12-18-2013

Jackknife Empirical Likelihood-Based Confidence Intervals for Low Income Proportions with Missing Data

YANAN YIN
GEORGIA STATE UNIVERSITY

JACKKNIFE EMPIRICAL LIKELIHOOD-BASED CONFIDENCE INTERVALS FOR LOW INCOME PROPORTIONS WITH MISSING DATA

by

YANAN YIN

Under the Direction of Gengsheng Qin

ABSTRACT

The estimation of low income proportions plays an important role in comparisons of poverty in different countries. In most countries, the stability of the society and the development of economics depend on the estimation of low income proportions. An accurate estimation of a low income proportion has a crucial role for the development of the natural economy and the improvement of people's living standards. In this thesis, the Jackknife empirical likelihood method is employed to construct confidence intervals for a low income proportion when the observed data had missing values. Comprehensive simulation studies are conducted to compare the relative performances of two Jackknife empirical likelihood-based confidence intervals for low income proportions in terms of coverage probability. A real data example is used to illustrate the application of the proposed methods.

INDEX WORDS: Jackknife, Empirical Likelihood, Missing data, Auxiliary information, Coverage Probability, Low Income Proportion, Income distribution

JACKKNIFE EMPIRICAL LIKELIHOOD-BASED INFERENCES FOR A LOW INCOME PROPORTION WITH MISSING DATA

by

YANAN YIN

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

Master of Sciences

in the College of Arts and Sciences

Georgia State University

2013

JACKKNIFE EMPIRICAL LIKELIHOOD-BASED INFERENCES FOR A LOW INCOME PROPORTION WITH MISSING DATA

by

YANAN YIN

Committee Chair: Gengsheng Qin

Committee: Xin Qi

Ruiyan Luo

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2013

DEDICATION

This thesis is dedicated to Dr. Gengsheng Qin, my parents, Chenxue Li, and all my best friends.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of all the people supports me in different ways. I would like to gratefully to express my sincere appreciation to each and every of them. First and foremost, I would like to express my gratitude to my advisor Dr. Gengsheng Qin, for all his supports, guidance and patience. He's shearing his wealth of knowledge in statistics with me, and giving me a great encouragement to challenge myself. I will never forget his insightful knowledge, his patience and encouragement, all of them are inspiring and highly appreciated. Second of all, I would like to thank my thesis committee, Dr. Ruiyan Luo and Dr. Xin Qi for taking some valuable time out of their schedule to read my thesis and providing me valuable suggestions. Third of all, I would like to appreciate all the professors in the Department of Mathematics and Statistics at Georgia State University who taught me during my graduate study. They all help me a lot to develop my statistical knowledge and skills. Besides, I would also like to give a special thank to my best friend Chenxue Li, for taking her precious time to help me with my thesis, giving me great ideas and encouragement. Moreover, I want to thank my parents, unconditionally pay and encourage to me, accompany me to complete my research paper. In addition, I need to thank my friends and classmates Zhujun Li, Bing Liu, Siyu Tian, Jun Xia, Wen Zhou, Hao Fan for their useful help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

- EU - Europe
- EL - Empirical Likelihood
- JEL - Jackknife Empirical Likelihood
- HT - Horvitz and Thompson
- HJ - H*á*jek
- PSID - Panel Study of Income Dynamics

PART 1

INTRODUCTION

1.1 Low Income Proportion

Low income proportion plays a very important role in social and economic studies. In many countries economic policies depend heavily on the low income proportion estimation. The low income proportion is defined as the proportion of the population's incomes falling below a given fraction α $(0 < \alpha < 1)$ of the β-th $(0 < \beta < 1)$ quantile of the income distribution. It is a significant index in comparisons of poverty in different countries, which directly reflects the stability of the society. For example, the proportion below half $(\alpha = 0.5)$ of the national median income $(\beta = 0.5)$ was used as a basis for comparison of poverty in seven countries(Smeeding et al.,1990)[1]. For another example, low-wage earners in EU countries (Eurostat, 2010)[2] are defined as those employees earning two thirds $(\alpha = 2/3)$ or less of the national median hourly earnings. Governments pay attention to low income proportions, because a high low income proportion can lead to social instability and other social problems.

