\left\{r\pi^{-1}U - (r\pi^{-1} - 1)h^*\right\}^{\otimes 2} + E\left\{(r\pi^{-1} - 1)h^* - (r\pi^{-1} - 1)(B\tilde{h} + CS^{-1}q)\right\}^{\otimes 2} \\
&\geq E\left\{r\pi^{-1}U - (r\pi^{-1} - 1)h^*\right\}^{\otimes 2}.
\end{aligned}$$

The equality holds when $h^* - B\tilde{h} - CS^{-1}q = 0$, which implies $q = S(C'C)^{-1}C'(h^* - B\tilde{h})$. Hence, the optimal choice of q is $\alpha' q = h^* - B\tilde{h}$ for some α.

F: Proof of Theorem 3.5.1

For simplicity, we assume $q = 2$ in the following proof. In order to prove Theorem 3.5.1, we first prove the following Lemma:

Lemma B.2 *Let* $\hat{\pi}_{i,H}$ *be the kernel estimator of* $\pi(x_i)$ *which is defined in (3.24). For the choice of* $h^* = E(U|X)$*, where* U *is the estimating function defined in (3.1), we have*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_{i,H}}U_i = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i}(U_i - h_i^*) + \frac{1}{n}\sum_{i=1}^{n}h_i^* + o_p(n^{-1/2}). \tag{F.1}$$

Proof. By using the standard arguments in kernel smoothing method, we have

$$E\left\{\frac{1}{n}\sum_{j=1}^{n}K_H(X_i, X_j)\right\} = f(X_i) + O(H^2) \tag{F.2}$$

and

$$E\left\{\frac{1}{n}\sum_{j=1}^{n}r_j K_H(X_i, X_j)\right\} = \pi(X_i)f(X_i) + O(H^2). \tag{F.3}$$

According to (F.2), (F.3) and by using Taylor expansion, we have

$$\begin{aligned}
\frac{\sum_{j=1}^{n}K_H(X_i, X_j)}{\sum_{j=1}^{n}r_j K_H(X_i, X_j)} &= \frac{1}{\pi(X_i)} + \frac{1}{\pi(X_i)f(X_i)}\left\{\frac{1}{n}\sum_{j=1}^{n}K_H(X_i, X_j) - f(X_i)\right\} \\
&- \frac{1}{\pi^2(X_i)f(X_i)}\left\{\frac{1}{n}\sum_{j=1}^{n}r_j K_H(X_i, X_j) - \pi(X_i)f(X_i)\right\} + O(H^2) \\
&= \frac{1}{\pi(X_i)} + \frac{1}{n}\sum_{j=1}^{n}\frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)}\left\{1 - \frac{r_j}{\pi(X_i)}\right\} + O(H^2). \qquad \text{(F.4)}
\end{aligned}$$

By (F.4) and because of $nH^4 \to 0, nH^2 \to \infty$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_{i,H}}U_i &= \frac{1}{n}\sum_{i=1}^{n}r_i\frac{\sum_{j=1}^{n}K_H(X_i, X_j)}{\sum_{j=1}^{n}r_j K_H(X_i, X_j)}U_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{r_i U_i}{\pi(X_i)} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}r_i U_i\frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)}\left\{1 - \frac{r_j}{\pi(X_i)}\right\} + O(H^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{r_i U_i}{\pi(X_i)} + \frac{1}{n^2 H}\sum_{i=1}^{n}U_i\frac{K(0)}{\pi(X_i)f(X_i)}\left\{r_i - \frac{r_i}{\pi(X_i)}\right\} \\
&+ \frac{1}{n(n-1)}\sum_{i\neq j}r_i U_i\frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)}\left\{1 - \frac{r_j}{\pi(X_i)}\right\} + O(H^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{r_i U_i}{\pi(X_i)} + \frac{1}{n(n-1)}\sum_{i\neq j}r_i U_i\frac{K_H(X_i, X_j)}{\pi(X_i)f(X_i)}\left\{1 - \frac{r_j}{\pi(X_i)}\right\} + o_p(n^{-1/2})
\end{aligned}$$

so,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\hat{\pi}_{i,H}}U_i = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i U_i}{\pi(X_i)} + \frac{1}{n(n-1)}\sum_{i\neq j}h(Z_i, Z_j) + o_p(n^{-1/2}), \tag{F.5}$$

where $Z_i = (X_i, Y_i, r_i)$ and

$$\begin{aligned} h(Z_i, Z_j) &= \frac{1}{2}\left[r_i U_i \frac{K_H(X_i, X_j)}{\pi(X_i) f(X_i)}\left\{1 - \frac{r_j}{\pi(X_i)}\right\} + r_j U_j \frac{K_H(X_j, X_i)}{\pi(X_j) f(X_j)}\left\{1 - \frac{r_i}{\pi(X_j)}\right\}\right] \\ &\triangleq \frac{1}{2}(\zeta_{ij} + \zeta_{ji}). \end{aligned} \tag{F.6}$$

According to (F.5) and (F.6), we know that $\sum_{i \neq j} h(Z_i, Z_j)/\{n(n-1)\}$ is the U-statistics. Let $s = (X_j - X_i)/H$, by $nH^2 \to \infty, nH^4 \to 0$ and according to Taylor expansion, we have

$$\begin{aligned} E(\zeta_{ij}|Z_i) &= \frac{r_i U_i}{\pi(X_i) f(X_i)} \frac{1}{H^{d_x}} \int K\left(\frac{X_j - X_i}{H}\right)\left\{1 - \frac{\pi(X_j)}{\pi(X_i)}\right\} f(X_j) dX_j \\ &= \frac{r_i U_i}{\pi(X_i) f(X_i)} \int K(s)\left\{1 - \frac{\pi(X_i + Hs)}{\pi(X_i)}\right\} f(X_i + Hs) ds \\ &= O(H^2), \end{aligned} \tag{F.7}$$

and

$$\begin{aligned} E(\zeta_{ji}|Z_i) &= \frac{1}{H^{d_x}} \int \frac{U_j}{f(X_j)} K\left(\frac{X_j - X_i}{H}\right)\left\{1 - \frac{r_i}{\pi(X_j)}\right\} f(X_j, Y_j) dX_j dY_j \\ &= \int \frac{U_j}{f(X_i + Hs)} K(s)\left\{1 - \frac{r_i}{\pi(X_i + Hs)}\right\} f(X_i + Hs, Y_j) ds dY_j \\ &= \left\{1 - \frac{r_i}{\pi(X_i)}\right\} h_i^* + O(H^2). \end{aligned} \tag{F.8}$$

According to (F.5), (F.6), (F.7), (F.8) and by the theory of U-statistics, see serfling (1980), chapter 5, we have

$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i = \frac{1}{n}\sum_{i=1}^n \frac{r_i U_i}{\pi(X_i)} + \frac{1}{n}\sum_{i=1}^n \left\{1 - \frac{r_i}{\pi(X_i)}\right\} h_i^* + o_p(n^{-1/2}).$$

■

Similarly to Lemma B.2, it can be shown that

$$\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} \tilde{h}_i = \frac{1}{n}\sum_{i=1}^n \tilde{h}_i + o_p(n^{-1/2}). \tag{F.9}$$

By the same argument for proof of Theorem 3.4.1, under certain conditions, it can be shown that $\hat{\theta}_{h3} \to^p \theta_0$, and the asymptotic normality property:

$$\hat{\theta}_{h3} - \theta_0 = S_{11}^{-1}\Big\{ -\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\hat{\pi}_{i,H}} U_i(\theta_0) + B\frac{1}{n}\sum_{i=1}^n (\frac{r_i}{\hat{\pi}_{i,H}} - 1)\tilde{h}_i(\theta_0)\Big\} + o_p(n^{-1/2}),$$

where B and S_{11} are defined in the proof of Theorem 3.4.1. By using (F.1) and (F.9), we have

$$\hat{\theta}_{h3} - \theta_0 = S_{11}^{-1}\Big\{ -\frac{1}{n}\sum_{i=1}^n \frac{r_i}{\pi_i} U_i(\theta_0) + \frac{1}{n}\sum_{i=1}^n (\frac{r_i}{\pi_i} - 1)h_i^*(\theta_0)\Big\} + o_p(n^{-1/2}).$$

G: Proof of Theorem 3.5.2

Let

$$\hat{Q}_3(\theta,\lambda)=\frac{1}{n}\sum_{r_i=1}\frac{\hat{\pi}_{i,H}^{-1}U_i(\theta)}{1+\lambda'\hat{\pi}_{i,H}^{-1}U_i(\theta)}.$$

We can write

$$R_n(\theta_0)=2\left\{\sum_{r_i=1}\log(1+\lambda_0^{'}\hat{\pi}_{i,H}^{-1}U_i(\theta_0))-\sum_{r_i=1}\log(1+\hat{\lambda}^{'}\hat{\pi}_{i,H}^{-1}U_i(\hat{\theta}_{h3}))\right\},$$

where λ_0 is the solution of $\hat{Q}_3(\theta_0,\lambda)=0$, by using a similar argument from Theorem 3.3.1, we have $\lambda_0\rightarrow^p 0$. Thus, by using Taylor expansion around 0, we have

$$0=\hat{Q}_3(\theta_0,\lambda_0)=\hat{Q}_3(\theta_0,0)+\frac{\partial\hat{Q}_3(\theta_0,0)}{\partial\lambda_0}\lambda_0+o_p(||\lambda_0||).$$

According to Lemma B.2 in the proof of Theorem 3.5.1, we have $\hat{Q}_3(\theta_0,0)=n^{-1}\sum_{r_i=1}\hat{\pi}_{i,H}^{-1}U_i(\theta_0)=O_p(n^{-1/2})$, so $||\lambda_0||=O_p(n^{-1/2})$. Hence, we have

$$\lambda_0=-S_{11}^{*-1}\hat{Q}_3(\theta_0,0)+o_p(n^{-1/2}),\qquad\text{(G.1)}$$

where $S_{11}^{*}=-E(UU^{'}/\pi)$. Also, by using Taylor expansion around $\lambda_0=0$, we have

$$2\sum_{r_i=1}\log(1+\lambda_0^{'}\hat{\pi}_{i,H}^{-1}U_i(\theta_0))=2\sum_{r_i=1}\lambda_0^{'}\hat{\pi}_{i,H}^{-1}U_i(\theta_0)-\sum_{r_i=1}\lambda_0^{'}\hat{\pi}_{i,H}^{-2}U_i(\theta_0)U_i^{'}(\theta_0)\lambda_0+o_p(1).\qquad\text{(G.2)}$$

By the existence of moments, we have

$$\partial\hat{Q}(\theta_0,0)/\partial\lambda=-n^{-1}\sum_{r_i=1}\hat{\pi}_{i,H}^{-2}U_i(\theta_0)U_i^{'}(\theta_0)\rightarrow^p S_{11}^{*}.\qquad\text{(G.3)}$$

By plugging (G.1) into (G.2) and according to (G.3), we have

$$2\sum_{r_i=1}\log(1+\lambda_0^{'}\hat{\pi}_{i,H}^{-1}U_i(\theta_0))=-n\hat{Q}_3^{'}(\theta_0,0)S_{11}^{*-1}\hat{Q}_3(\theta_0,0)+o_p(1).\qquad\text{(G.4)}$$

