5-9-2015

Empirical Likelihood Confidence Intervals for the Population Mean Based on Incomplete Data

Jose Manuel Valdovinos Alvarez

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE POPULATION MEAN BASED ON INCOMPLETE DATA

by

JOSE M. VALDOVINOS ALVAREZ

Under the Direction of Yichuan Zhao, PhD

ABSTRACT

The use of doubly robust estimators is a key for estimating the population mean response in the presence of incomplete data. Cao et al. (2009) proposed an alternative doubly robust estimator which exhibits strong performance compared to existing estimation methods. In this thesis, we apply the jackknife empirical likelihood, the jackknife empirical likelihood with nuisance parameters, the profile empirical likelihood, and an empirical likelihood method based on the influence function to make an inference for the population mean. We use these methods to construct confidence intervals for the population mean, and compare the coverage probabilities and interval lengths using both the "usual" doubly robust estimator and the alternative estimator proposed by Cao et al. (2009). An extensive simulation study is carried out to compare the different methods. Finally, the proposed methods are applied to two real data sets.

INDEX WORDS: Missing data, Doubly robust estimation, Empirical likelihood, Jackknife empirical likelihood, Profile empirical likelihood, Coverage probability, Interval length

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE POPULATION MEAN BASED ON INCOMPLETE DATA

by

JOSE M. VALDOVINOS ALVAREZ

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2015

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR THE POPULATION MEAN BASED ON INCOMPLETE DATA

by

JOSE M. VALDOVINOS ALVAREZ

Committee Chair: Yichuan Zhao

Committee: Jing Zhang

Xin Qi

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2015

DEDICATION

Dedicado a mis padres Jose Valdovinos y Maria Cruz Alvarez.

Todo es gracias a ustedes.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my committee chair Dr. Yichuan Zhao. His guidance, support, knowledge, and encouragement have helped me carry out the most important academic project of my life to date. It has been an extraordinary experience which will help me succeed for the rest of my life.

I also want to thank Dr. Jing Zhang and Dr. Xin Qi for agreeing to be part of my committee. Thank you for your helpful questions and sharing your knowledge with me.

Finally, I want to thank all of my professors at GSU. I have obtained an inmense amount of knowledge from each and every one of you. I am very happy for choosing this school, and getting the opportunity to meet you all.

TABLE OF CONTENTS

LIST OF TABLES

1 INTRODUCTION

As it is well known, missing data is a common problem which can affect inferences obtained from data. Some settings in which missing outcomes arise include non-response in sample surveys and patient dropouts during clinical trials. Further, making a causal inference on a treatment mean from an experiment, or observational study, may be viewed as a missing data problem (Kang and Schafer, 2007).

Doubly robust estimators have been proposed to alleviate the bias obtained from using the naive sample mean based on the complete cases only. These estimators require the specification of an outcome regression model to describe the population of responses, and a propensity scores model to describe the missingness mechanism observed in the data. Although these estimators are consistent as long as one of the two models is correctly specified (Scharfstein et al., 1999), Kang and Schafer (2007) revealed that the usual doubly robust estimator can be severely biased if both models are even mildly incorrectly specified.

In the remainder of this chapter, we introduce the basic concepts that were used for this research. First, we present the ideas behind the USUAL doubly robust estimator and the alternative estimator proposed by Cao et al. (2009) which we denote as the "PROJ" estimator. Next, we present the methodology of empirical likelihood (EL), acknowledge some of its applications, and introduce methods based on EL. Finally, we conclude this chapter with a description of the organization that the thesis will follow.

1.1 Doubly Robust estimators

Consider a population of interest for which we have a random sample of n observations. Let X_i denote the vector of covariates and Y_i be the response or outcome of observation i. Now, consider the case where Y_i is missing for some subjects; in this case, we can introduce a dummy variable, R_i, to indicate whether the response was observed ($R_i = 1$) or is missing ($R_i = 0$). Hence, the data observed is independent and identically distributed (R_iY_i, R_i, X_i) for $i = 1, \ldots, n$. As in Rubin (1978), we assume that the data is missing at random (i.e. Y_i and R_i are conditionally independent given X_i) in order to estimate the population mean μ.

Denote the propensity score and the outcome regression by $P(R = 1|X)$ and $E(Y|X)$ respectively. Since the true propensity scores are rarely known, we can use a logistic regression model $\pi(X, \gamma) = \{1 + exp(-\tilde{X}\gamma)\}^{-1}$, where $\tilde{X} = (1, X)$, to estimate them. Similarly, we can adopt a model $m(X, \beta)$ for $E(Y|X)$, where we estimate β using only the complete cases $\{i|R_i = 1\}$. Thus, by combining both models, we can obtain the USUAL doubly robust estimator

$$\hat{\mu}_{USUAL} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{R_i Y_i}{\pi(X_i, \hat{\gamma})} - \frac{R_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} m(X_i, \hat{\beta}) \right\}, \tag{1.1}$$

where γ is estimated by maximum likelihood and β is estimated using ordinary or weighted least squares. Scharfstein et al. (1999) noted that this estimator is consistent as long as at least one of the two models is correctly specified but is inconsistent otherwise.

Using a thorough simulation scenario, Kang and Schafer (2007) noted that the USUAL doubly robust estimator may exhibit poor performance when some of the estimated propensity scores are close to 0. Among several strategies proposed by Tsiatis and Davidian (2007) to improve the performance of this estimator, these authors used semiparametric theory to argue that the method used to estimate β may be a strong influence to the poor performance of the estimator.

In order to find an estimator for μ, in the form of eqn. (1.1), that is (i) doubly robust and (ii) has the smallest asymptotic variance when the propensity model is correct, Cao et al. (2009) proposed to estimate β by solving the following equation jointly in (β, c)

$$\sum_{i=1}^{n} \left[\frac{R_i}{\pi(X_i, \hat{\gamma})} \frac{1 - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} \left\{ \frac{\frac{m_\beta(X_i, \beta)}{\pi_\gamma(X_i, \hat{\gamma})}}{1 - \pi(X_i, \hat{\gamma})} \right\} \left\{ Y_i - m(X_i, \beta) - c^T \frac{\pi_\gamma(X_i, \hat{\gamma})}{1 - \pi(X_i, \hat{\gamma})} \right\} \right] = 0. \tag{1.2}$$

Thus, if we take $\hat{\beta}^*$ as the solution to eqn. (1.2), $\hat{\gamma}$ as the maximum likelihood estimator for the propensity scores model, and plug these estimators into eqn. (1.1), we obtain the alternative doubly robust estimator $\hat{\mu}_{PROJ}$.