Let the population income Y be a non-negative random variable with distribution function $F(y)$. Assume that $F(y)$ has the density function $f(y)$. Denote $\xi_\beta = F^{-1}(\beta)$ as the β-th $(0 < \beta < 1)$ quantile of $F(y)$. The fraction α of the β-th quantile of the income distribution, $\alpha\xi_\beta$, is defined as the low income line. Then the low income proportion is

$$\Omega_{\alpha\beta} = P(Y \leq \alpha\xi_\beta) = F(\alpha\xi_\beta) \qquad (\star)$$

Usually, the income distribution $F(y)$ is rarely known. In order to estimate the low income proportion $\theta_{\alpha\beta}$, we have to estimate the low income line $\alpha\xi_\beta$ and the unknown income distribution $F(y)$.

There are many ways to measure how many people are poor in a given country, and there are many methods to make inferences for a low income proportion. One of the measurement criteria is estimating the low income proportion to compare poverty in different countries. In some literature, the relative poverty line is used instead of the low income proportion, and has been become increasingly popular in poverty studies. Zheng (2001)[3] derived a statistical inference procedure for poverty measure with relative poverty line, but this inference has not been well developed for the low income proportion. Yves and Chris (2003)[4] proposed variance estimation for a low income proportion based on the Family Expenditure Survey. Shao and Rao (1993)[5] proposed a linearization method to get variance estimation for a low income proportion based on the case where $\alpha = 0.5$ and $\beta = 0.5$. Preston(1995)[6] discussed the reliability of estimating a low income proportion based on simple random sample. These methods have poor finite sample performances when income data are skewed or have outliers because most of the above existing methods and deduced inferences are based on the asymptotic normal distribution. Because of the limitations of the data type in economics study, most of the income data are skewed data, which inspired us to propose better inferences for low income proportions by developing new statistical methods.

1.2 Missing data

Missing data occur when the data value is missing for a variable of interest in an observation. In statistics, missing data are a very common practical problem, which arises often in areas of study such as economics, social science and political sciences studies. Missing data exist in different patterns such as covariant partially missing, response partially missing, and both response variables and covariant variables partially missing. The excellent textbook by Little and Rubin(2002)[7] provided a comprehensive overview on missing data problems. In this thesis, the response variable is an income variable Y, which is assumed as missing at random. The covariant variables can be represented by different types such as different countries, the level of the education, and the age of people in a population. We will deal with problems on income data with missing values.

With missing data, people normally choose to delete all observations with missing values from the whole data set and use complete data to make inferences. However, the direct application of complete data inference procedures to missing data problems may produce biased estimation and lose efficiency. Therefore, we need to find effective ways to manage missing data in order to reduce the deviation and improve performances of the statistical analysis. Based on previous literature, Little and Rubin(2002)[7] provided likelihood-based approaches to the analysis of missing data. Rubin(1976)[8] focused primarily on multiple imputation approach to estimation of incomplete data regression models. Moreover, Ibrahim et al.(2005)[9] examined missing at random data by Bayesian approach. In this thesis, we will use the famous weighting methods motivated by Horvitz and Thompson's estimator[10] and Hájek estimator[11] to develop a new method for inferences on low income proportions with missing data.

1.3 Jackknife Empirical Likelihood

Parametric likelihood method is powerful for inferences of unknown parameter with correct parameter models. But sometimes this method can lead to biased inferences if the underlying parametric model is misspecified. For instance, if we use a normal model to analyze a data set which is a random sample from a Cauchy distribution, incorrect claims will appear. To avoid and reduce the risk of model mis-specification, we prefer to use non-parametric methods. Empirical Likelihood(EL), introduced by Owen (1988)[12], is a nonparametric method traditionally used for providing confidence intervals for the mean, without assuming a specified distribution for the underlying population. According to Wilks' theorem, EL ratio tends to the Chi-square distribution (Owen 1988)[12]. Many inferences and widely used applications based on the EL method are found in many occasions, such as public health studies, econometrics and sampling. During recent years, many cases and examples have proven that the EL approach is effective in many applications. For instance, Chen & Qin (1993) [13] applied EL method on finite populations with the auxiliary information. Besides, Yang et at.(2010) [14] proposed various EL-based inferences for the low income

proportion. Also, in Zhou, Qin, Lin & Li (2006)'s[15] paper, they demonstrated that EL-based method performed extraordinarily well on analyzing highly skewed health care cost data.