Similarly, by the same argument of Qin and Lawless (1994) and by using Taylor expansion around $\hat{\lambda}=0$, we have

$$\begin{aligned}2\sum_{r_i=1}\log(1+\hat{\lambda}^{'}\hat{\pi}_{i,H}^{-1}U_i(\hat{\theta}))&=-n\hat{Q}_3^{'}(\theta_0,0)S_{11}^{*-1}\hat{Q}_3(\theta_0,0)\\&+n\hat{Q}_3^{'}(\theta_0,0)S_{11}^{*-1}S_{12}^{*}S_{22.1}^{*-1}S_{21}^{*}S_{11}^{*-1}\hat{Q}_3(\theta_0,0)+o_p(1),\end{aligned}\qquad\text{(G.5)}$$

where $S_{12}^* = E\left\{\partial U(\theta_0)/\partial\theta\right\}, S_{21}^* = S_{12}^{*'}$, and $S_{22.1}^* = S_{21}^* S_{11}^{*-1} S_{12}^*$. So, by (G.4) and (G.5), we have

$$R_n(\theta_0) = -n\hat{Q}_3'(\theta_0, 0) S_{11}^{*-1} S_{12}^* S_{22.1}^{*-1} S_{21}^* S_{11}^{*-1} \hat{Q}_3(\theta_0, 0) + o_p(1).$$

In addition, by the existence of moments and $h^* = E\left\{U(\theta_0)|X\right\} = 0$, we have

$$\sqrt{n}\hat{Q}_3(\theta_0, 0) \to^d N(0, V_{Q_3}),$$

where $V_{Q_3} = E(UU'/\pi) = -S_{11}^*$. Hence, we have

$$R_n(\theta_0) = -\sqrt{n}\hat{Q}_3'(\theta_0, 0) S_{11}^{*-1/2} S_{11}^{*-1/2} S_{12}^* S_{22.1}^{*-1} S_{21}^* S_{11}^{*-1/2} \sqrt{n} S_{11}^{*-1/2} \hat{Q}_3(\theta_0, 0) + o_p(1).$$

Because $-\sqrt{n} S_{11}^{*-1/2} \hat{Q}_3(\theta_0, 0) \to^d N(0, I)$, and $-S_{11}^{*-1/2} S_{12}^* S_{22.1}^{*-1} S_{21}^* S_{11}^{*-1/2}$ is an idempotent matrix with trace p, we have $R_n(\theta_0) \to^d \chi_p^2$.

F: Proof of Theorem 3.6.1

Let $\eta = (\theta, \mu_2, \lambda, \nu)$ and define

$$U_1(\eta) = \frac{1}{N}\sum_{i \in A_2} \frac{\pi_{2i}^{-1} U_i(\theta)}{1 + \nu' G_i(\mu_2)}, \quad U_2(\eta) = \frac{1}{N}\sum_{i \in A_2} \frac{G_i(\mu_2)}{1 + \nu' G_i(\mu_2)},$$

$$U_3(\eta) = \frac{1}{N}\sum_{i \in A_1} \frac{\pi_{1i}^{-1}(h_{2i} - \mu_2)}{1 + \lambda' \pi_{1i}^{-1}(h_{1i} - \bar{h}_1)}, \quad U_4(\eta) = \frac{1}{N}\sum_{i \in A_1} \frac{\pi_{1i}^{-1}(h_{1i} - \bar{h}_1)}{1 + \lambda' \pi_{1i}^{-1}(h_{1i} - \bar{h}_1)}.$$

Hence, the proposed estimator $\hat{\theta}_{EL}$ can be obtained by solving

$$U_1(\eta) = U_2(\eta) = U_3(\eta) = U_4(\eta) = 0.$$

By using similar argument as the proof of Theorem 3.3.1 in the paper, it can be shown that

$$\hat{\eta} = (\hat{\theta}_{EL}, \hat{\mu}_2', \hat{\lambda}', \hat{\nu}')' \to^p (\theta_0, \mu_{2,0}', 0', 0')' = \eta_0 \tag{F.1}$$

and

$$\hat{\lambda} = O_p(n^{-1/2}), \quad \hat{\nu} = O_p(n^{-1/2}). \tag{F.2}$$

We have

$$U_1(\eta_0) = \frac{1}{N}\sum_{i \in A_2} \pi_{2i}^{-1} U_i(\theta_0), \quad U_2(\eta_0) = \frac{1}{N}\sum_{i \in A_2} G(\mu_{2,0}), \tag{F.3}$$

$$U_3(\eta_0) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}(h_{2i}-\bar{h}_2), \quad U_4(\eta_0) = \frac{1}{N}\sum_{i\in A_1}\pi_{1i}^{-1}(h_{1i}-\bar{h}_1), \tag{F.4}$$

$$\frac{\partial U_1(\eta_0)}{\partial\theta} \to^p \frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i(\theta_0)}{\partial\theta} \triangleq S_{11}, \quad \frac{\partial U_1(\eta_0)}{\partial\mu_2} = \frac{\partial U_1(\eta_0)}{\partial\lambda} = 0, \tag{F.5}$$

$$\frac{\partial U_1(\eta_0)}{\partial\nu} \to^p -\frac{1}{N}\left(\sum_{i=1}^{N}\frac{U_i(h_{1i}-\bar{h}_1)'}{\pi_{2i}}, \sum_{i=1}^{N}\frac{U_i(h_{2i}-\bar{h}_2)'}{\pi_{2i}}\right) \triangleq -S_{14}, \tag{F.6}$$

$$\frac{\partial U_2(\eta_0)}{\partial\theta} = \frac{\partial U_2}{\partial\lambda} = 0, \quad \frac{\partial U_2(\eta_0)}{\partial\mu_2} \to^p -\begin{pmatrix} 0_{s\times s} \\ I_{s\times s}\end{pmatrix} \triangleq S_{22}, \tag{F.7}$$

where we assume $\dim(h_1) = \dim(h_2) = s$,

$$\begin{aligned}\frac{\partial U_2(\eta_0)}{\partial\nu} &\to^p -\begin{bmatrix} N^{-1}\sum_{i=1}^{N}\pi_{2i}^{-1}(h_{1i}-\bar{h}_1)^{\otimes 2} & N^{-1}\sum_{i=1}^{N}\pi_{2i}^{-1}(h_{1i}-\bar{h}_1)(h_{2i}-\bar{h}_2)' \\ N^{-1}\sum_{i=1}^{N}\pi_{2i}^{-1}(h_{2i}-\bar{h}_2)(h_{1i}-\bar{h}_1)' & N^{-1}\sum_{i=1}^{N}\pi_{2i}^{-1}(h_{2i}-\bar{h}_2)^{\otimes 2}\end{bmatrix} \\ &\triangleq -S_{24},\end{aligned} \tag{F.8}$$

$$\frac{\partial U_3(\eta_0)}{\partial\theta} = \frac{U_3(\eta_0)}{\partial\nu} = 0, \quad \frac{\partial U_3(\eta_0)}{\partial\mu_2} \to^p -I_{s\times s} \triangleq S_{32}, \tag{F.9}$$

$$\frac{\partial U_4(\eta_0)}{\partial\theta} = \frac{\partial U_4(\eta_0)}{\partial\mu_2} = \frac{\partial U_4(\eta_0)}{\partial\nu} = 0, \quad \frac{\partial U_4(\eta_0)}{\partial\lambda} \to^p -\frac{1}{N}\sum_{i=1}^{N}\frac{(h_{1i}-\bar{h}_1)^{\otimes 2}}{\pi_{1i}^{-1}} \triangleq S_{43}. \tag{F.10}$$

According to (F.1)-(F.10), by using Taylor linearization, we have

$$0 = U_1(\hat{\eta}) = U_1(\eta_0) + S_{11}(\hat{\theta}-\theta_0) + S_{14}\hat{\nu} + o_p(n^{-1/2}), \tag{F.11}$$

$$0 = U_2(\hat{\eta}) = U_2(\eta_0) + S_{22}(\hat{\mu}_2-\mu_{2,0}) + S_{24}\hat{\nu} + o_p(n^{-1/2}), \tag{F.12}$$

$$0 = U_3(\hat{\eta}) = U_3(\eta_0) + S_{32}(\hat{\mu}_2-\mu_{2,0}) + S_{33}\hat{\lambda} + o_p(n^{-1/2}), \tag{F.13}$$

$$0 = U_4(\hat{\eta}) = U_4(\eta_0) + S_{43}\hat{\lambda} + o_p(n^{-1/2}). \tag{F.14}$$

By (F.11)-(F.14), after some algebra, we have

$$\hat{\theta}-\theta_0 = -S_{11}^{-1}\left\{U_1(\eta_0) - S_{14}S_{24}^{-1}U_2(\eta_0) - S_{14}S_{24}^{-1}S_{22}U_3(\eta_0) + S_{14}S_{24}^{-1}S_{22}S_{33}S_{43}^{-1}U_4(\eta_0)\right\} + o_p(n^{-1/2}),$$

which can be simplified to (3.28) in Theorem 3.6.1.

APPENDIX C. PROOFS FOR CHAPTER 4

A: Proof of Theorem 4.3.1

We first prove the consistency of $(\hat{\theta}, \hat{\lambda})$. Let

$$\hat{Q}_1(\theta, \lambda) = \frac{1}{n_B}\sum_{i=1}^{N}\frac{I_i\pi_i^{-1}f_N U_i(\theta)}{1+\lambda'\Psi_i}, \quad \hat{Q}_2(\theta, \lambda) = \frac{1}{n_B}\sum_{i=1}^{N}\frac{\Psi_i}{1+\lambda'\Psi_i}, \tag{A.1}$$

where $f_N = n_B/N$, $\Psi_i = f_N\big((I_i\pi_i^{-1}-1), I_i\pi_i^{-1}(h_i' - \bar{h}_N')\big)'$. Let $\hat{\lambda} = \rho\delta$, where $||\delta|| = 1$, so according to (A.1), we have

$$\begin{aligned}
0 &= ||n_B^{-1}\sum_{i=1}^{N}\frac{\Psi_i}{1+\hat{\lambda}'\Psi_i}|| \geq \big|n_B^{-1}\sum_{i=1}^{N}\frac{\delta'\Psi_i}{1+\rho\delta'\Psi_i}\big| \\
&= \big|n_B^{-1}\sum_{i=1}^{N}\frac{\delta'\Psi_i(1+\rho\delta'\Psi_i-\rho\delta'\Psi_i)}{1+\rho\delta'\Psi_i}\big| \\
&= \big|n_B^{-1}\sum_{i=1}^{N}\delta'\Psi_i - n_B^{-1}\sum_{i=1}^{N}\frac{\rho\delta'\Psi_i\Psi_i'\delta}{1+\rho\delta'\Psi_i}\big| \\
&\geq \big||n_B^{-1}\sum_{i=1}^{N}\delta'\Psi_i| - |n_B^{-1}\sum_{i=1}^{N}\frac{\rho\delta'\Psi_i\Psi_i^T\delta}{1+\rho\delta'\Psi_i}|\big|.
\end{aligned}$$

Hence,

$$\begin{aligned}
|\frac{1}{n_B}\sum_{i=1}^{N}\delta'\Psi_i| &= |\frac{1}{n_B}\sum_{i=1}^{N}\frac{\rho\delta'\Psi_i\Psi_i'\delta}{1+\rho\delta'\Psi_i}| \\
&\geq \big|\delta'\frac{1}{n_B}\sum_{i=1}^{N}\Psi_i\Psi_i'\delta\big|\frac{|\rho|}{1+|\rho|u^*},
\end{aligned} \tag{A.2}$$

where $u^* = \max_{i\in A}||\Psi_i||$.