1.2 Empirical likelihood

Empirical Likelihood is a nonparametric methodology which was first introduced by Owen (1988, 1990). The method can be used to construct confidence regions and perform hypothesis tests without any distributional assumptions. Furthermore, EL can incorporate side information in the form of constraints to the likelihood function, and it has the ability to construct confidence regions for the parameter of interest without estimating complicated covariance matrices. For a more thorough review, we refer the reader to Owen (2001).

Due to its simplicity and attractive properties, EL has found applications in areas such as in regression models (Chen and Van Keilegom, 2009), quantile estimation (Chen and Hall, 1993), the accelerated failure time model (Zhao, 2011), and continuous scale diagnostic tests in the presence of verification bias (Wang and Qin, 2013), to name only a few. Of particular importance are the papers by Qin and Lawless (1994), and Hjort et al. (2009), which linked the concepts of empirical likelihood, general estimating equations, and nuisance parameters. Since then, EL has been applied to many different contexts.

Regarding missing outcome data problems, Qin and Zhang (2007) proposed an EL based estimator for a response mean with the double robustness property when the outcomes might be missing at random. Chan (2012) studied modifications of the EL estimator of Qin and Zhang (2007) that attains uniform improvements in asymptotic efficiency. Xue and Xue (2011) used a bias-correction technique to construct EL ratios to study a semi-parametric model with missing response data. Similarly, Tang and Zhao (2013) developed inferences for a semi-parametric nonlinear regression model for longitudinal data in the presence of missing responses. Zhao et al. (2013) developed an EL approach to obtain inference for mean functionals with nonignorably missing data. Further, Tang et al. (2014) developed EL inferences on parameters in generalized estimating equations with nonignorably missing response data.

In order to overcome the computational difficulties that EL encounters when handling nonlinear statistics, Jing et al. (2009) proposed a new approach: the jackknife empirical

likelihood (JEL). This method converts the statistic of interest into a sample mean of jackknife pseudo-values (Quenouille, 1956). The attractiveness of this methodology lies behind its simplicity, as it merely combines the EL method of Owen to the sample mean of the jackknife pseudo-values.

Zhong and Chen (2014) proposed JEL methods for constructing confidence intervals for a population mean with regression imputation or non-ignorable missingness. Likewise, Gong et al. (2010) proposed a smoothed JEL method to construct confidence intervals for the receiver operating characteristic (ROC) curve. Further applications of JEL include a test for the equality of 2 high dimensional means (Wang et al., 2013), the accelerated failure time model with censored data (Bouadoumou et al., 2015), and confidence intervals for the difference between two ROC curves (Yang and Zhao, 2013).

Motivated by Jing et al. (2009), Li et al. (2011) proposed a JEL method to construct confidence regions for a parameter of interest in the presence of nuisance parameters. This method allows the computation of the nuisance parameter through a subset of estimating equations. Furthermore, it retains the property of the standard chi-square limiting distribution of the EL ratio. Peng (2012) proposed an approximate JEL method to reduce the computation of the JEL method when the parameters cannot be explicitly estimated.

1.3 Organization

In this thesis, we adapt the methodologies of Jing et al. (2009), Li et al. (2011), and Wang and Qin (2013) to construct confidence intervals for a population mean in the presence of incomplete data. We also develop a new approach based on the influence functions of the doubly robust estimators to construct confidence intervals for the parameter of interest. In Chapter 2, we give a detailed overview of the methodologies and develop their application to our scenario. In Chapter 3, we carry out an extensive simulation study to compare the efficiencies of the proposed methods in terms of coverage probabilities and average lengths of the confidence intervals. Next, in Chapter 4, we apply the proposed methods to two real data sets. Finally, we give a concluding discussion of our work in Chapter 5.

2 INFERENCE METHODS

2.1 Jackknife empirical likelihood

In order to apply the JEL technique of Jing et al. (2009), we first need to estimate the unknown quantities to plug into eqn.(1.1). As mentioned earlier, propensity scores and predicted outcomes can be estimated by using a logistic regression model and a linear model respectively.

Define $\hat{\pi}_i = \pi(x_i, \hat{\gamma})$, where $\hat{\gamma}$ is the MLE for γ; $\hat{m}_i = m(x_i, \hat{\beta})$, where $\hat{\beta}$ is the OLS or WLS estimator for β; and $\hat{m}_i{}^* = m(x_i, \hat{\beta}^*)$, where $\hat{\beta}^*$ is the solution to eqn.(1.2). We illustrate the proposed method by using the results of the USUAL doubly robust estimation methodology. Inferences for the methodology proposed by Cao et al. (2009) are obtained by replacing $\hat{m}_i$ with $\hat{m}_i{}^*$.

Let $Z_1, \ldots, Z_n$ be independent random variables, where

$$Z_i = (R_i Y_i, R_i, x_i, \hat{m}_i, \hat{\pi}_i).$$

A consistent estimator for μ is given by

$$T_n = n^{-1} \sum_{i=1}^{n} h(Z_i) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{R_i Y_i}{\hat{\pi}_i} - \frac{R_i - \hat{\pi}_i}{\hat{\pi}_i} \hat{m}_i \right\}.$$

Define the jackknife pseudo-values by

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

where $T_{n-1}^{(-i)}$ is the statistic T_{n-1} computed from the $n-1$ observations from the original data set after deleting the ith data value in which $\hat{\gamma}$ and $\hat{\beta}$ are obtained from the full sample as in T_n. Thus, we can obtain the jackknife estimator of μ

$$\hat{T}_{n,jack} = \frac{1}{n} \sum_{i=1}^{n} \hat{V}_i,$$

which is just the sample average of asymptotically independent random variables $\hat{V}_i$'s (Shi, 1984).

We can now apply Owen's EL. Let $\mathbf{p} = (p_1, \ldots, p_n)$ be a vector of weights such that $\sum_{i=1}^{n} p_i = 1$ and $p_i > 0$ for $i = 1, \ldots, n$. The empirical likelihood evaluated at μ is

$$L(\mu) = max \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i \hat{V}_i = \mu, \sum_{i=1}^{n} p_i = 1, p_i > 0 \right\},$$

from which we can obtain the jackknife empirical likelihood ratio at μ

$$R(\mu) = \frac{L(\mu)}{n^{-n}} = max \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i \hat{V}_i = \mu, \sum_{i=1}^{n} p_i = 1, p_i > 0 \right\}.$$

Using the method of Lagrange multipliers, we have

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda(\hat{V}_i - \mu)},$$

where λ is the solution to

$$f(\lambda) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{V}_i - \mu}{1 + \lambda(\hat{V}_i - \mu)} = 0.$$

By taking the logarithm of $R(\mu)$ and plugging in the computed values of p_i into $logR(\mu)$, we obtain the nonparametric jackknife empirical log-likelihood ratio

$$logR(\mu) = -\sum_{i=1}^{n} log\{1 + \lambda(\hat{V}_i - \mu)\}.$$

Using the techniques in Jing et al. (2009), we can prove the following Wilks' theorem,

Theorem 1. *Let μ_0 be the true value of μ. Under the regularity conditions that $Eh^2(Z) < \infty$ and $\sigma_h^2 > 0$,*

$$-2logR(\mu_0) \xrightarrow{d} \chi_1^2.$$

Based on the previous theorem, we can construct asymptotic $(1-\alpha)100\%$ JEL confidence

intervals of the form

$$I_c = \{\mu \mid -2logR(\mu) \leq \chi_1^2(\alpha)\},$$

where $\chi_1^2(\alpha)$ is the upper αth quantile of the χ_1^2 distribution.