The empirical likelihood method is a procedure to maximize nonparametric likelihood function under constraints on the parameters. The maximization process goes smoothly if those constraints are linear. However, if the constraints involve nonlinear statistics, it runs into serious computational difficulties. In order to improve the computational efficiency, a new simple nonparametric method, called Jackknife Empirical Likelihood (JEL) was proposed by Jing et al.(2009) [16]. The JEL method is potentially useful for nonlinear statistics. As well known, "Jackknife, as a kind of re-sampling method, is applied with empirical likelihood and named as jackknife empirical likelihood, which surprisingly transforms nonlinear estimation equation as linear's and multi-variable optimization problem as simple-variables" Jing et al.(2009) [16]. Using the JEL methods, it's can reduce the computational intensity and maintain the computational accuracy. The most significant idea of the JEL is to turn the statistics of interest into a sample mean based on the jackknife pseudo-values (Quenouille, 1956)[17]. In this thesis, we will develop JEL-based methods for the inference of the low income proportion with missing data.

This thesis is organized as follows: In Section 2, we introduce the main methodology of the jackknife empirical likelihood for the low income proportion with the missing data. In Section 3, simulation studies are conducted to evaluate the coverage probabilities of the JEL-based confidence intervals for the low income proportion with missing data. Some data analysis with a real data example is given in Section 4. In Section 5, we make conclusion and discussion.

PART 2

METHODOLOGY

2.1 The low income proportion with missing data

We assume the income variable in this thesis is Y, the covariant variable is denoted as X and the missing indicator variable is D. Y is missing when $D = 0$. In this thesis we assume X always is observable, and the probability of observing y is as follows:

$$P(D = 1|x) = \omega(x, \eta) \tag{2.1}$$

where ω is a specified probability distribution function. We normally choose ω as the logistic regression, e.g., $exp(\eta_0+\eta_1 x)/[1+exp(\eta_0+\eta_1 x)]$, where $\eta = (\eta_0, \eta_1)$ is an unknown parameter vector.

Let $(y_i, x_i, d_i), i = 1, 2, ..., n$, be the observed data for (Y, X, D). Under assumption (2.1), the unknown parameter η can be estimated by maximizing the binomial likelihood:

$$L = \prod_{i=1}^{n} [\omega(x_i, \eta)]^{d_i} \{1 - \omega(x_i, \eta)\}^{1-d_i} \tag{2.2}$$

We denote the estimator as $\hat{\eta} = (\hat{\eta_0}, \hat{\eta_1})$.

In the income data analysis, the income distribution $F(y)$ of y is hardly known. Therefore, both the low income line $\alpha\xi_\beta$ and low income proportion $\Omega_{\alpha\beta}$ are unknown. The main purpose of this thesis is to construct the confidence intervals for low income proportion with missing data. The low income proportion defined as:

$$\Omega_{\alpha\beta} = P(Y \leq \alpha\xi_\beta) = F(\alpha\xi_\beta), \tag{2.3}$$

2.2 The estimation of income distribution and low income proportions with missing data

In this section, we consider two estimators for the income distribution and low income proportions.

2.2.1 The Horvitz and Thompson's estimator

The Horvitz and Thompson (1952)(HT)[10] proposed a method for estimating the population total and mean. They introduced an unbiased estimator for the population total which works for any design under a finite population setting (Fikri and Yaprak, 2012)[18]. Normally, we consider a finite population $U = 1, 2, ..., N$ and let π_i be the probability that the $i - th$ unit of the population is included in the sample. We measure a response z_i on each unit i, and wish to estimate:

$$\tau_Z = \sum_{i=1}^{N} z_i,$$

The Horvitz-Thompson estimator is defined as:

$$\hat{\tau}_{HT} = \sum_{i=1}^{v} \frac{z_i}{\pi_i}$$

where the sum is taken over the v distinct units in the sample.

In this thesis, motivated by the HT estimator, we can estimate the income distribution as follows:

$$\hat{F}_{HT}(y) = n^{-1} \sum_{i=1}^{n} \frac{D_i I(y_i \leq y)}{\omega(x_i, \hat{\eta})}, \tag{2.4}$$

where $\omega(x_i, \hat{\eta})$ estimate of $\omega(x_i, \eta)$ is a probability distribution function. And the low income proportion based on the Horvitz-Thompson estimator is:

$$\hat{\Omega}_{HT} = n^{-1} \sum_{i=1}^{n} \frac{D_i I(y_i \leq \alpha \hat{\xi}_{HT})}{\omega(x_i, \hat{\eta})}, \tag{2.5}$$

where $\hat{\xi}_{HT} = \hat{F}_{HT}^{-1}(\beta)$ is the β-th quantile of the $\hat{F}_{HT}(y)$.