Under assumption (C3), we have $n_B^{-1}\sum_{i=1}^{N}\Psi_i\Psi_i' = \Sigma_\Psi + o_p(1)$, and Σ_Ψ is a positive definite matrix. Let λ_p be the smallest eigenvalue of Σ_Ψ, then $\lambda_p > 0$. So, the following holds

$$\big|\delta' n_B^{-1}\sum_{i=1}^{N}\Psi_i\Psi_i'\delta\big| \geq \lambda_p + o_p(1). \tag{A.3}$$

In addition, according to Assumption (C3),

$$\frac{1}{n_B}\sum_{i=1}^{N}\delta'\Psi_i = O_p(n_B^{-1/2}). \tag{A.4}$$

By (A.2), (A.3), (A.4) and assumptions (C5), (C6),

$$\lambda_p|\rho| = O_p(n_B^{-1/2}) + o_p(|\rho|).$$

Thus, we have $|\rho| = O_p(n_B^{-1/2})$, which means $||\hat{\lambda}|| = O_p(n_B^{-1/2})$.

Because $\max_{i\in A}|\hat{\lambda}'\Psi_i| = O_p(n_B^{-1/2})o_p(n_B) = o_p(1)$ and assumption (C4), we can apply Taylor expansion and get

$$\begin{aligned} 0 &= \frac{1}{n_B}\sum_{i=1}^{N}\frac{f_N I_i \pi_i^{-1} U_i(\hat{\theta})}{1+\hat{\lambda}'\Psi_i} \\ &= \frac{1}{n_B}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\hat{\theta}) - \left\{\frac{1}{n_B}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\hat{\theta})\Psi_i'\right\}\hat{\lambda} \\ &+ O_p(n_B^{-1}). \end{aligned} \tag{A.5}$$

By assumption (C4), it can be shown that

$$n_B^{-1}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\hat{\theta})\Psi_i' = O_p(1), \tag{A.6}$$

and

$$\sup_{\theta\in\Theta}||\frac{1}{n_B}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\theta) - \frac{1}{N}\sum_{i=1}^{N} U_i(\theta)|| \to^p 0, \tag{A.7}$$

so according to (A.5), (A.6) and (A.7),

$$\begin{aligned} 0 &= p\lim\left|\frac{1}{n_B}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\hat{\theta})\right| \\ &= p\lim\left|\frac{1}{n_B}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\hat{\theta}) - \frac{1}{N}\sum_{i=1}^{N} U_i(\hat{\theta}) + \frac{1}{N}\sum_{i=1}^{N} U_i(\hat{\theta})\right| \\ &\geq p\lim\left||\frac{1}{n_B}\sum_{i=1}^{N} f_N I_i \pi_i^{-1} U_i(\hat{\theta}) - \frac{1}{N}\sum_{i=1}^{N} U_i(\hat{\theta})| - |\frac{1}{N}\sum_{i=1}^{N} U_i(\hat{\theta})|\right| \\ &= \left|\frac{1}{N}\sum_{i=1}^{N} U_i(\hat{\theta})\right|. \end{aligned} \tag{A.8}$$

By (A.8), assumptions (C1) and (C2), we have $\hat{\theta} \to^p \theta_0$. Hence,

$$(\hat{\theta}_{POEL}, \hat{\lambda}) \to^p (\theta_0, 0). \tag{A.9}$$

According to (A.9), assumptions (C2) and (C4), we can apply the standard arguments using Taylor expansion to get

$$0 = \hat{Q}_1(\hat{\theta}, \hat{\lambda}) = \hat{Q}_1(\theta_0, 0) + \frac{\partial \hat{Q}_1(\theta_0, 0)}{\partial \theta'}(\hat{\theta} - \theta_0) + \frac{\partial \hat{Q}_1(\theta_0, 0)}{\partial \lambda'}(\hat{\lambda} - 0) + o_p(\delta_n),$$

and

$$0 = \hat{Q}_2(\hat{\theta}, \hat{\lambda}) = \hat{Q}_2(\theta_0, 0) + \frac{\partial \hat{Q}_2(\theta_0, 0)}{\partial \theta'}(\hat{\theta} - \theta_0) + \frac{\partial \hat{Q}_2(\theta_0, 0)}{\partial \lambda'}(\hat{\lambda} - 0) + o_p(\delta_n),$$

with $\delta_n = ||\hat{\theta} - \theta_0|| + ||\hat{\lambda}||$. Let

$$S_n = \begin{pmatrix} \partial \hat{Q}_1(\theta_0, 0)/\partial \lambda & \partial \hat{Q}_1(\theta_0, 0)/\partial \theta \\ \partial \hat{Q}_2(\theta_0, 0)/\partial \lambda & \partial \hat{Q}_2(\theta_0, 0)/\partial \theta \end{pmatrix}.$$

Under the existence of moments, we can obtain

$$S_n \rightarrow^p \begin{pmatrix} S_{11}^* & S_{12}^* \\ S_{21}^* & 0 \end{pmatrix},$$

and

$$||S_{11}^* - S_{11}|| = o_p(1), \quad ||S_{12}^* - S_{12}|| = o_p(1), \quad ||S_{21}^* - S_{21}|| = o_p(1), \tag{A.10}$$

where

$$S_{11} = -\Big(\frac{1}{N}\sum_{i=1}^N f_N(\frac{1}{\pi_i} - 1)U_i, \frac{1}{N}\sum_{i=1}^N f_N \frac{1}{\pi_i} U_i (h_i - \bar{h}_N)'\Big), \quad S_{12} = N^{-1}\sum_{i=1}^N \frac{\partial U_i(\theta_0)}{\partial \theta}, \tag{A.11}$$

and

$$S_{21} = -\begin{pmatrix} N^{-1}\sum_{i=1}^N f_N(\pi_i^{-1} - 1) & N^{-1}\sum_{i=1}^N f_N(\pi_i^{-1} - 1)(h_i - \bar{h}_N)' \\ N^{-1}\sum_{i=1}^N f_N(\pi_i^{-1} - 1)(h_i - \bar{h}_N) & N^{-1}\sum_{i=1}^N f_N \pi_i^{-1}(h_i - \bar{h}_N)^{\otimes 2} \end{pmatrix}. \tag{A.12}$$

According to assumption (C3),

$$\hat{Q}_1(\theta_0, 0) = N^{-1}\sum_{i=1}^N I_i \pi_i^{-1} U_i(\theta_0) = O_p(n_B^{-1/2}), \quad \hat{Q}_2(\theta_0, 0) = N^{-1}\sum_{i=1}^N \Psi_i(\theta_0, 0) = O_p(n_B^{-1/2}),$$

so we have $\delta_n = O_p(n_B^{-1/2})$. Also, according to (A.10), (A.11) and (A.12), after some algebra,

$$\hat{\lambda} = -S_{21}^{-1}\hat{Q}_2(\theta_0, 0) + o_p(n_B^{-1/2}), \tag{A.13}$$

and

$$\begin{aligned}\hat{\theta}-\theta_0 &= -S_{12}^{-1}\left\{\hat{Q}_1(\theta_0,0)-S_{11}S_{21}^{-1}\hat{Q}_2(\theta_0,0)\right\}+o_p(n_B^{-1/2})\\ &= -\tau\left\{\hat{Q}_1(\theta_0,0)-B^*\hat{Q}_2(\theta_0,0)\right\}+o_p(n_B^{-1/2}),\end{aligned}\tag{A.14}$$

where $\tau = S_{12}$, $B^* = \Omega_1\Omega_2^{-1}$, $\Omega_1 = -(N\alpha_N)^{-1}S_{11}$ and $\Omega_2 = -(N\alpha_N)^{-1}S_{21}$. Hence,(4.16) in Theorem 4.3.1 is proved. Result (4.19) can be obtained by (A.14) and assumptions (C3), (C4).

B: Proof of Theorem 4.3.2

Maximizing (4.10) subject to

$$\sum_{i=1}^N \omega_i = 1,\quad \sum_{i=1}^N \omega_i f_N(\frac{I_i}{\pi_i}-1)=0,\quad \sum_{i=1}^N \omega_i\frac{I_i}{\pi_i}f_N(h_i-\bar{h}_N)=0,$$

leads, after some algebra, to

$$l(\hat{\theta})=\sum_{i=1}^N \log(\omega_i(\hat{\theta})) = -N\log(N)-\sum_{i=1}^N\log(1+\hat{\lambda}'\Psi_{i1}),$$

where $\Psi_{i1} = (f_N(I_i\pi_i^{-1}-1), I_i\pi_i^{-1}f_N(h_i-\bar{h}_N)')'$. Similarly, consider maximizing (3.1) subject to

$$\sum_{i=1}^N \omega_i = 1,\quad \sum_{i=1}^N \omega_i f_N(\frac{I_i}{\pi_i}-1)=0,\quad \sum_{i=1}^N \omega_i\frac{I_i}{\pi_i}f_N(h_i-\bar{h}_N)=0,$$

and

$$\sum_{i=1}^N \omega_i f_N r_i = 0,$$

with $r_i = I_i\pi_i^{-1}U_i(\theta_0) - B_1^*(I_i\pi_i^{-1}-1) - B_2^* I_i\pi_i^{-1}(h_i-\bar{h}_N)$ and $B^* = (B_1^*, B_2^*) = S_{11}S_{21}^{-1}$, where S_{11}, S_{21} are defined in (A.11) and (A.12) of the proof of Theorem 4.3.1. After some algebra,

$$l(\theta_0)=\sum_{i=1}^N \log(\omega_i(\theta_0)) = -N\log(N)-\sum_{i=1}^N\log(1+\lambda_0'\Psi_{i2}),$$

where $\Psi_{i2} = (f_N(I_i\pi_i^{-1}-1), I_i\pi_i^{-1}f_N(h_i-\bar{h}_N)', f_N r_i')'$. So, we can write

$$R_n(\theta_0) = 2\big\{\sum_{i=1}^N \log(1+\lambda_0'\Psi_{i2}) - \sum_{i=1}^N\log(1+\hat{\lambda}'\Psi_{i1})\big\},\tag{B.1}$$

and λ_0 is the solution of $\hat{Q}_3(\theta_0,\lambda)=0$ with

$$\hat{Q}_3(\theta_0,\lambda) = \frac{1}{n_B}\sum_{i=1}^N\frac{\Psi_{i2}(\theta_0)}{1+\lambda'\Psi_{i2}(\theta_0)}.$$

By the same argument for (A.9), we have $\lambda_0 \to^p 0$. We can apply a Taylor expansion to get

$$0 = \hat{Q}_3(\theta_0, \lambda_0) = \hat{Q}_3(\theta_0, 0) + \frac{\partial \hat{Q}_3(\theta_0, 0)}{\partial \lambda}\lambda_0 + o_p(||\lambda_0||).$$

According to assumption (C3), $\hat{Q}_3(\theta_0, 0) = n_B^{-1} \sum_{i=1}^N \Psi_{i2}(\theta_0, 0) = O_p(n_B^{-1/2})$, hence $||\lambda_0|| = O_p(n_B^{-1/2})$, so

$$\lambda_0 = -S^{-1}\hat{Q}_3(\theta_0, 0) + o_p(n_B^{-1/2}), \tag{B.2}$$

with

$$S = \begin{pmatrix} S_{21} & 0 \\ 0 & S_r \end{pmatrix}, \tag{B.3}$$

where

$$\begin{aligned} S_r &= f_N\{-NV_{poi}(\bar{r}_N) - N^{-1}\sum_{i=1}^N U_i^{\otimes 2} - B_2^* N^{-1}\sum_{i=1}^N (h_i - \bar{h}_N)^{\otimes 2} B_2^{*'} \\ &+ N^{-1}\sum_{i=1}^N U_i (h_i - \bar{h}_N)' B_2^{*'} + B_2^* N^{-1}\sum_{i=1}^N (h_i - \bar{h}_N) U_i'\}, \end{aligned} \tag{B.4}$$

and $\bar{r}_N = N^{-1}\sum_{i=1}^N r_i$, S_{21} is defined in (A.12) in the proof of Theorem 4.3.1 and V_{poi} is the variance under Poisson sampling.