2.2 Jackknife empirical likelihood with nuisance parameters

Let $\xi = (\mu, \theta^T)^T$ be the collection of unknown parameters involved in the estimation of μ, where θ denotes the vector of nuisance parameters. In particular, $\theta = (\gamma^T, \beta^T)^T$ for $\hat{\mu}_{USUAL}$ and $\theta = (\gamma^T, \beta^T, c^T)^T$ for $\hat{\mu}_{PROJ}$. Since we are only interested in μ, we can estimate θ through a subset of estimating equations. Estimation of both $\hat{\mu}_{USUAL}$ and $\hat{\mu}_{PROJ}$ involves solving jointly a set of M-estimating equations (Stefanski and Boos, 2002). $\hat{\mu}_{USUAL}$ is found by solving the score equation for γ, the least squares equation for β, and the estimating equation implied by eqn. (1.1). On the other hand, $\hat{\mu}_{PROJ}$ is found by again solving the score equation for γ, the two estimating equations implied by eqn. (1.2) for β and c, and the estimating equation implied by eqn. (1.1).

As in the previous section, the method is illustrated using the USUAL doubly robust estimation methodology. Results for the Cao et al. (2009) estimator are obtained by replacing $\tilde{\beta}$ with $\tilde{\beta}^*$, which is the corresponding estimator for β obtained by solving eqn. (1.2).

Let $\tilde{\theta}$ denote the corresponding estimator for θ and set

$$T_n(\mu) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi(x_i, \tilde{\gamma})} \left(Y_i - m(x_i, \tilde{\beta}) \right) + m(x_i, \tilde{\beta}) - \mu \right\}.$$

Next, let $\tilde{\theta}^{(-j)}$ denote the estimator for θ obtained by first removing the jth observation from the original data set. Similarly, define

$$T_n^{(-j)}(\mu) = (n-1)^{-1} \sum_{i=1, i \neq j}^{n} \left\{ \frac{R_i}{\pi(x_i, \tilde{\gamma}^{(-j)})} \left(Y_i - m(x_i, \tilde{\beta}^{(-j)}) \right) + m(x_i, \tilde{\beta}^{(-j)}) - \mu \right\}.$$

The jackknife pseudo sample is then defined as

$$V_i(\mu) = nT_n(\mu) - (n-1)T_n^{(-j)}(\mu), \qquad \text{for } i = 1, \ldots, n.$$

Since the jackknife pseudo sample is expected to be asymptotically independent (Tukey, 1958), we may apply the standard empirical likelihood method to these values in order to construct confidence intervals for μ. Hence, we define the JEL ratio for μ as

$$L^J(\mu) = sup\left\{\prod_{i=1}^{n}(np_i) : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i V_i(\mu) = 0, p_i > 0\right\}.$$

Using the Lagrange multiplier technique, the above ratio is maximized at

$$p_i = \frac{1}{n}\frac{1}{1+\lambda V_i(\mu)},$$

and the empirical log-likelihood ratio is

$$\ell^J(\mu) = -logL^J(\mu) = \sum_{i=1}^{n} log\left\{1 + \lambda V_i(\mu)\right\},$$

where λ satisfies

$$n^{-1}\sum_{i=1}^{n}\frac{V_i(\mu)}{1+\lambda V_i(\mu)} = 0.$$

We have the following theorem:

Theorem 2. *Let μ_0 be the true value of μ. Under regularity conditions (A1)-(A7) given in Li et al. (2011), we have that*

$$2\ell^J(\mu_0) \xrightarrow{d} \chi_1^2.$$

Based on Theorem 2, a $(1-\alpha)100\%$ confidence interval for μ is given by

$$I_\alpha = \left\{\mu : 2\ell^J(\mu) \leq \chi_1^2(\alpha)\right\},$$

where $\chi^2_1(\alpha)$ is the upper αth quantile of the χ^2_1 distribution.

2.3 Profile empirical likelihood

As in the introduction, let $Z_i = (R_iY_i, R_i, X_i)$ for $i = 1, \ldots, n$ be the data actually observed. Further, suppose that $Z_i \sim F_Z$ for some unknown distribution F_Z. As in the previous section, we are interested in obtaining information about μ in the presence of a nuisance parameter θ, where $\theta = (\gamma^T, \beta^T)^T$ for $\hat{\mu}_{USUAL}$ and $\theta = (\gamma^T, \beta^T, c^T)^T$ for $\hat{\mu}_{PROJ}$. Let $U(Z, \mu, \theta)$ denote the unbiased estimating equation which relates μ and F_Z. Similarly, let $V(Z, \theta)$ represent the vector of unbiased estimating functions relating to θ and F_Z.

In order to streamline the presentation of the results, let $\pi_i = \pi(X_i, \gamma)$ denote the propensity score model, and $m_i = m(X_i, \beta)$ represent the outcome regression model. Furthermore, write $\pi_{\gamma_i} = \dfrac{\partial}{\partial\gamma}\pi(X_i, \gamma)$ and $m_{\beta_i} = \dfrac{\partial}{\partial\beta}m(X_i, \beta)$. Finally, let S_{γ_i} denote the score function corresponding to the propensity score model.

Based on the USUAL doubly robust estimator, we have the following estimating equation for μ

$$U_{USUAL}(Z_i, \mu, \theta) = \frac{R_i}{\pi_i}(Y_i - m_i) + m_i - \mu.$$

Also, since propensity scores and outcome regression can be modeled with a logistic and a linear regression model respectively, we have the following estimating equations for θ

$$V_{USUAL}(Z_i, \theta) = \begin{pmatrix} R_i(Y_i - m_i)m_{\beta_i} \\ \dfrac{R_i - \pi_i}{\pi_i(1-\pi_i)}\pi_{\gamma_i} \end{pmatrix},$$

where the first equation corresponds to the least squares estimation method based on the complete cases, and the second equation corresponds to the maximum likelihood estimation method.