2.2.2 The Hájek estimator

Another well known and popular estimator attributed to Hájek (1971)[19] is defined by:

$$\hat{\tau}_{HJ} = \frac{\sum_{i=1}^{v} \frac{z_i}{\pi_i}}{\sum_{i=1}^{v} \frac{1}{\pi_i}},$$

where π_i is the inclusion probability, z_i is the response on unit i, and v denotes as distinct units in the sample.

Based on the Hájek estimator, we could estimate the low income distribution as:

$$\hat{F}_{HJ}(y) = \frac{\sum_{i=1}^{n} D_i \frac{I(y_i \leq y)}{\omega(x_i, \hat{\eta})}}{\sum_{i=1}^{n} D_i \frac{1}{\omega(x_i, \hat{\eta})}}. \tag{2.6}$$

And the low income proportion based on this estimator is:

$$\hat{\Omega}_{HJ} = \frac{\sum_{i=1}^{n} \frac{D_i I(y \leq \alpha \hat{\xi}_{HJ})}{\omega(x_i, \hat{\eta})}}{\sum_{i=1}^{n} \frac{D_i}{\omega(x_i, \hat{\eta})}}. \tag{2.7}$$

Where $\hat{\xi}_{HJ} = \hat{F}_{HJ}^{-1}(\beta)$ is the β-th quantile of the $\hat{F}_{HJ}(y)$.

2.3 The Jackknife Empirical Likelihood for the low income proportion

Jackknife empirical likelihood (JEL) combines two popular approaches: the jackknife and the empirical likelihood. The key idea of the JEL is to turn the statistic of interest into a sample mean based on jackknife pseudo-values (Quenouille, 1956)[17]. We can apply Owen's empirical likelihood for the mean of the jackknife pseudo-values when these values are asymptotically independent (Jing, 2009)[16].

We describe the JEL method in a general way, let Ω_n be a consistent estimate for a low

income proportion

$$\Omega_n = \Omega((Y_1, X_1, D_1), (Y_2, X_2, D_2), ..., (Y_n, X_n, D_n)), \tag{2.8}$$

for example, we can take $\Omega_n = \hat{\Omega}_{HT}$, or $\hat{\Omega}_{HJ}$.

Define the jackknife pseudo-values by:

$$\hat{V}_{J,i} = n\Omega_n - (n-1)\Omega_{n-1}^{(-i)}, \tag{2.9}$$

where $\Omega_{n-1}^{(-i)} = \Omega((Y_1, X_1, D_1), ..., (Y_{i-1}, X_{i-1}, D_{i-1}), (Y_{i+1}, X_{i+1}, D_{i+1}), ..., (Y_n, X_n, D_n))$ which is the statistic Ω_n computed on the sample, by deleting the $i-th$ observation from the original data set.

In this thesis, first of all, we should define the *empirical likelihood function* for the $\Omega_{\alpha\beta}$. The next step is to define the *jackknife empirical likelihood function* at $\Omega_{\alpha\beta}$. Finally, we can define the *jackknife empirical likelihood ratio* at $\Omega_{\alpha\beta}$ by using Lagrange multipliers.

(Yang et al.,2010)[14] proposed a plug-in EL for low income proportion. Let $Y_1, Y_2, ..., Y_n$ be a random sample from an unknown income distribution $F(y)$. From (2.3), for any α and β, we have $E[I(Y \leq \alpha\xi_\beta)] - \Omega_{\alpha\beta} = 0$. Based on this equation, and the Empirical Likelihood function for the low income proportion $\Omega_{\alpha\beta}$ can be defined as:

$$\tilde{L}(\Omega_{\alpha\beta}) = sup_p\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i V_i = 0\}, \tag{2.10}$$

where $p = (p_1, p_2, ..., p_n)$ is a probability vector, $V_i = I(Y_i \leq \alpha\xi_\beta) - \Omega_{\alpha\beta}, i = 1, 2, ..., n$.