According to assumption (C6), $n_B/N = o(1)$ and (B.4), it can be shown that

$$||S_r + f_N N V_{poi}(\bar{r}_N)|| = o(1). \tag{B.5}$$

Similarly, by a Taylor expansion with respect to $\lambda_0 = 0$,

$$2\sum_{i=1}^N \log(1 + \lambda_0' \Psi_{i2}) = 2\sum_{i=1}^N \lambda_0' \Psi_{i2} - \sum_{i=1}^N \lambda_0' \Psi_{i2} \Psi_{i2}' \lambda_0 + o_p(1). \tag{B.6}$$

According to (B.2), we have

$$\partial \hat{Q}_3(\theta_0, 0)/\partial \lambda = -n_B^{-1}\sum_{i=1}^N \Psi_{i2}(\theta_0)\Psi_{i2}'(\theta_0) \to^p S. \tag{B.7}$$

By plugging (B.2) into (B.6) and according to (B.7), we have

$$2\sum_{i=1}^N \log(1 + \lambda_0' \Psi_{i2}(\theta_0)) = -n_B \hat{Q}_3'(\theta_0, 0) S^{-1} \hat{Q}_3(\theta_0, 0) + o_p(1). \tag{B.8}$$

Similarly, according to (A.13) and by using a Taylor expansion around $\hat{\lambda} = 0$,

$$2\sum_{i=1}^N \log(1 + \hat{\lambda}' \Psi_{i1}(\hat{\theta})) = -n_B \hat{Q}_2'(\theta_0, 0) S_{21}^{-1} \hat{Q}_2(\theta_0, 0) + o_p(1). \tag{B.9}$$

By assumption (C3) and (C4), we can apply the central limit theorem to get

$$V_{poi}^{-1/2}(\bar{r}_N)\bar{r}_N \to^d N(0, I). \tag{B.10}$$

Therefore, plugging (B.8) and (B.9) into (B.1) and by (B.3), we have

$$\begin{aligned} R_n(\theta_0) &= -n_B\hat{Q}_3'(\theta_0, 0)S^{-1}\hat{Q}_3(\theta_0, 0) + n_B\hat{Q}_2'(\theta_0, 0)S_{21}^{-1}\hat{Q}_2(\theta_0, 0) + o_p(1) \\ &= \bar{r}_N(-n_B^{-1}S_r)^{-1}(\bar{r}_N)' + o_p(1). \end{aligned} \tag{B.11}$$

According to (B.5), (B.11) and (B.10),

$$R_n(\theta_0) = \bar{r}_N \left\{V_{poi}(\bar{r}_N)\right\}^{-1} (\bar{r}_N)' + o_p(1) \to^d \chi_p^2,$$

where $\bar{r}_N = N^{-1}\sum_{i=1}^N r_i$, and p is the dimension of θ_0.

C: Proof of Theorem 4.4.1

Similar as the proof of Theorem 4.3.1, $\hat{\theta}$ can be obtained by solving $\hat{Q}_1(\theta, \lambda) = 0$ and $\hat{Q}_2(\theta, \lambda) = 0$, where $\hat{Q}_1(\theta, \lambda)$ and $\hat{Q}_2(\theta, \lambda)$ are defined in (A.1) with π_i, Ψ_i replaced by p_i, Ψ_i^*, and $\Psi_i^* = \left(f_N(I_ip_i^{-1} - 1)z_i^*, f_NI_ip_i^{-1}(h_i' - \bar{h}_N')\right)'$, with $z_i^* = (1, z_i')'$. Without loss of generality, we assume $\bar{z}_N = 0$ and $n_BN^{-2}\sum_{i=1}^N(1 - p_i)p_i^{-1}z_i^2 = 1$. Hence, according to assumption (C9) and (C10) in Section 4,

$$\begin{aligned} \pi_i &= Pr(i \in s|\hat{Q}_{p,n} \le \gamma^2) = \frac{Pr(\hat{Q}_{p,n} \le \gamma^2|i \in s)Pr(i \in s)}{Pr(\hat{Q}_{p,n} \le \gamma^2)} \\ &= p_i\left\{1 + C_\gamma\eta_i + o_p(n_B^{-1})\right\}, \end{aligned} \tag{C.1}$$

and

$$\begin{aligned} \pi_{ij} &= Pr(i, j \in s|\hat{Q}_{p,n} \le \gamma^2) = \frac{Pr(\hat{Q}_{p,n} \le \gamma^2|i, j \in s)Pr(i, j \in s)}{Pr(\hat{Q}_{p,n} \le \gamma^2)} \\ &= p_{ij}\left\{1 + C_\gamma(\eta_i + \eta_j) + o_p(n_B^{-1})\right\}, \end{aligned} \tag{C.2}$$

with $C_\gamma = g_{1N}(\gamma^2)G_N^{-1}(\gamma^2)$.

According to (C.1), (C.2) and by using a similar argument as the proof of Theorem 4.3.1, it can be shown that $(\hat{\theta}, \hat{\lambda}) \to^p (\theta_0, 0)$. After some algebra,

$$\hat{\theta} - \theta_0 = -S_{12}^{-1}\left\{\hat{Q}_1(\theta_0, 0) - S_{11}S_{21}^{-1}\hat{Q}_2(\theta_0, 0)\right\} + o_p(n_B^{-1/2}), \tag{C.3}$$

where

$$S_{11} = -(N^{-1}\sum_{i=1}^{N} f_N(\pi_i p_i^{-2} - \pi_i p_i^{-1})U_i z_i^{*\prime}, N^{-1}\sum_{i=1}^{N} f_N \pi_i p_i^{-2} U_i(h_i - \bar{h}_N)'),$$

$$S_{12} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi_i}{p_i}\frac{\partial U_i(\theta_0)}{\partial \theta}$$

and

$$S_{21} = -N^{-1}\begin{pmatrix} \sum_{i=1}^{N} f_N(\pi_i p_i^{-2} - 2\pi_i p_i^{-1} + 1) z_i^* z_i^{*\prime} & \sum_{i=1}^{N} f_N(\pi_i p_i^{2} - \pi_i p_i^{-1}) z_i^*(h_i - \bar{h}_N)' \\ \sum_{i=1}^{N} f_N(\pi_i p_i^{2} - \pi_i p_i^{-1})(h_i - \bar{h}_N) z_i^{*\prime} & \sum_{i=1}^{N} f_N \pi_i p_i^{-2}(h_i - \bar{h}_N)^{\otimes 2} \end{pmatrix}.$$

By (C.1) and (C.2),

$$||\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{p_i}U_i(\theta_0) - \frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}U_i(\theta_0)|| = o_p(n_B^{-1/2}).$$

Hence,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{p_i}U_i(\theta_0) = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}U_i(\theta_0) + o_p(n_B^{-1/2}). \tag{C.4}$$

Similarly, it can be shown that

$$\frac{1}{N}\sum_{i=1}^{N}(\frac{I_i}{p_i} - 1)z_i^* = \frac{1}{N}\sum_{i=1}^{N}(\frac{I_i}{\pi_i} - 1)z_i^* + o_p(n_B^{-1/2}), \tag{C.5}$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{p_i}(h_i - \bar{h}_N) = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}(h_i - \bar{h}_N) + o_p(n_B^{-1/2}), \tag{C.6}$$

$$||S_{12} - S_{12}^*|| = o_p(1), \quad ||S_{11} - S_{11}^*|| = o_p(1), \quad ||S_{21} - S_{21}^*|| = o_p(1),$$

where

$$S_{11}^* = -(N^{-1}\sum_{i=1}^{N} f_N(\pi_i^{-1} - 1)U_i z_i^{*\prime}, N^{-1}\sum_{i=1}^{N} f_N \pi_i^{-1} U_i(h_i - \bar{h}_N)'),$$

$$S_{12}^* = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i(\theta_0)}{\partial \theta}$$

and

$$S_{21}^* = -\begin{pmatrix} N^{-1}\sum_{i=1}^{N} f_N(\pi_i^{-1} - 1) z_i^* z_i^{*\prime} & N^{-1}\sum_{i=1}^{N} f_N(\pi_i^{-1} - 1) z_i^*(h_i - \bar{h}_N)' \\ N^{-1}\sum_{i=1}^{N} f_N(\pi_i^{-1} - 1)(h_i - \bar{h}_N) z_i^{*\prime} & N^{-1}\sum_{i=1}^{N} f_N \pi_i^{-1}(h_i - \bar{h}_N)^{\otimes 2} \end{pmatrix}.$$

Hence according to previous derivations and (C.3),

$$\hat{\theta} - \theta_0 = -\tau \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} U_i(\theta_0) - B(\frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} \eta_i - \bar{\eta}_N) \right\} + o_p(n_B^{-1/2}) \tag{C.7}$$

with $\tau = S_{12}^*$, $\eta = (z_i^*, (h - \bar{h}_N)')'$, $B = \Omega_1 \Omega_2^{-1}$, $\Omega_1 = -(Nf_N)^{-1} S_{11}^*$, $\Omega_2 = -(Nf_N)^{-1} S_{21}^*$. Thus, (4.32) in Theorem 4.4.1 is proved.

Let $e_i = U_i - B\eta_i$ and $\hat{\bar{e}}_p = N^{-1} \sum_{i=1}^{N} I_i p_i^{-1} e_i$. Next we want to prove

$$||V_{rej}(\hat{\bar{e}}_p) - V_{poi}(\hat{\bar{e}}_p)|| = o_p(n_B^{-1}), \tag{C.8}$$

where V_{rej} and V_{poi} denote the variances under rejective Poisson sampling and Poisson sampling, respectively. According to (C.1) and (C.2),

$$\begin{aligned}
V_{rej}(\hat{\bar{e}}_p) &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{p_i p_j} e_i e_j' \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \frac{\pi_i - \pi_i^2}{p_i^2} e_i^{\otimes 2} + \frac{1}{N^2} \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{p_i p_j} e_i e_j' \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - p_i}{p_i} e_i^{\otimes 2} + \frac{1}{N^2} \sum_{i=1}^{N} (1 - p_i) p_i^{-2} n_B N^{-2} z_i^2 e_i^{\otimes 2} \\
&+ \frac{1}{N^2} \sum_{i \neq j} p_i^{-1} p_j^{-1} o_p(\frac{n_B}{N^2}) e_i e_j' + o_p(n_B^{-1}) \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \frac{1 - p_i}{p_i} e_i^{\otimes 2} + o_p(n_B^{-1}) = V_{poi}(\hat{\bar{e}}_p) + o_p(n_B^{-1}).
\end{aligned}$$

So, (C.8) is proved. Together with (C.4), (C.5) and (C.6),

$$||V_{rej}(\hat{\bar{e}}_{HT}) - V_{poi}(\hat{\bar{e}}_p)|| = o_p(n_B^{-1}), \tag{C.9}$$

where $\hat{\bar{e}}_{HT} = N^{-1} \sum_{i=1}^{N} I_i \pi_i^{-1} e_i$. Hence, result (4.35) in Theorem 4.4.1 can be obtained by (C.7), (C.9) and assumptions (C3), (C4).