On the other hand, based on Cao et al. (2009), we have the following estimating equation for μ

$$U_{PROJ}(Z_i, \mu, \theta) = \frac{R_i}{\pi_i}(Y_i - m_i) + m_i - c^T S_{\gamma_i} - \mu,$$

where the additional term corresponds to the projection onto the propensity score tangent space (Tsiatis, 2007). Further, based on eqn. (1.2), it is easily seen that we have the following estimating equations for θ

$$V_{PROJ}(Z_i,\theta) = \begin{pmatrix} \frac{R_i(1-\pi_i)}{\pi_i^2} m_{\beta_i}\left(Y_i - m_i - c^T \frac{\pi_{\gamma_i}}{1-\pi_i}\right) \\ \frac{R_i(1-\pi_i)}{\pi_i^2}\frac{\pi_{\gamma_i}}{1-\pi_i}\left(Y_i - m_i - c^T \frac{\pi_{\gamma_i}}{1-\pi_i}\right) \\ \frac{R_i-\pi_i}{\pi_i(1-\pi_i)}\pi_{\gamma_i} \end{pmatrix},$$

where the first and second equations correspond to the estimation of β and c, respectively. The last equation yields an estimator for γ based on the score equation obtained by using logistic regression. The subsequent results are applicable to both $\hat{\mu}_{PROJ}$ and $\hat{\mu}_{USUAL}$ by replacing $U(Z,\mu,\theta)$ and $V(Z,\theta)$ with the corresponding pair of estimating equations for the estimator of interest.

We define the empirical likelihood for (μ,θ) as

$$L(\mu,\theta) = sup\left\{\prod_{i=1}^{n} p_i : p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i U(Z_i,\mu,\theta) = 0\right\}. \tag{2.1}$$

Once we obtain a consistent estimator $\hat{\theta}$ of θ, we can plug it into eqn. (2.1) to obtain a profile empirical likelihood for μ:

$$\hat{L}(\mu) = sup\left\{\prod_{i=1}^{n} p_i : p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i U(Z_i,\mu,\hat{\theta}) = 0\right\}.$$

Furthermore, the profile empirical likelihood ratio for μ is defined as:

$$\hat{R}(\mu) = \frac{\hat{L}(\mu)}{n^{-n}} = \prod_{i=1}^{n}\left\{1 + \lambda U(Z_i,\mu,\hat{\theta})\right\}^{-1},$$

where λ is the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{U(Z_i,\mu,\hat{\theta})}{1+\lambda U(Z_i,\mu,\hat{\theta})} = 0.$$

Then, the empirical log-likelihood ratio for μ is given by

$$\hat{\ell}(\mu) = -2\sum_{i=1}^{n} log\left\{1 + \lambda U(Z_i, \mu, \hat{\theta})\right\}. \tag{2.2}$$

By applying the general framework provided by Wang and Qin (2013) to the estimating equations defined above, we obtain the following results. Let

$$Q_{1n}(Z_i, \lambda, \mu, \theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{U(Z_i, \mu, \theta)}{1 + \lambda U(Z_i, \mu, \theta)}.$$

We have the following theorem:

Theorem 3. *Assume that* (μ_0, θ_0) *is the true value of* (μ, θ), $E\left[\frac{\partial U(Z, \mu_0, \theta_0)}{\partial \mu}\right]$ *and* $E\left[\frac{\partial V(Z, \theta_0)}{\partial \theta}\right]$ *are negatively definite. Then,*

$$\hat{\ell}(\mu_0) \xrightarrow{d} k\chi_1^2,$$

where k *is obtained as follows:*

$$k = (-S_{11})^{-1}\left(I, \quad S_{12}(-S_{22})^{-1}\right) S^* \begin{pmatrix} I \\ (-S_{22})^{-1}S_{12}^T \end{pmatrix},$$

$$S^* = Cov\left([U(Z, \mu_0, \theta_0), V^T(Z, \theta_0)]^T\right),$$

$$S_{11} = E\left[\frac{\partial Q_1(Z, 0, \mu_0, \theta_0)}{\partial \lambda}\right],$$

$$S_{12} = E\left[\frac{\partial Q_1(Z, 0, \mu_0, \theta_0)}{\partial \theta^T}\right],$$

$$S_{22} = E\left[\frac{\partial V(Z, \theta_0)}{\partial \theta^T}\right].$$

Making use of Theorem 3, we can construct $(1 - \alpha)100\%$ profile empirical likelihood

confidence intervals for μ of the form

$$I_\alpha = \{\mu : \hat{\ell}(\mu) \leq c_\alpha\},$$

where c_α is the $(1-\alpha)^{th}$ quantile of the $k\chi^2_1$ distribution.

2.4 Influence function based empirical likelihood

In this section, we propose an empirical likelihood method based on the influence function for the estimator of interest. The proposed method inherits the standard χ^2 limiting distribution. This property, in turn, significantly improves computation time compared to the jackknife and profile empirical likelihood methods described previously.

In the foregoing section, we defined the estimating equations for both $\hat{\mu}_{USUAL}$ and $\hat{\mu}_{PROJ}$. For development purposes, let us consider inferences for $\hat{\mu}_{PROJ}$.

Assuming that γ_0 is the true value of γ, the influence functions corresponding to estimators of the form given by eqn. (1.1) defined in this thesis have the form:

$$\frac{RY}{\pi(X,\gamma_0)} - \frac{R-\pi(X,\gamma_0)}{\pi(X,\gamma_0)}\left\{m(X,\beta) + c^T\frac{\pi_{\gamma_0}(X,\gamma_0)}{1-\pi(X,\gamma_0)}\right\} - \mu,$$

where $\pi_{\gamma_0}(X,\gamma_0) = \partial/\partial\gamma_0\{\pi(X,\gamma_0)\}$, see Cao et al. (2009).

Let $S_\gamma(R,X,\gamma) = \{R-\pi(X,\gamma)\}[\pi(X,\gamma)\{1-\pi(X,\gamma)\}]^{-1}\pi_\gamma(X,\gamma)$ be the score function for gamma, where $\pi_\gamma(X,\gamma) = \partial/\partial\gamma\{\pi(X,\gamma)\}$. Using the influence function defined above, we consider

$$W_{ni}(\mu) = \frac{R_i}{\hat{\pi}_i}(Y_i - \hat{m}_i{}^*) + \hat{m}_i{}^* - \hat{c}^T\hat{S}_{\gamma_i} - \mu,$$

where $\hat{\pi}_i = \pi(X_i,\hat{\gamma})$ denotes the estimated propensity score, $\hat{m}_i{}^* = m(X_i,\hat{\beta}^*)$ represents the predicted outcome, and $\hat{S}_{\gamma_i} = S_\gamma(R_i,X_i,\hat{\gamma})$ denotes the estimated value of the score functions associated with the propensity score model. Further, in order to obtain $\hat{\beta}^*$ and $\hat{c}$, we need to solve equation (1.2) jointly in (β,c). To estimate γ, we use the method of maximum likelihood.