However, V_i depends on the unknown population quantile ξ_β, $L(\Omega_{\alpha\beta})$ is still unknown. Therefore, they used the sample quantile $\hat{\xi}_\beta$ to replace ξ_β in equation (2.10), and get the following profile Empirical Likelihood function for $\Omega_{\alpha\beta}$ (Yang et al.,2010)[14]:

$$L(\Omega_{\alpha\beta}) = sup_p\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i = 0\}, \tag{2.11}$$

where $\hat{V}_i = I(Y_i \leq \alpha\hat{\xi}_\beta) - \Omega_{\alpha\beta}, i = 1, 2, ..., n.$

Based on Yang el at.(2010)[14] showed the limiting distribution of empirical log-likelihood ratio is a scaled chi-square distribution which makes the inference complicated. To avoid the estimation of unknown scaled constant, we proposed JEL-based method for the low income proportion. The *Jackknife Empirical Likelihood function* for $\Omega_{\alpha\beta}$ is defined as

$$L(\Omega_{\alpha\beta}) = sup_p\{\prod_{i=1}^{n} p_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(\hat{V}_{J,i} - \Omega_{\alpha\beta}) = 0\}, \tag{2.12}$$

where $p = (p_1, p_2, ..., p_n)$ is a probability vector. By using the Lagrange multipliers, we obtain

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(\hat{V}_{J,i} - \Omega_{\alpha\beta})} \tag{2.13}$$

where λ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_{J,i} - \Omega_{\alpha\beta}}{1 + \lambda(\hat{V}_{J,i} - \Omega_{\alpha\beta})} = 0 \tag{2.14}$$

Note that $\prod_{i=1}^{n} p_i$, under constraints $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for all i, attains its maximum n^{-n} at $p_i = n^{-1}$. So we can define the *jackknife empirical likelihood ratio* for $\Omega_{\alpha\beta}$ as:

$$R(\Omega_{\alpha\beta}) = \frac{L(\Omega_{\alpha\beta})}{n^{-n}} = max\{\prod_{i=1}^{n}(np_i) : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i(\hat{V}_{J,i} - \Omega_{\alpha\beta}) = 0\} \tag{2.15}$$

Pluging the p_i's (2.13) into (2.15) and taking the logarithm of $R(\Omega_{\alpha\beta})$, we obtain *jackknife empirical log-likelihood ratio*:

$$logR(\Omega_{\alpha\beta}) = -\sum_{i=1}^{n} log\{1 + \lambda(\hat{V}_{J,i} - \Omega_{\alpha\beta})\} \tag{2.16}$$

We conjecture that the jackknife empirical log-likelihood ratio has a limiting chi-square

distribution:

$$-2logR(\Omega_{\alpha\beta}) \xrightarrow{d} \chi_1^2 \tag{2.17}$$

where χ_1^2 denotes a chi-squared distribution with one degree of freedom, and $\Omega_{\alpha\beta}$ represent the true value of the low income proportion.

Based on the conjecture, we can construct an approximate ρ-th level JEL-based confidence interval for $\Omega_{\alpha\beta}$:

$$I_\rho = \{\Omega_{\alpha\beta} : -2logR(\Omega_{\alpha\beta}) \leq c\}, \tag{2.18}$$

where c is chosen to satisfy $P(\chi_1^2 \leq c) = \rho$. If (2.17) holds, the $\lim_{n\to\infty} P\{\Omega_{\alpha\beta} \in I_c\} = P(\chi_1^2 \leq c) = \rho$. That is, the interval I_ρ gives asymptotically correct coverage probability.

PART 3

SIMULATION

In this section, we study the finite-sample performance of our JEL method on low income proportions with missing data. We conduct simulation studies to evaluate the coverage accuracy of the JEL-based intervals with $\Omega_n = \hat{\Omega}_{HT}$ or $\hat{\Omega}_{HJ}$ when $\alpha = 0.4, 0.5$ and $\beta = 0.4, 0.5$, respectively. We generate $B = 1000$ samples, $(y_1^{(b)}, x_1^{(b)}, d_1^{(b)}), ..., (y_i^{(b)}, x_i^{(b)}, d_i^{(b)}), ..., (y_n^{(b)}, x_n^{(b)}, d_n^{(b)})$, $b = 1, 2, ..., B$, from the underlying models, for each sample, we use our proposed JEL method to construct 95% JEL-based intervals I_ρ^b, $\rho = 0.95$,b=1,2,...,B.