D: Proof of Theorem 4.4.2

By using the argument similar to the proof of Theorem 4.3.2, it can be shown that

$$R_n(\theta_0) = \bar{r}_N \left\{ V_{poi}(\bar{r}_N) \right\}^{-1} (\bar{r}_N)' + o_p(1), \tag{D.1}$$

where $\bar{r}_N = \hat{Q}_1(\theta_0, 0) - S_{11}S_{21}^{-1}\hat{Q}_2(\theta_0, 0)$, and $\hat{Q}_1(\theta_0, 0), \hat{Q}_2(\theta_0, 0), S_{11}, S_{21}$ are defined in (C.3) of the proof for Theorem 4.4.1. $\bar{r}_N = N^{-1}\sum_{i=1}^{N} r_i$, and p is the dimension of θ_0. According to (C.8) in the proof of Theorem 4.4.1 and (D.1), we have $R_n(\theta_0) \to^d \chi_p^2$.

APPENDIX D. PROOFS FOR CHAPTER 5

A. Use of population auxiliary information

In this section, we assume population information $\bar{\mathbf{X}}_N$ is available. If we want to incorporate both population and sample level information and obtain the optimal estimators, similar to Section 3, the regression estimator $\hat{\theta}_{1,Reg}$ of θ_1 can be written as

$$\begin{aligned}\hat{\theta}_{1,Reg} &= \hat{\theta}_1 - \hat{B}_{Reg}(\hat{\theta}_{x,1} - \bar{\mathbf{X}}_N) \\ &\cong \hat{\theta}_1 - B_{Reg}(\hat{\theta}_{x,1} - \bar{\mathbf{X}}_N),\end{aligned} \tag{A.1}$$

where $\hat{\theta}_1$ is defined in (5.4), $\hat{\theta}_{x,1} = N^{-1}\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}\mathbf{x}_i$ and $\hat{B}_{Reg} = \sum_{i\in A_1}\hat{\pi}_{1i}^{-1}y_{1i}\mathbf{x}_i'(\sum_{i\in A_1}\hat{\pi}_{1i}^{-1}\mathbf{x}_i\mathbf{x}_i')^{-1}$ and $B_{Reg} = p\lim \hat{B}_{Reg}$. After ignoring the higher order terms, it can be shown that

$$\hat{\theta}_{1,Reg} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\delta_{1i}}{\pi_{1i}}(y_{1i} - B_{Reg}\mathbf{x}_i) - B_{1,Reg}\delta_{1i}(\frac{\delta_{2i}}{\pi_{2i}} - 1)\mathbf{h}_{1i} - B_{2,Reg}(\frac{\delta_{1i}}{\pi_{1i}} - 1) + B_{Reg}\bar{\mathbf{X}}_N\right\},$$

where $B_{1,Reg} = B_{1,y} - B_{Reg}B_{1,x},\ B_{2,Reg} = B_{2,y} - B_{Reg}B_{2,x}$,

$$(B_{1,x}, B_{2,x}) = \sum_{i=1}^{N}(1-\pi_{1i})\mathbf{x}_i(0, \mathbf{x}_i', y_{1i}, 1)\begin{pmatrix}\sum_{i=1}^{N}\pi_{1i}(1-\pi_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i=1}^{N}(1-\pi_{1i})(0, \mathbf{x}_i', y_{1i}, 1)\end{pmatrix}^{-1}.$$

The asymptotic variance can be estimated as follows

$$\hat{V}_{1,Reg} = \frac{1}{N^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}(y_{1i} - \hat{B}_{Reg}\mathbf{x}_i - \hat{B}_{2,Reg})^2 + \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}(\hat{B}_{1,Reg}\mathbf{h}_{1i})^2,$$

where $\hat{B}_{1,Reg} = \hat{B}_{1,y} - \hat{B}_{Reg}\hat{B}_{1,x},\ \hat{B}_{2,Reg} = \hat{B}_{2,y} - \hat{B}_{Reg}\hat{B}_{2,x}$,

$$(\hat{B}_{1,x}, \hat{B}_{2,x}) = \sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})\mathbf{x}_i(0, \mathbf{x}_i', y_{1i}, 1)\begin{pmatrix}\sum_{i\in A_2}\hat{\pi}_{2i}^{-1}(1-\hat{\pi}_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1} \\ \sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1)\end{pmatrix}^{-1}.$$

Specifically, the optimal estimator of θ_1 can be written as

$$\hat{\theta}_{1,opt} = \hat{\theta}_1 - \hat{B}_{opt}(\hat{\theta}_{1,x} - \bar{\mathbf{X}}_N), \tag{A.2}$$

where $\hat{B}_{opt} = \hat{Cov}(\hat{\theta}_1, \hat{\theta}_{x,1})\hat{V}^{-1}(\hat{\theta}_{x,1})$, with

$$\hat{Cov}(\hat{\theta}_1, \hat{\theta}_{x,1}) = \frac{1}{N^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}\hat{\boldsymbol{\eta}}_i^* + \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}\hat{B}_{1,y}\mathbf{h}_{1i}(\hat{B}_{1,x}\mathbf{h}_{1i})',$$

$$\hat{V}(\hat{\theta}_{x,1}) = \frac{1}{N^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}(\mathbf{x}_i - \hat{B}_{2,x})^{\otimes 2} + \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}(\hat{B}_{1,x}\mathbf{h}_{1i})^{\otimes 2},$$

and $\hat{\boldsymbol{\eta}}_i^* = \mathbf{x}_i' y_{1i} - y_{1i}\hat{B}_{2,x}' - \hat{B}_{2,y}\mathbf{x}_i' + \hat{B}_{2,y}\hat{B}_{2,x}'$. $B_{1,y}$, $B_{2,y}$, $\hat{B}_{1,y}$ and $\hat{B}_{2,y}$ are defined in Section 3. Similarly, the regression estimator of θ_2 can be written as

$$\begin{aligned}\hat{\theta}_{2,Reg} &= \hat{\theta}_2 - \hat{B}_{1,Reg}^*(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}) - \hat{B}_{2,Reg}^*(\hat{\theta}_{x,2} - \bar{\mathbf{X}}_N)\\ &= \hat{\theta}_2 - B_{1,Reg}^*(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}) - B_{2,Reg}^*(\hat{\theta}_{x,2} - \bar{\mathbf{X}}_N),\end{aligned}$$

where $\hat{\theta}_{x,2} = N^{-1}\sum_{i\in A_2}\hat{\pi}_{1i}^{-1}\hat{\pi}_{2i}^{-1}\mathbf{x}_i$. $(\hat{B}_{1,Reg}^*, \hat{B}_{2,Reg}^*)$ is the regression coefficient. It can be shown that

$$\hat{\theta}_{2,Reg} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\delta_{1i}\delta_{2i}}{\pi_{1i}\pi_{2i}}(y_{2i} - B_{2,Reg}^*\mathbf{x}_i) - D_{1i,Reg}^*\delta_{1i}(\frac{\delta_{2i}}{\pi_{2i}} - 1)\mathbf{h}_{1i} - D_{2,Reg}^*(\frac{\delta_{1i}}{\pi_{1i}} - 1) + B_{2,Reg}^*\bar{\mathbf{X}}_N\right\},$$

where

$$D_{1i,Reg}^* = \left\{D_1 + B_{1,Reg}^*(\pi_{1i}^{-1} - D_{1,h_1} + B_{1,h_1}) - B_{2,Reg}^* D_{1,x}\right\},$$

$$D_{2,Reg}^* = \left\{D_2 - B_{1,Reg}^*(D_{2,h_1} - B_{2,h_1}) - B_{2,Reg}^* D_{2,x}\right\},$$

and

$$\begin{aligned}(D_{1,x}, D_{2,x}) &= \sum_{i\in A_2}\frac{\mathbf{x}_i}{\hat{\pi}_{1i}\hat{\pi}_{2i}}\left\{(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1) + (1-\hat{\pi}_{2i})(1, \mathbf{x}_i', y_{2i}, 0)\right\}\\ &\times \begin{pmatrix}\sum_{i\in A_2}\hat{\pi}_{2i}^{-1}(1-\hat{\pi}_{2i})\mathbf{h}_{1i}\mathbf{h}_{2i}', \mathbf{0}_{r\times 1}\\ \sum_{i\in A_1}\hat{\pi}_{1i}^{-1}(1-\hat{\pi}_{1i})(0, \mathbf{x}_i', y_{1i}, 1)\end{pmatrix}^{-1}.\end{aligned}$$

The asymptotic variance can be estimated by

$$\hat{V}_{2,Reg} = \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2\hat{\pi}_{2i}}(y_{2i} - \hat{B}_{2,Reg}^*\mathbf{x}_i - \hat{D}_{2,Reg}^*)^2 + \frac{1}{N^2}\sum_{i\in A_2}\frac{1-\hat{\pi}_{2i}}{\hat{\pi}_{1i}^2\hat{\pi}_{2i}^2}(y_{2i} - \hat{D}_{1i,Reg}^*\hat{\pi}_{1i}\mathbf{h}_{1i})^2,$$

where

$$\hat{D}_{1i,Reg}^* = \left\{\hat{D}_1 + \hat{B}_{1,Reg}^*(\hat{\pi}_{1i}^{-1} - \hat{D}_{1,h_1} + \hat{B}_{1,h_1}) - \hat{B}_{2,Reg}^*\hat{D}_{1,x}\right\}$$

and

$$\hat{D}_{2,Reg}^* = \left\{\hat{D}_2 - \hat{B}_{1,Reg}^*(\hat{D}_{2,h_1} - \hat{B}_{2,h_1}) - \hat{B}_{2,Reg}^*\hat{D}_{2,x}\right\},$$

with other estimators defined before. Specifically, the optimal estimator of θ_2 can be written as

$$\hat{\theta}_{2,opt} = \hat{\theta}_2 - \hat{B}^*_{1,opt}(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}) - \hat{B}^*_{2,opt}(\hat{\theta}_{x,2} - \bar{\mathbf{X}}_N), \tag{A.3}$$

where

$$(\hat{B}^*_{1,opt}, \hat{B}^*_{2,opt}) = \hat{C}\left\{\hat{\theta}_2, (\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}, \hat{\theta}_{x,2})\right\}\left[\hat{V}\left\{(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}, \hat{\theta}_{x,2})\right\}\right]^{-1}.$$