Next, we apply Owen's empirical likelihood method using $W_{ni}(\mu)$, to define the EL ratio at μ as

$$R_{IF}(\mu) = sup\left\{\prod_{i=1}^{n} np_i : p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i W_{ni}(\mu) = 0\right\}.$$

Define the log-empirical likelihood ratio as

$$\ell_{IF}(\mu) = -2logR_{IF}(\mu).$$

Using the method of Lagrange multipliers, one has that

$$\ell_{IF}(\mu) = 2\sum_{i=1}^{n} log\left\{1 + \lambda(\mu)W_{ni}(\mu)\right\},$$

where λ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \frac{W_{ni}(\mu)}{1 + \lambda(\mu)W_{ni}(\mu)} = 0.$$

Similar to Owen (2001), we have the following Wilk's theorem:

Theorem 4. *Suppose that μ_0 is the true value of μ, then*

$$\ell_{IF}(\mu_0) \xrightarrow{d} \chi_1^2.$$

Based on Theorem 4, we may construct asymptotic $(1-\alpha)100\%$ empirical likelihood confidence intervals for μ as follows:

$$I_\alpha = \{\mu : \ell_{IF}(\mu) \leq \chi_1^2(\alpha)\},$$

where $\chi_1^2(\alpha)$ denotes the $(1-\alpha)^{th}$ quantile of the χ_1^2 distribution.

Inferences for μ_{USUAL} work in the exact same way by replacing $W_{ni}(\mu)$ above with

$$W_{ni}^*(\mu) = \frac{R_i}{\hat{\pi}_i}(Y_i - \hat{m}_i) + \hat{m}_i - \mu,$$

where, in this case, γ is estimated using maximum likelihood, and β is estimated using the method of least squares.

3 SIMULATION STUDY

Based on the methodology proposed in the previous chapters, an extensive simulation study is carried out to compute the coverage probabilities and average lengths of confidence intervals. We compare the results obtained from the EL methods to those of the normal approximations. Our simulations are identical to those conducted by Kang and Schafer (2007), whose design leads to the discovery that the USUAL doubly robust estimator may be severely biased when both models are incorrectly specified. For samples of size $n = 50$, $n = 100$, $n = 200$, $n = 500$, and $n = 1000$, we consider the four possible combinations of correct and erroneous model specifications. All simulation results are obtained using 1000 repetitions. For each estimator, nominal 95% confidence intervals for μ are calculated.

For each i $(i = 1, \ldots, n)$, let $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})^T$ be generated as a standard multivariate normal random variable. Also, let the elements of $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^T$ be defined as $X_{i1} = exp(Z_{i1}/2)$, $X_{i2} = Z_{i2}/\{1 + exp(Z_{i1})\} + 10$, $X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$, and $X_{i4} = (Z_{i2} + Z_{i4} + 20)^2$. Define $Y_i = m_0(Z_i) + \epsilon_i$ for $\epsilon_i \sim N(0, 1)$ and

$$m_0(Z_i) = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4}.$$

Further, let $R_i \sim Bernoulli(\pi_0(Z_i))$, such that

$$\pi_0(Z_i) = sigmoid(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4}),$$

where $sigmoid(x) = e^x/(1 + e^x)$. Correct models are obtained when a linear regression of Y_i on Z_i and a logistic regression of R_i on Z_i, respectively, are fitted. Thus, $m(Z, \beta)$ and $\pi(Z, \gamma)$ represent the correct models. Incorrect models are obtained by replacing Z_i with X_i; hence, $m(X, \beta)$ and $\pi(X, \gamma)$ denote the incorrect models. Furthermore, the true value of the mean is $\mu_0 = 210$.

Tables 3.1 - 3.4 display the results of the coverage probabilities and average interval lengths for each of the proposed methods. In general, the performance of the methods improves as the sample size increases.

Table (3.1) Coverage probabilities for $\hat{\mu}_{PROJ}$

METHOD	OR	PS	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
NA	Correct	Correct	0.941	0.946	0.945	0.950	0.953
	Correct	Incorrect	0.938	0.947	0.943	0.952	0.952
	Incorrect	Correct	0.906	0.931	0.940	0.953	0.954
	Incorrect	Incorrect	0.619	0.819	0.903	0.876	0.815
JEL	Correct	Correct	0.946	0.950	0.949	0.951	0.954
	Correct	Incorrect	0.943	0.948	0.944	0.952	0.955
	Incorrect	Correct	0.957	0.972	0.979	0.988	0.995
	Incorrect	Incorrect	0.956	0.982	0.981	0.966	0.925
JELN	Correct	Correct	0.919	0.948	0.948	0.951	0.954
	Correct	Incorrect	0.936	0.947	0.947	0.950	0.953
	Incorrect	Correct	0.899	0.939	0.947	0.949	0.955
	Incorrect	Incorrect	0.512	0.878	0.905	0.858	0.664
PEL	Correct	Correct	0.943	0.949	0.948	0.950	0.955
	Correct	Incorrect	0.947	0.950	0.946	0.950	0.954
	Incorrect	Correct	0.917	0.935	0.942	0.952	0.947
	Incorrect	Incorrect	*	*	*	*	*
IFEL	Correct	Correct	0.937	0.947	0.948	0.950	0.954
	Correct	Incorrect	0.942	0.947	0.946	0.952	0.954
	Incorrect	Correct	0.910	0.939	0.940	0.953	0.955
	Incorrect	Incorrect	0.725	0.860	0.923	0.896	0.839

NOTE:

NA: Normal Approximation

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

IFEL: Influence Function Empirical Likelihood

OR: Outcome Regression Model

PS: Propensity Score Model

Table (3.2) Coverage probabilities for $\hat{\mu}_{USUAL}$

METHOD	OR	PS	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
NA	Correct	Correct	0.941	0.947	0.943	0.951	0.954
	Correct	Incorrect	0.940	0.946	0.945	0.947	0.953
	Incorrect	Correct	0.930	0.939	0.942	0.940	0.948
	Incorrect	Incorrect	0.931	0.936	0.915	0.813	0.619
JEL	Correct	Correct	0.945	0.948	0.946	0.951	0.954
	Correct	Incorrect	0.945	0.948	0.946	0.948	0.947
	Incorrect	Correct	0.932	0.943	0.948	0.959	0.956
	Incorrect	Incorrect	0.912	0.884	0.828	0.628	0.355
JELN	Correct	Correct	0.944	0.948	0.946	0.951	0.954
	Correct	Incorrect	0.945	0.949	0.947	0.947	0.955
	Incorrect	Correct	0.949	0.959	0.952	0.958	0.960
	Incorrect	Incorrect	0.944	0.961	0.930	0.858	0.664
PEL	Correct	Correct	0.948	0.949	0.946	0.951	0.954
	Correct	Incorrect	0.946	0.948	0.947	0.947	0.949
	Incorrect	Correct	0.929	0.934	0.933	0.933	0.941
	Incorrect	Incorrect	0.939	0.921	0.838	0.574	0.329
IFEL	Correct	Correct	0.945	0.948	0.946	0.951	0.954
	Correct	Incorrect	0.945	0.948	0.946	0.948	0.947
	Incorrect	Correct	0.932	0.943	0.948	0.959	0.956
	Incorrect	Incorrect	0.912	0.884	0.828	0.628	0.355

NOTE:

NA: Normal Approximation

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

IFEL: Influence Function Empirical Likelihood

OR: Outcome Regression Model

PS: Propensity Score Model

Table (3.3) Average length of confidence intervals for $\hat{\mu}_{PROJ}$

METHOD	OR	PS	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
NA	Correct	Correct	20.817	14.180	10.030	6.354	4.493
	Correct	Incorrect	19.816	14.137	10.023	6.353	4.490
	Incorrect	Correct	19.858	14.239	10.110	6.407	4.532
	Incorrect	Incorrect	22.037	15.586	10.887	6.726	4.676
JEL	Correct	Correct	21.106	14.529	10.166	6.395	4.511
	Correct	Incorrect	20.304	14.303	10.088	6.375	4.507
	Incorrect	Correct	25.590	17.768	12.539	8.011	5.707
	Incorrect	Incorrect	69.653	37.601	19.557	9.807	6.169
JELN	Correct	Correct	22.792	14.575	10.043	6.357	4.515
	Correct	Incorrect	20.721	14.277	10.027	6.358	4.520
	Incorrect	Correct	27.237	16.208	10.517	6.487	4.563
	Incorrect	Incorrect	97.419	39.978	18.321	7.916	5.541
PEL	Correct	Correct	21.312	14.636	10.196	6.403	4.513
	Correct	Incorrect	20.652	14.388	10.115	6.387	4.512
	Incorrect	Correct	20.581	14.563	10.231	6.451	4.552
	Incorrect	Incorrect	*	*	*	*	*
IFEL	Correct	Correct	20.144	14.256	10.068	6.364	4.497
	Correct	Incorrect	20.123	14.256	10.066	6.364	4.498
	Incorrect	Correct	20.358	14.393	10.158	6.419	4.535
	Incorrect	Incorrect	27.732	17.632	11.507	6.913	4.843

NOTE:

NA: Normal Approximation

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

IFEL: Influence Function Empirical Likelihood

OR: Outcome Regression Model

PS: Propensity Score Model

Table (3.4) Average length of confidence intervals for $\hat{\mu}_{USUAL}$

METHOD	OR	PS	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
NA	Correct	Correct	19.824	14.146	10.025	6.353	4.493
	Correct	Incorrect	19.853	14.184	10.073	6.425	4.882
	Incorrect	Correct	25.531	18.048	12.576	8.258	5.992
	Incorrect	Incorrect	30.578	30.740	30.367	28.708	128.507
JEL	Correct	Correct	20.125	14.268	10.071	6.365	4.498
	Correct	Incorrect	20.127	14.283	10.117	6.470	4.726
	Incorrect	Correct	26.596	19.912	14.358	9.432	6.800
	Incorrect	Incorrect	26.751	22.564	18.329	16.551	15.977
JELN	Correct	Correct	20.049	14.211	9.979	6.354	4.509
	Correct	Incorrect	20.165	14.229	10.070	6.415	5.551
	Incorrect	Correct	24.524	17.925	10.714	7.968	5.877
	Incorrect	Incorrect	27.670	23.336	18.379	14.311	11.662
PEL	Correct	Correct	20.339	14.342	10.096	6.372	4.500
	Correct	Incorrect	20.371	14.387	10.161	6.462	4.736
	Incorrect	Correct	26.389	18.630	13.005	8.515	6.178
	Incorrect	Incorrect	30.043	27.908	28.521	15.969	13.749
IFEL	Correct	Correct	20.125	14.268	10.071	6.365	4.498
	Correct	Incorrect	20.127	14.283	10.117	6.470	4.726
	Incorrect	Correct	26.596	19.912	14.358	9.432	6.800
	Incorrect	Incorrect	26.751	22.564	18.329	16.551	15.977

NOTE:

NA: Normal Approximation

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

IFEL: Influence Function Empirical Likelihood

OR: Outcome Regression Model

PS: Propensity Score Model

For $\hat{\mu}_{USUAL}$, all of the proposed methods perform similarly as long as one of the models is correctly specified. Regarding coverage probability, all methods slightly undercover when $n = 50$. As the sample size is increased, the coverage probability for all the methods converges to nominal level. In the case when both models are incorrect, the JELN method performs the best compared to other methods; however, all of the proposed techniques undercover, and the performance decreases as the sample size increases. This trend is expected since according to Scharfstein et al. (1999), $\hat{\mu}_{USUAL}$ may be severely biased.

In terms of average interval lengths, when at least one model is correctly specified, the NA method has the shortest average lengths followed by JELN in general. When both models are incorrect, the NA method has the largest lengths. In this scenario, the JELN method has the best performance.

In the case of $\hat{\mu}_{PROJ}$, all coverage probabilities for the proposed methods, except the JEL method, converge to nominal level as n increases. Similar to the results of Cao et al. (2009), when both models are incorrect and the sample size is large, our simulations show that the coverage probabilities for the normal approximations are vastly improved upon by using the $\hat{\mu}_{PROJ}$ estimator as opposed to $\hat{\mu}_{USUAL}$. In terms of coverage probabilities, the IFEL method performs the best overall.

Regarding average lengths of confidence intervals, it is seen that the JEL method has the longest lengths compared to the other methods. This is the main reason for the observed inflation with respect to coverage probabilities for this method. The JELN method has the second longest average lengths, but we observe close to nomial level in most settings. Similar to the case of using $\hat{\mu}_{USUAL}$, the NA method has the shortest interval lengths closely followed by the IFEL method.