$\hat{C}(\hat{\theta}_2, \hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1})$, $\hat{V}(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1})$ are defined in (5.15) and (5.16),

$$\hat{C}(\hat{\mathbf{h}}_{2,1} - \hat{\mathbf{h}}_{1,1}, \hat{\theta}_{x,2}) = \frac{1}{N^2}\sum_{i \in A_2} \hat{\boldsymbol{\eta}}_i' \left\{\mathbf{x}_i' \hat{\pi}_{1i}^{-1} - (\hat{D}_{1,x}\mathbf{h}_{1i})'\right\} \frac{1 - \hat{\pi}_{2i}}{\hat{\pi}_{2i}^2}.$$

$$\hat{C}(\hat{\theta}_2, \hat{\theta}_{x,2}) = \frac{1}{N^2}\sum_{i \in A_2} \frac{1 - \hat{\pi}_{3i}}{\hat{\pi}_{3i}^2}\mathbf{x}_i' y_{2i} - \frac{1}{N^2}\sum_{i \in A_2} \hat{\boldsymbol{\zeta}}_i' \frac{1 - \hat{\pi}_{2i}}{\hat{\pi}_{2i}^2} - \frac{1}{N^2}\sum_{i \in A_2} \hat{\boldsymbol{\xi}}_i' \frac{1 - \hat{\pi}_{1i}}{\hat{\pi}_{1i}\hat{\pi}_{3i}},$$

$$\hat{V}(\hat{\theta}_{x,2}) = \frac{1}{N^2}\sum_{i \in A_2} \frac{1 - \hat{\pi}_{1i}}{\hat{\pi}_{1i}^2 \hat{\pi}_{2i}}(\mathbf{x}_i - \hat{D}_{2,x})^{\otimes 2} + \frac{1}{N^2}\sum_{i \in A_2} \frac{1 - \hat{\pi}_{2i}}{\hat{\pi}_{1i}^2 \hat{\pi}_{2i}^2}(\mathbf{x}_i - \hat{D}_{1,x}\hat{\pi}_{1i}\mathbf{h}_{1i})^{\otimes 2},$$

where $\hat{\boldsymbol{\eta}}_i$ is defined in (5.16), $\hat{\boldsymbol{\zeta}}_i = y_{2i}(\hat{D}_{1,x}\mathbf{h}_{1i})\hat{\pi}_{1i}^{-1} + (\hat{D}_1\mathbf{h}_{1i})\mathbf{x}_i\hat{\pi}_{1i}^{-1} - (\hat{D}_{1,y}\mathbf{h}_{1i})(\hat{D}_{1,x}\mathbf{h}_{1i})'$ and $\hat{\boldsymbol{\xi}}_i = y_{2i}\hat{D}_{2,x} + \hat{D}_{2,y}\mathbf{x}_i - \hat{D}_{2,y}\hat{D}_{2,x}$.

B. Use of sample and population auxiliary information for Section 4

To discuss asymptotic properties of the two-phase regression estimator in (5.20), by Taylor linearization, we have

$$\hat{\mathbf{h}}_1^* \cong \frac{1}{N}\sum_{i=1}^N \left\{\frac{\delta_{1i}}{\pi_{1i}}(\mathbf{h}_i^* - \theta_{h^*}) - B_{1,h^*}(\delta_{1i} - \frac{\delta_i^*}{p_i}\pi_{1i})\mathbf{h}_i - B_{2,h^*}(\delta_{2i} - \frac{\delta_i^*}{p_i}\pi_{2i})\mathbf{h}_i\right\} \tag{B.1}$$

and

$$\hat{\mathbf{h}}_2^* \cong \frac{1}{N}\sum_{i=1}^N \left\{\frac{\delta_{2i}}{\pi_{2i}}(\mathbf{h}_i^* - \theta_{h^*}) - D_{1,h^*}(\delta_{1i} - \frac{\delta_i^*}{p_i}\pi_{1i})\mathbf{h}_i - D_{2,h^*}(\delta_{2i} - \frac{\delta_i^*}{p_i}\pi_{2i})\mathbf{h}_i\right\}, \tag{B.2}$$

where $\theta_{h^*} = E(\mathbf{h}^*)$,

$$\begin{aligned}(B_{1,h^*}, B_{2,h^*}) &= \sum_{i=1}^N (1 - \pi_{1i})(\mathbf{h}_i^* - \theta_{h^*})(\mathbf{h}_i', \mathbf{0}_{1\times r}) \\ &\times \left\{\sum_{i=1}^N p_i^{-1}\pi_{1i}\pi_{2i}\begin{pmatrix}(1 - \pi_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1 - \pi_{1i})(1 - \pi_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1 - \pi_{1i})(1 - \pi_{2i})\mathbf{h}_i\mathbf{h}_i' & (1 - \pi_{2i})\mathbf{h}_i\mathbf{h}_i'\end{pmatrix}\right\}^{-1}\end{aligned}$$

and

$$
\begin{aligned}
(D_{1,h^*}, D_{2,h^*}) &= \sum_{i=1}^{N}(1-\pi_{2i})(\mathbf{h}_i^* - \theta_{h^*})(\mathbf{0}_{1\times r}, \mathbf{h}_i') \\
&\times \left\{ \sum_{i=1}^{N} p_i^{-1}\pi_{1i}\pi_{2i} \begin{pmatrix} (1-\pi_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1-\pi_{1i})(1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1-\pi_{1i})(1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i' & (1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i' \end{pmatrix} \right\}^{-1}.
\end{aligned}
$$

Hence, according to (5.18), (B.1) and (B.2), we have

$$
\begin{aligned}
\hat{\theta}_{Reg} - \theta^* &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\delta_{1i}}{\pi_{1i}}\left\{y_i - \theta^* - B_{Reg}(\mathbf{h}_i^* - \theta_{h^*})\right\} + B_{Reg}\frac{\delta_{2i}}{\pi_{2i}}(\mathbf{h}_i^* - \theta_{h^*})\right] \\
&- \frac{1}{N}\sum_{i=1}^{N}\left\{ b_1(\delta_{1i} - \frac{\delta_i^*}{p_i}\pi_{1i})\mathbf{h}_i + b_2(\delta_{2i} - \frac{\delta_i^*}{p_i}\pi_{2i})\mathbf{h}_i \right\},
\end{aligned}
$$

where $b_1 = B_1 - B_{Reg}(B_{1,h^*} - D_{1,h^*})$ and $b_2 = B_2 - B_{Reg}(B_{2,h^*} - D_{2,h^*})$. Then, the variance of $\hat{\theta}_{Reg}$ can be written as

$$
\begin{aligned}
V(\hat{\theta}_{Reg}) &\cong V\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{1i}}{\pi_{1i}}\left\{y_i - \theta^* - B_{Reg}(\mathbf{h}_i^* - \theta_{h^*}) - b_1 p_i^{-1}\pi_{1i}\pi_{2i}\mathbf{h}_i + b_2 p_i^{-1}\pi_{1i}\pi_{2i}(1-\pi_{2i})\mathbf{h}_i\right\}\right] \\
&+ E\left[\frac{1}{N^2}\sum_{i=1}^{N}\pi_{2i}(1-\pi_{2i})\left\{B_{Reg}\pi_{2i}^{-1}(\mathbf{h}_i^* - \theta_{h^*}) + b_1 p_i^{-1}(1-\delta_{1i})\pi_{1i}\mathbf{h}_i - b_2 p_i^{-1}\pi_{1i}^*\mathbf{h}_i\right\}^2\right],
\end{aligned}
$$

where $\pi_{1i}^* = \pi_{1i} - \pi_{1i}\pi_{2i} + \delta_{1i}\pi_{2i}$. The consistent estimator of $V(\hat{\theta}_{Reg})$ can be written as

$$
\hat{V}(\hat{\theta}_{Reg}) = \frac{1}{\hat{N}^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}\hat{\eta}_{1i}^2 + \frac{1}{\hat{N}^2}\sum_{i\in A_1\cup A_2}\frac{\hat{\pi}_{2i}(1-\hat{\pi}_{2i})}{\hat{p}_i}\hat{\eta}_{2i}^2, \tag{B.3}
$$

where $\hat{N} = \sum_{i\in A_1\cup A_2}\hat{p}_i^{-1}$, $\hat{\eta}_{1i} = y_i - \hat{\theta}_{Reg} - \hat{B}_{Reg}(\mathbf{h}_i^* - \hat{\theta}_{h^*}) - \hat{b}_1\hat{p}_i^{-1}\hat{\pi}_{1i}\hat{\pi}_{2i}\mathbf{h}_i + \hat{b}_2\hat{p}_i^{-1}\hat{\pi}_{1i}\hat{\pi}_{2i}(1-\hat{\pi}_{2i})\mathbf{h}_i$ and $\hat{\eta}_{2i} = \hat{B}_{Reg}\hat{\pi}_{2i}^{-1}(\mathbf{h}_i^* - \hat{\theta}_{h^*}) + \hat{b}_1\hat{p}_i^{-1}(1-\delta_{1i})\hat{\pi}_{1i}\mathbf{h}_i - \hat{b}_2\hat{p}_i^{-1}\hat{\pi}_{1i}^*\mathbf{h}_i$, $\hat{\pi}_{1i}^* = \hat{\pi}_{1i} - \hat{\pi}_{1i}\hat{\pi}_{2i} + \delta_{1i}\hat{\pi}_{2i}$, $\hat{\theta}_{h^*} = \sum_{i\in A_1\cup A_2}\hat{p}_i^{-1}\mathbf{h}_i^* / \sum_{i\in A_1\cup A_2}\hat{p}_i^{-1}$, $\hat{b}_1 = \hat{B}_1 - \hat{B}_{Reg}(\hat{B}_{1,h^*} - \hat{D}_{1,h^*})$ and $\hat{b}_2 = \hat{B}_2 - \hat{B}_{Reg}(\hat{B}_{2,h^*} - \hat{D}_{2,h^*})$ with

$$
\begin{aligned}
(\hat{B}_{1,h^*}, \hat{B}_{2,h^*}) &= \sum_{i\in A_1\cup A_2}\frac{1-\hat{\pi}_{1i}}{\hat{p}_i}(\mathbf{h}_i^* - \hat{\theta}_{h^*})(\mathbf{h}_i', \mathbf{0}_{1\times r}) \\
&\times \left\{ \sum_{i\in A_1\cup A_2}\hat{p}_i^{-2}\hat{\pi}_{1i}\hat{\pi}_{2i} \begin{pmatrix} (1-\hat{\pi}_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' & (1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \end{pmatrix} \right\}^{-1}
\end{aligned}
$$

and

$$(\hat{D}_{1,h^*}, \hat{D}_{2,h^*}) = \sum_{i \in A_1 \cup A_2} \frac{1-\hat{\pi}_{2i}}{\hat{p}_i}(\mathbf{h}_i^* - \hat{\theta}_{h^*})(\mathbf{0}_{1\times r}, \mathbf{h}_i') \times \left\{ \sum_{i \in A_1 \cup A_2} \hat{p}_i^{-2}\hat{\pi}_{1i}\hat{\pi}_{2i} \begin{pmatrix} (1-\hat{\pi}_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' & (1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \end{pmatrix} \right\}^{-1}.$$

$\hat{B}_{Reg}$ is defined in Section 4. Define $U_{h1} = N^{-1}\sum_{i=1}^N \delta_{1i}\pi_{1i}^{-1}(\mathbf{h}_i^* - \theta_{h^*})$ and $U_{h2} = N^{-1}\sum_{i=1}^N \delta_{2i}\pi_{2i}^{-1}(\mathbf{h}_i^* - \theta_{h^*})$, with $\theta_{h^*} = E(\mathbf{h}^*)$. The optimal estimator which incorporates sample based information can be written as

$$\hat{\theta}_{opt} = \hat{\theta} - \hat{B}_{opt}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*), \tag{B.4}$$

where $\hat{B}_{opt} = \hat{C}(\hat{\theta}, \hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*)\left\{\hat{V}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*)\right\}^{-1}$ with

$$\begin{aligned} \hat{C}(\hat{\theta}, \hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*) &= \hat{C}(U_p, U_{h1}) + \hat{C}(U_p, S_C)(\hat{B}_{h2} - \hat{B}_{h1})' - \hat{B}\hat{C}(S_C, U_{h1}) \\ &+ \hat{B}\hat{C}(S_C, U_{h2}) - \hat{B}\hat{V}(S_C)(\hat{B}_{h2} - \hat{B}_{h1})'. \end{aligned} \tag{B.5}$$

$$\begin{aligned} \hat{V}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*) &= \hat{V}(U_{h1}) + \hat{V}(U_{h2}) + (\hat{B}_{h2} - \hat{B}_{h1})\hat{V}(S_C)(\hat{B}_{h2} - \hat{B}_{h1})' \\ &+ \hat{C}(U_{h1}, S_C)(\hat{B}_{h2} - \hat{B}_{h1})' + (\hat{B}_{h2} - \hat{B}_{h1})\hat{C}(S_C, U_{h1}) \\ &- \hat{C}(U_{h2}, S_C)(\hat{B}_{h2} - \hat{B}_{h1})' - (\hat{B}_{h2} - \hat{B}_{h1})\hat{C}(S_C, U_{h2}), \end{aligned} \tag{B.6}$$

where

$$\hat{C}(U_p, U_{h1}) = \frac{1}{\hat{N}^2}\sum_{i \in A_1} \frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}(y_i - \hat{\theta})(\mathbf{h}_i^* - \hat{\theta}_{h^*})'.$$

$$\hat{C}(S_{C1}, U_{h1}) = \frac{1}{\hat{N}}\sum_{i \in A_1 \cup A_2}(\frac{1}{\hat{p}_i} - \frac{\hat{\pi}_{1i}}{\hat{p}_i^2})\mathbf{h}_i(\mathbf{h}_i^* - \hat{\theta}_{h^*})', \quad \hat{C}(S_{C2}, U_{h1}) = \frac{1}{\hat{N}}\sum_{i \in A_1 \cup A_2}(\frac{\hat{\pi}_{2i}}{\hat{p}_i} - \frac{\hat{\pi}_{2i}}{\hat{p}_i^2})\mathbf{h}_i(\mathbf{h}_i^* - \hat{\theta}_{h^*})'.$$

$$\hat{C}(S_{C1}, U_{h2}) = \frac{1}{\hat{N}}\sum_{i \in A_1 \cup A_2}(\frac{\hat{\pi}_{1i}}{\hat{p}_i} - \frac{\hat{\pi}_{1i}}{\hat{p}_i^2})\mathbf{h}_i(\mathbf{h}_i^* - \hat{\theta}_{h^*})', \quad \hat{C}(S_{C2}, U_{h2}) = \frac{1}{\hat{N}}\sum_{i \in A_1 \cup A_2}(\frac{1}{\hat{p}_i} - \frac{\hat{\pi}_{2i}}{\hat{p}_i^2})\mathbf{h}_i(\mathbf{h}_i^* - \hat{\theta}_{h^*})'.$$

$$\hat{V}(U_{h1}) = \frac{1}{\hat{N}^2}\sum_{i \in A_1 \cup A_2}\frac{1-\hat{\pi}_{1i}}{\hat{p}_i\hat{\pi}_{1i}}(\mathbf{h}_i^* - \hat{\theta}_{h^*})^{\otimes 2}, \quad \hat{V}(U_{h2}) = \frac{1}{\hat{N}^2}\sum_{i \in A_1 \cup A_2}\frac{1-\hat{\pi}_{2i}}{\hat{p}_i\hat{\pi}_{2i}}(\mathbf{h}_i^* - \hat{\theta}_{h^*})^{\otimes 2},$$

with $\hat{B} = (\hat{B}_1, \hat{B}_2)$, and $\hat{B}_1$, $\hat{B}_2$ defined in Section 4, $\hat{\theta}_{h^*} = \hat{N}^{-1}\sum_{i \in A_1 \cup A_2}\hat{p}_i^{-1}\mathbf{h}_i^*$, $\hat{N} = \sum_{i \in A_1 \cup A_2}\hat{p}_i^{-1}$,

$$\hat{B}_{h1} = -\frac{1}{\hat{N}}\sum_{i \in A_1 \cup A_2}\frac{1-\hat{\pi}_{1i}}{\hat{p}_i}(\mathbf{h}_i^* - \hat{\theta}_{h^*})(\mathbf{h}_i', \mathbf{0}_{1\times r})\left\{\hat{E}(\frac{\partial S_C}{\partial \Phi})\right\}^{-1},$$

$$\hat{B}_{h2} = -\frac{1}{\hat{N}} \sum_{i \in A_1 \cup A_2} \frac{1-\hat{\pi}_{2i}}{\hat{p}_i} (\mathbf{h}_i^* - \hat{\theta}_{h^*})(\mathbf{0}_{1\times r}, \mathbf{h}_i') \left\{ \hat{E}(\frac{\partial S_C}{\partial \Phi}) \right\}^{-1},$$

and $\hat{E}(\partial S_C/\partial \Phi)$ is defined in Section 4. Next, we want to derive the asymptotic properties for $\hat{\theta}^*_{Reg}$. By linearization, we have

$$\hat{\theta}_{x,1} - \bar{\mathbf{X}}_N \cong \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\delta_{1i}}{\pi_{1i}} (\mathbf{x}_i - \bar{\mathbf{X}}_N) - B_{1,x}(\delta_{1i} - \frac{\delta_i^*}{p_i}\pi_{1i})\mathbf{h}_i - B_{2,x}(\delta_{2i} - \frac{\delta_i^*}{p_i}\pi_{2i})\mathbf{h}_i \right\}, \quad \text{(B.7)}$$

where

$$\begin{aligned} (B_{1,x}, B_{2,x}) &= \sum_{i=1}^{N} (1-\pi_{1i})(\mathbf{x}_i - \bar{\mathbf{X}}_N)(\mathbf{h}_i', \mathbf{0}_{1\times r}) \\ &\times \left\{ \sum_{i=1}^{N} p_i^{-1} \pi_{1i}\pi_{2i} \begin{pmatrix} (1-\pi_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1-\pi_{1i})(1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1-\pi_{1i})(1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i' & (1-\pi_{2i})\mathbf{h}_i\mathbf{h}_i' \end{pmatrix} \right\}^{-1}. \end{aligned}$$

According to (5.18), (5.21), (B.1), (B.2) and (B.7), we have

$$\begin{aligned} \hat{\theta}^*_{Reg} - \theta^* &= \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\delta_{1i}}{\pi_{1i}} \left\{ y_i - \theta^* - B_{1,Reg}(\mathbf{h}_i^* - \theta_{h^*}) - B_{2,Reg}(\mathbf{x}_i - \bar{\mathbf{X}}_N) \right\} + B_{1,Reg} \frac{\delta_{2i}}{\pi_{2i}} (\mathbf{h}_i^* - \theta_{h^*}) \right] \\ &- \frac{1}{N} \sum_{i=1}^{N} \left\{ b_1^*(\delta_{1i} - \frac{\delta_i^*}{p_i}\pi_{1i})\mathbf{h}_i + b_2^*(\delta_{2i} - \frac{\delta_i^*}{p_i}\pi_{2i})\mathbf{h}_i \right\}, \end{aligned}$$

where $b_1^* = B_1 - B_{1,Reg}(B_{1,h^*} - D_{1,h^*}) - B_{2,Reg}B_{1,x}$ and $b_2^* = B_2 - B_{1,Reg}(B_{2,h^*} - D_{2,h^*}) - B_{2,Reg}B_{2,x}$. As before, the consistent estimator of $V(\hat{\theta}^*_{Reg})$ can be written as

$$\hat{V}(\hat{\theta}^*_{Reg}) = \frac{1}{\hat{N}^2} \sum_{i \in A_1} \frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2} \hat{\eta}_{1i}^{*2} + \frac{1}{\hat{N}^2} \sum_{i \in A_1 \cup A_2} \frac{\hat{\pi}_{2i}(1-\hat{\pi}_{2i})}{\hat{p}_i} \hat{\eta}_{2i}^{*2}, \quad \text{(B.8)}$$

where $\hat{\eta}^*_{1i} = y_i - \hat{\theta}^*_{Reg} - \hat{B}_{1,Reg}(\mathbf{h}_i^* - \hat{\theta}_{h^*}) - \hat{B}_{2,Reg}(\mathbf{x}_i - \bar{\mathbf{X}}_N) - \hat{b}_1^* \hat{p}_i^{-1} \hat{\pi}_{1i}\hat{\pi}_{2i}\mathbf{h}_i + \hat{b}_2^* \hat{p}_i^{-1} \hat{\pi}_{1i}\hat{\pi}_{2i}(1-\hat{\pi}_{2i})\mathbf{h}_i$ and $\hat{\eta}_{2i} = \hat{B}_{1,Reg}\hat{\pi}_{2i}^{-1}(\mathbf{h}_i^* - \hat{\theta}_{h^*}) + \hat{b}_1^*\hat{p}_i^{-1}(1-\delta_{1i})\hat{\pi}_{1i}\mathbf{h}_i - \hat{b}_2^*\hat{p}_i^{-1}\hat{\pi}^*_{1i}\mathbf{h}_i$, $\hat{b}_1^* = \hat{B}_1 - \hat{B}_{1,Reg}(\hat{B}_{1,h^*} - \hat{D}_{1,h^*}) - \hat{B}_{2,Reg}\hat{B}_{1,x}$ and $\hat{b}_2^* = \hat{B}_2 - \hat{B}_{1,Reg}(\hat{B}_{2,h^*} - \hat{D}_{2,h^*}) - \hat{B}_{2,Reg}\hat{B}_{2,x}$, with

$$\begin{aligned} (\hat{B}_{1,x}, \hat{B}_{2,x}) &= \sum_{i \in A_1 \cup A_2} \frac{1-\hat{\pi}_{1i}}{\hat{p}_i} (\mathbf{x}_i - \bar{\mathbf{X}}_N)(\mathbf{h}_i', \mathbf{0}_{1\times r}) \\ &\times \left\{ \sum_{i \in A_1 \cup A_2} \hat{p}_i^{-2} \hat{\pi}_{1i}\hat{\pi}_{2i} \begin{pmatrix} (1-\hat{\pi}_{1i})\mathbf{h}_i\mathbf{h}_i' & -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \\ -(1-\hat{\pi}_{1i})(1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' & (1-\hat{\pi}_{2i})\mathbf{h}_i\mathbf{h}_i' \end{pmatrix} \right\}^{-1}. \end{aligned}$$