Our simulation results show that the weights obtained for the PEL method are either negative or very large when both models are incorrect and the $\hat{\mu}_{PROJ}$ methodology is used. In this case, it appears that a consistent estimator of the Λ matrix does not exist which in turn causes the weight for the limiting distribution to be incorrect. For this reason, care must be taken when applying the PEL method under this scenario. Moreover, as tables

3.2 and 3.4 illustrate, the JEL and IFEL methods have the same coverage probabilities and average lengths when $\hat{\mu}_{USUAL}$ is used.

4 REAL DATA ANALYSIS

In this chapter, the proposed methods are applied to two real data sets. For each data set, both the $\hat{\mu}_{USUAL}$ and the $\hat{\mu}_{PROJ}$ estimators are obtained, and 95% confidence intervals are constructed using each of the proposed methods. As in the simulation studies, the coverage probabilities and average interval lengths are the same for the JEL and IFEL methods when the $\hat{\mu}_{USUAL}$ estimator is used.

The first data set, "Hitters", was taken from the StatLib library at the Carnegie Mellon University. The data set consists of 322 observations describing career statistics for baseball players. The second data set, "Acupuncture", consists of 401 observations from a study comparing the effects of acupuncture on the treatment of chronic headache. This data set was obtained from the article titled "Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial" published in March 2004 by the *BMJ* journal.

4.1 Hitters data analysis

The "Hitters" data set consists of 322 observations with 18 variables representing various performance statistics for major league baseball players during 1986, and a factor variable representing the player's league in the beginning of 1987. Our interest is in estimating the 1987 average annual salary on opening day in thousands of dollars for all baseball players. Salary information is missing for 59 observations.

Applying ordinary least squares regression with Bayes information criterion (BIC) as the model selection technique, we find that the significant variables to be included in the outcome regression model include the number of times at bat, hits, walks, and put outs in 1986. Also, the number of hits during his professional career, and the player's division at the end of 1986 were selected into the model.

To model the missingness mechanism observed in this data set, a logistic regression model was fitted using the BIC criterion for variable selection. Only two variables were selected into the final model: the number of runs and the number of assists made in 1986.

The results of applying the proposed methods are given in Table 4.1. The first thing to

note is that the PFL method yields incorrect confidence intervals using Cao's doubly robust estimation technique. The weight obtained is 183.844, which explains the extraordinary length of the confidence interval. Both estimators are close to each other with a percent difference of less than 0.3%. In terms of length, the results are consistent with the simulations; we see that normal approximations perform the best, but are closely followed by the IFEL technique.

Table (4.1) 95% C.I. for the Hitters Data Set

$\hat{\mu}_{USUAL} = 515.725$			
	LB	UB	Length
NA	463.663	567.787	104.124
IFEL	467.237	571.632	104.395
JEL	467.237	571.632	104.395
JELN	465.454	570.761	105.307
PEL	466.798	572.220	105.422
$\hat{\mu}_{PROJ} = 517.063$			
	LB	UB	Length
NA	466.057	568.069	102.012
IFEL	468.987	572.102	103.115
JEL	467.601	573.605	106.004
JELN	467.121	573.951	106.830
PEL	47.834	2327.771	2279.937

NOTE:

NA: Normal Approximation

IFEL: Influence Function Empirical Likelihood

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

LB: Lower Bound of 95% Confidence Interval

UB: Upper Bound of 95% Confidence Interval

Length: Length of Confidence Interval

4.2 Acupuncture data analysis

The "Acupuncture" data set contains information on 401 patients with chronic headache involved in a large, pragmatic, and randomized clinical trial. Subjects in the study were randomly allocated to receive acupuncture treatments over 3 months, or to a control intervention group receiving usual care. 196 subjects were allocated to control (56 dropouts), while 205 to treatment (44 dropouts). The main goal of this study was to assess the effects of receiving acupuncture treatments on headache scores versus usual interventions.

The data consists of 18 baseline covariates including demographics and results of the SF-36 questionnaire; a covariate which determines the treatment group membership, and the outcome of interest which is defined as a headache score at the one year follow up. A separate analysis is carried out for each group. The model selection process follows the same steps as described in the previous section for the analysis of the "Hitters" data set.

For the control group, headache scores are best modeled with an outcome regression model using the baseline headache score and the "rle" value of the SF-36 questionnaire. The final propensity scores model uses age and the "pf" value of the SF-36 questionnaire.

The results for this group are given in Table 4.2. The percent difference between the two estimators is only 0.1%. This time, the PEL method has very short lengths for $\hat{\mu}_{PROJ}$ since the estimate for the weight in Theorem 3 is $\hat{k} = 0.599$. Overall, the NA and IFEL methods have the best performance.

For the treatment group, the outcome regression model is chosen to include the baseline headache score as well as the "painmeds" variable, which appears to be a score for pain medications at baseline. The best model which can be used to describe the missingness mechanism observed in the data is the null model according to Bayes information criterion.

Results for the treatment group are given in Table 4.3. In this scenario, the IFEL and the JEL have the best performance. Inferences using the PEL seem to be correct since the weight is $\hat{k} = 1.141$. As seen in previous tables and the simulation results, the JELN method estimate appears to have the lowest efficiency in terms of lengths of confidence intervals.

Table (4.2) 95% C.I. for the Control Group

$\hat{\mu}_{USUAL} = 22.873$			
	LB	UB	Length
NA	20.339	25.407	5.068
IFEL	20.499	25.582	5.083
JEL	20.499	25.582	5.083
JELN	20.473	25.618	5.145
PEL	20.488	25.597	5.109
$\hat{\mu}_{PROJ} = 22.907$			
	LB	UB	Length
NA	20.335	25.479	5.144
IFEL	20.513	25.652	5.139
JEL	19.693	26.295	6.602
JELN	20.541	25.941	5.400
PEL	21.029	24.993	3.964

NOTE:

NA: Normal Approximation

IFEL: Influence Function Empirical Likelihood

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

LB: Lower Bound of 95% Confidence Interval

UB: Upper Bound of 95% Confidence Interval

Length: Length of Confidence Interval

Table (4.3) 95% C.I. for the Treatment Group

$\hat{\mu}_{USUAL} = 16.762$			
	LB	UB	Length
NA	14.497	19.027	4.530
IFEL	14.811	19.078	4.267
JEL	14.811	19.078	4.267
JELN	14.672	19.280	4.608
PEL	14.675	19.271	4.596
$\hat{\mu}_{PROJ} = 16.750$			
	LB	UB	Length
NA	14.528	18.972	4.444
IFEL	14.811	19.024	4.213
JEL	14.817	19.024	4.207
JELN	14.697	19.242	4.545
PEL	14.688	19.195	4.507

NOTE:

NA: Normal Approximation

IFEL: Influence Function Empirical Likelihood

JEL: Jackknife Empirical Likelihood

JELN: Jackknife Empirical Likelihood with Nuisance Parameters

PEL: Profile Empirical Likelihood

LB: Lower Bound of 95% Confidence Interval

UB: Upper Bound of 95% Confidence Interval

Length: Length of Confidence Interval

Since none of the two confidence intervals overlap with each other, we can reject the null hypothesis that there is no headache score difference between the 2 groups at the $\alpha = 0.05$ confidence level. Our results show that acupuncture leads to clinically relevant benefits for patients suffering from chronic headache. Furthermore, using our proposed methods, results are coherent to those published in Vickers et al. (2004).