Other terms are defined before. Define $U_{x1} = N^{-1} \sum_{i \in A_1} \pi_{1i}^{-1}(\mathbf{x}_i - \bar{\mathbf{X}}_N)$, then the optimal estimator can be written as

$$\hat{\theta}^*_{opt} = \hat{\theta} - \hat{B}_{1,opt}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*) - \hat{B}_{2,opt}(\hat{\theta}_{x,1} - \bar{\mathbf{X}}_N), \quad \text{(B.9)}$$

where

$$(\hat{B}_{1,opt}, \hat{B}_{2,opt}) = \hat{C}\left\{\hat{\theta}, (\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*, \hat{\theta}_{x,1})\right\}\left[\hat{V}\left\{(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*, \hat{\theta}_{x,1})\right\}\right]^{-1},$$

with $\hat{C}(\hat{\theta}, \hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*)$, $\hat{V}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*)$ defined in (B.5) and (B.6),

$$\hat{C}(\hat{\theta}, \hat{\theta}_{x,1}) = \hat{C}(U_p, U_{x1}) - \hat{C}(U_p, S_C)\hat{B}_x' - \hat{B}\hat{C}(S_C, U_{x1}) + \hat{B}\hat{V}(S_C)\hat{B}_x',$$

$$\begin{aligned} \hat{C}(\hat{\mathbf{h}}_1^* - \hat{\mathbf{h}}_2^*, \hat{\theta}_{x,1}) &= \hat{C}(U_{h1}, U_{x1}) - \hat{C}(U_{h1}, S_C)\hat{B}_x' + \hat{C}(U_{h2}, S_C)\hat{B}_x' \\ &+ (\hat{B}_{h2} - \hat{B}_{h1})\hat{C}(S_C, U_{x1}) - (\hat{B}_{h2} - \hat{B}_{h1})\hat{V}(S_C)\hat{B}_x', \end{aligned}$$

$$\hat{V}(\hat{\theta}_{x,1}) = \hat{V}(U_{x1}) - \hat{C}(U_{x1}, S_C)\hat{B}_x' - \hat{B}_x\hat{C}(S_C, U_{x1}) + \hat{B}_x\hat{V}(S_C)\hat{B}_x',$$

where

$$\hat{C}(U_p, U_{x1}) = \frac{1}{\hat{N}^2}\sum_{i\in A_1}\frac{1-\hat{\pi}_{1i}}{\hat{\pi}_{1i}^2}(y_i - \hat{\theta})(\mathbf{x}_i - \bar{\mathbf{X}}_N)',$$

$$\hat{C}(S_{C1}, U_{x1}) = \frac{1}{\hat{N}}\sum_{i\in A_1\cup A_2}(\frac{1}{\hat{p}_i} - \frac{\hat{\pi}_{1i}}{\hat{p}_i^2})\mathbf{h}_i(\mathbf{x}_i - \bar{\mathbf{X}}_N)', \quad \hat{C}(S_{c2}, U_{x1}) = \frac{1}{\hat{N}}\sum_{i\in A_1\cup A_2}(\frac{\hat{\pi}_{2i}}{\hat{p}_i} - \frac{\hat{\pi}_{2i}}{\hat{p}_i^2})\mathbf{h}_i(\mathbf{x}_i - \bar{\mathbf{X}}_N)',$$

$$\hat{C}(U_{h1}, U_{x1}) = \frac{1}{\hat{N}^2}\sum_{i\in A_1\cup A_2}\frac{1-\hat{\pi}_{1i}}{\hat{p}_i\hat{\pi}_{1i}}(\mathbf{h}_i^* - \hat{\theta}_{h^*})(\mathbf{x}_i - \bar{\mathbf{X}}_N)', \quad \hat{V}(U_{x1}) = \frac{1}{\hat{N}^2}\sum_{i\in A_1\cup A_2}\frac{1-\hat{\pi}_{1i}}{\hat{p}_i\hat{\pi}_{1i}}(\mathbf{x}_i - \bar{\mathbf{X}}_N)^{\otimes 2},$$

$$\hat{B}_x = -\hat{N}^{-1}\sum_{i\in A_1\cup A_2}\frac{1-\hat{\pi}_{1i}}{\hat{p}_i}(\mathbf{x}_i - \bar{\mathbf{X}}_N)(\mathbf{h}_i', \mathbf{0}_{1\times r})\left\{\hat{E}(\partial S_C/\partial\Phi)\right\}^{-1},$$

with other terms defined before.

BIBLIOGRAPHY

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, *46*, 623-635.

Amstrup, S. C., McDonald, T. L. and Manly, B. F. J. (2005). Handbook of capture-recapture analysis. *Princeton University Press, Princeton, NJ.*

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, *10*, 1100-1120.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*, 200-203.

Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika*, *85*, 535-547.

Berger, Y. G., and De La Riva Torres, O. (2012). A unified theory of empirical likelihood ratio confidence intervals for survey data with unequal probabilities and non negligible sampling fractions. *http : //eprints.soton.ac.uk/*337688/ *Southampton Statistical Sciences Research Institute (S3RI Methodology Working Papers)*, 24pp.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279-292.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, *155*, 138-154.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*, 305-334.

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, *95*, 555-571.

Chaudhuri, S., Handcock, M. S., and Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B*, *70*, 311–328.

Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, *21*, 1166-1181.

Chen, S. X. and Cui, H. J. (2006). On bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, *93*, 215-220.

Chen, S. X., Leung, D. H. Y. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of Royal Statistical Society: Series B*, *70*, 803-823.

Chen, S. X., Peng, L. and Qin, Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, *96*, 711-722.

Chen, S. X., Tang, C. Y., and Mule, V. T. (2010). Local post-stratification and diagnostics in dual system accuracy and coverage evaluation for the U.S. Census. *Journal of the American Statistical Association*, *105*, 105-119.

Chen, S. X. and Tang, C. Y. (2011). Properties of census dual system population size estimators. *International Statistical Review*, *79*, 336-361.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, *80*, 107-116.

Chen, J. H. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, *9*, 385-406.

Chen, J. H. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* *80*, 107-116.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, *89*, 81-87.

Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B*, *46*, 440-464.

Da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, *35*, 165-176.

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

DiCiccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is bartlett-correctable. *The Annals of Statistics*, *19*, 1053-1061.

Dupont, F. (1995). Alternative adjustment where there are several levels of auxiliary information. *Survey Methodology*, *21*, 125-135.

Duncan, K. B. and Stasny, E. A. (2001). Using propensity scores to control coverages bias in telephone surveys. *Survey Methodology*, *27*, 121-130.

Durrant, G. B. and Skinner, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Survey Methodology*, *32*, 25-36.

Fang, F., Hong, Q. and Shao, J. (2010). Empirical likelihood estimation for samples with nonignorable nonresponse. *Statistica Sinica*, *20*, 263-280.

Fuller, W. A. and Isaki, C. (1981). Survey design under superpopulation models. *Current Topics in Survey Sampling*. Academic press, New York, 199-226.

Fuller, W. A., Loughin, M. M. and Baker, H. D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*, *20*, 75-85.

Fuller, W. A. (2002). Regression estimation for sample surveys. *Survey Methodology*, *28*, 5-23.

Fuller, W. A. (2009a). Some design properties of a rejective sampling procedure. *Biometrika*, *96*, 933-944.

Fuller, W. A. (2009b). *Sampling Statistics*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Godambe, V. P. and Joshi, V. M. (1965). Admissibility and bayes estimation in sampling finite populations. *The Annals of Mathematical Statistics*, *36*, 1707-1722.

Groves, R., Eleanor S. and Amy C. (2000). Leverage-salience theory of survey participation: descriptiona and illustration. *Public Opinion Quarterly*, *64*, 299-308.

Groves, R., Presser, S. and Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *68*, 2-31.

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

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica*, *64*, 891-916.

Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, *35*, 1491-1523.

Hajek, J. (1981). Sampling from a Finite Population. *Marcel Dekker, New York.*

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, *55*, 547557.

Hidiroglou, M. A. and Sarndal, C. E. (1998). Use of auxiliary information for twophase sampling. *Survey Methodology*, *24*, 11-20.

Hidiroglou. M. A. (2001). Double sampling. *Survey Methodology*, *27*, 143-154.

Hirano, K., Imbens, G., Ridder, G., and Rubin, D. (1998). Combining panel data sets with attrition and refreshment samples. *NBER Technical Working Paper No. 230.*

Hirano, K., Imbens, G, W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161-1189.

Hjort, N. L., McKeague, I., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, *37*, 1079-1111.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, *76*, 133-140.

Huggins, R. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, *47*, 725-732.

Huggins, R. and Hwang, W. H. (2011). A review of the use of conditional likelihood in capture-recapture experiments. *International Statistical Review*, *79*, 385400.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies, Wiley Blackwell*, *61*, 655-680.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Economic Studies*, *64*, 359-383.

Imbens, G. W., Spady, R. H., and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, *66*, 333-357.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, *20*, 493-506.

Jing, B. Y., Yuan, J. Q. and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, *104*, 1224-1232.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, *35*, 501-514.

Kim, J. K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, *19*, 145-157.

Kim, J. K. and Rao, J. N. K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, *96*, 917-932.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, *78*, 21-39.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, *98*, 119-132.

Kim, J. K. and Yu, C. L. (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data. *Journal of the American Statistical Association*, *106*, 157-165.

Kim, J. K., and Yu, C. L. (2011). Replication variance estimation under two-phase sampling. *Survey Methodology*, *37*, 67-74.

Kim, J. K. and Riddles, M. (2012). Some theory for propensity scoring adjustment estimator. *Survey Methodology, Accepted for publication*.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodol*, *32*, 13342.

Kott, P. S and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, *105*, 1265-1275.

Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, *65*, 861-874.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, *25*, 2084-2102.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *44*, 226-233.

Liu, Y. and Chen, J. (2010). Adjusted empirical likelihood with high-order precision. *Annals of Statistics*, *38*, 1341-1362.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*, 329-349.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods* & *Research*, *37*, 319-43.

Legg, J. C. and Fuller, W. A. (2009). Two-phase sampling. In: D. Pfeffermann and C. R. Rao (eds), *Sample Surveys: Theory, Methods and Inference.* Wiley, New York.

Montanari, G. E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, *55*, 191-202.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, *99*, 1131-1139.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, *72*, 219-255.

Nordman, D. J., Sibbertsen, P. and Lahiri, S. N. (2007). Empirical likelihood for the mean under long-range dependence. *Journal of Time Series Analysis*, *28*, 576-599.

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. *Annals of Statistics*, *18*, 90-120.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, New York.

Pollock, K. H., Hines, J. E. and Nichols, J. D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics* *40*, 329-340.

Qin, J. (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics*, *21*, 1182-1196.

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

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, *87*, 484-490.

Qin, J, Leung, D and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, *97*, 193-200.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of Royal Statistical Society: Series B*, *69*, 101-122.

Qin, J, Zhang, B and Leung D. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association*, *104*, 1492-1503.

Rao, C. R. (1973). *Linear statistical inference and its applications.* NY Wiley.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, *10*, 153-165.

Rao, J. N. K. and Wu. C. (2009). Empirical Likelihood Methods. *Sample Surveys: Theory, Methods and Inference, Handbook of Statistics, Vol. 29, ed. D. Pfeermann and C. R. Rao, Amsterdam: North Holland.*

Renssen, R. H. and Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, *92*, 368-374.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846-866.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*, 122-129.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387-394.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, *35*, 634-672.

Serfling, R. J. (1980). Approximation theorems of mathematical statistics. *New York: Wiley.*

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. *Analysis of Complex Surveys*, Wiley, New York, 59-87.

Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal*, *107*, 503-519.

Tan, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika*, *98*, 663-684.

Tang, C. Y. and Leng, C. (2010). Penalized high dimensional empirical likelihood. *Biometrika*, *97*, 905-920.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, *13*, 178-203.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*, 105-137.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, *30*, 896-924.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, *37*, 490-517.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, *32*, 15-26.

Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, *34*, 359-375.

Xue, L. (2009). Empirical likelihood for linear models with missing responses. *Journal of Multivariate Analysis*, *100*, 1353-1366.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, *85*, 986-1001.