5 SUMMARY AND FUTURE WORK

5.1 Summary

In this thesis, we developed four distinct methods to construct confidence intervals for a population mean using incomplete outcome data based on the empirical likelihood methodology. The confidence intervals are constructed using the estimating equations of both the $\hat{\mu}_{USUAL}$ and $\hat{\mu}_{PROJ}$ doubly robust estimators.

Our simulation results suggest that the confidence intervals for $\hat{\mu}_{PROJ}$ perform better than those for $\hat{\mu}_{USUAL}$ as the sample size increases. In terms of coverage probability, we note that most scenarios undercover when both models are incorrectly specified. When at least one model is correct, coverage probabilities of all methods converge to the nominal level with the exception of the JEL method applied to $\hat{\mu}_{PROJ}$, which actually overcovers. With respect to interval lengths, we note that all the methods have longer lengths when the outcome regression model is incorrectly specified. In this aspect, normal approximations have the best performance, but are closely followed by IFEL.

Results for the real data analysis also suggest that IFEL and NA have similar performance. For the Hitters data set, the normal approximations have the shortest lengths; however, for the Acupuncture data set, the IFEL method has the shortest lengths of all 5 methods.

Overall, it appears that using $\hat{\mu}_{PROJ}$ and constructing confidence intervals based on the IFEL method yields the best performance. Furthermore, IFEL is very computationally efficient, and avoids the complex formulation of the sandwich standard errors associated with the NA method. Based on our results, we suggest the use of this strategy to obtain inferences for a population mean when facing real data problems with missing outcome data.

5.2 Future Work

The main drawback of empirical likelihood lies in the difficulty of its practical implementation due to the complex optimizations involved. For this reason, the coordinate descent

algorithm of Tang and Wu (2014) as well as the self-concordance for EL proposed by Owen (2013) can lead to better computational efficiency.

In addition, the enhanced propensity score model proposed in Cao et al. (2009) was shown to reduce the bias of $\hat{\mu}_{USUAL}$ and $\hat{\mu}_{PROJ}$. We expect that combining $\hat{\mu}_{PROJ}$ with the enhanced propensity scores and the IFEL method may give the most robust inferences for the population mean when missing outcome data is present.

Bibliography

Bouadoumou, M., Zhao, Y., and Lu, Y. (2015). Jackknife empirical likelihood for the accelerated failure time model with censored data. *Communications in Statistics - Simulation and Computation*, 44(7):1818–1833.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Chan, K. C. G. (2012). Uniform improvement of empirical likelihood for missing response problem. *Electron. J. Statist.*, 6:289–302.

Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):1166–1181.

Chen, S. X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test*, 18(3):415–447.

Gong, Y., Peng, L., and Qi, Y. (2010). Smoothed jackknife empirical likelihood method for ROC curve. *Journal of Multivariate Analysis*, 101(6):1520 – 1531.

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics*, 37(3):370–384.

Jing, B.-Y., Yuan, J., and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association*, 104(487):1224–1232.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22(4):523–539.

Li, M., Peng, L., and Qi, Y. (2011). Reduce computation in profile empirical likelihood method. *Canadian Journal of Statistics*, 39(2):370–384.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

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

Owen, A. B. (2013). Self-concordance for empirical likelihood. *Canadian Journal of Statistics*, 41(3):387–397.

Peng, L. (2012). Approximate jackknife empirical likelihood method for estimating equations. *Canadian Journal of Statistics*, 40(1):110–123.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):101–122.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3-4):353–360.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Shi, X. (1984). The approximate independence of jackknife pseudo-values and the bootstrap methods. *Journal of Wuhan Institute Hydra-Electric Engineering*, 2:83–90.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Tang, C. Y. and Wu, T. T. (2014). Nested coordinate descent algorithms for empirical likelihood. *Journal of Statistical Computation and Simulation*, 84(9):1917–1930.

Tang, N.-S. and Zhao, P.-Y. (2013). Empirical likelihood semiparametric nonlinear regression analysis for longitudinal data with responses missing at random. *Annals of the Institute of Statistical Mathematics*, 65(4):639 – 665.

Tang, N.-S., Zhao, P.-Y., and Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica*, 24(2):723–747.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer.

Tsiatis, A. A. and Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):569–573.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *The Annals of Statistics*, 29(2):614.

Vickers, A. J., Rees, R. W., Zollman, C. E., McCarney, R., Smith, C. M., Ellis, N., Fisher, P., and Van Haselen, R. (2004). Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *BMJ*, 328(7442):744–747.

Wang, B. and Qin, G. (2013). Empirical likelihood confidence regions for the evaluation of continuous-scale diagnostic tests in the presence of verification bias. *Canadian Journal of Statistics*, 41(3):398–420.

Wang, R., Peng, L., and Qi, Y. (2013). Jackknife empirical likelihood test for the equality of two high dimensional means. *Statistica Sinica*, 23(2):667–690.

Xue, L. and Xue, D. (2011). Empirical likelihood for semiparametric regression model with missing response data. *Journal of Multivariate Analysis*, 102(4):723 – 740.

Yang, H. and Zhao, Y. (2013). Jackknife empirical likelihood confidence intervals for the difference of two ROC curves. *Journal of Multivariate Analysis*, 115:270 – 284.

Zhao, H., Zhao, P.-Y., and Tang, N.-S. (2013). Empirical likelihood inference for mean functionals with nonignorably missing response data. *Computational Statistics & Data Analysis*, 66(0):101 – 116.

Zhao, Y. (2011). Empirical likelihood inference for the accelerated failure time model. *Statistics & Probability Letters*, 81(5):603–610.

Zhong, P.-S. and Chen, S. (2014). Jackknife empirical likelihood inference with regression imputation and survey data. *Journal of Multivariate Analysis*, 129:193 – 205.