Therefore, the coverage probabilities of JEL-based interval is given by:

$$CP = \frac{1}{B}\sum_{b=1}^{B} I(\Omega_{\alpha\beta} \in I_\rho^b) \tag{3.1}$$

The sample sizes are chosen to be $n = 50, 100, 300$. We compute the coverage probabilities of 95% confidence intervals for the low income proportion with the missing data based on $B = 1000$ repetitions. The following examples will be used in our simulation studies.

3.1 Examples

In the first example, we generated $(y_j, x_{1j}, x_{2j})'s$ from the model: $y_j = 2+3x_{1j}+x_{2j}+\epsilon_j$, where $x_{1j} \sim N(0,1), x_{2j} \sim N(0,1)$, $\epsilon_j \sim t_2$ where t_2 is a t-distribution with degree of freedom 2. In the second example, the model is same as the first example, except that $\epsilon_j \sim lognorm(0,1)$. In both examples, we generate d_j's from the following logistic model:

$$P(D = 1|x_{1j}, x_{2j}) = \frac{exp(1 + 2x_{1j} + x_{2j})}{1 + exp(1 + 2x_{1j} + x_{2j})}$$

The simulation results on estimating the coverage probability for the JEL-based confidence intervals of low income proportions with the missing data are summarized in Tables

[3.1, 3.2].

Table (3.1) $\epsilon_j \sim t_2$: Coverage Probabilities of 95% confidence intervals for low income proportions with the missing data

n	α	β	HT	HJ
50	0.4	0.4	0.9572	0.9601
	0.4	0.5	0.9547	0.9596
	0.5	0.4	0.9528	0.9534
	0.5	0.5	0.9503	0.9518
100	0.4	0.4	0.9489	0.9502
	0.4	0.5	0.9466	0.9499
	0.5	0.4	0.9489	0.9487
	0.5	0.5	0.9493	0.9525
300	0.4	0.4	0.9296	0.9412
	0.4	0.5	0.9236	0.9446
	0.5	0.4	0.9178	0.9432
	0.5	0.5	0.9145	0.9503

Table (3.2) $\epsilon_j \sim lognorm(0,1)$: Coverage Probabilities of 95% confidence intervals for low income proportions with the missing data

n	α	β	HT	HJ
50	0.4	0.4	0.9379	0.9309
	0.4	0.5	0.9354	0.9344
	0.5	0.4	0.9338	0.9347
	0.5	0.5	0.9228	0.9304
100	0.4	0.4	0.9365	0.9357
	0.4	0.5	0.9280	0.9361
	0.5	0.4	0.9351	0.9357
	0.5	0.5	0.9248	0.9368
300	0.4	0.4	0.9237	0.9464
	0.4	0.5	0.9205	0.9448
	0.5	0.4	0.9054	0.9457
	0.5	0.5	0.9077	0.9439

3.2 Summary of the simulation results

From two tables, we can observe follows: First of all, JEL methods work well for both HT and HJ estimators when construct the confidence interval for low income proportion with missing data. In most cases, the coverage probabilities are close to the nominal level 95%. Secondly, The JEL-based HJ estimator performed better than JEL-based HT estimator in general. Third, as the sample size n increases, the coverage probability of HJ estimator is much closer to the nominal level. However, the coverage probability of the JEL-based HT interval appears to under-cover the true income proportions. In summary, we recommend to use the HJ estimator to construct JEL-based interval for the low income proportion with the missing value.

PART 4

A REAL EXAMPLE

In this section, we apply our JEL method to a real economic observational data set. Lalonde (1986)[20] collected the original data, and we choose a subset of the data which was used by Lalonde (1986)[20], Dehejia and Wahba (1999)[21]. We can get the data set from the website http://www. nber.org/%7Erdehejia/nswdata.html. An alternative is to install the package of MatchIt, and load the data(lalonde)in R programm. There are several variables included in this data set. Real earnings in 1974,1975 and 1978 ("re74", "re75", "re78", respectively), age ("age"), educated years ("educ"), indicators for race ("black"), for marital status ("married"), for Hispanic ("hispanic"), for high school degree ("nodegree"),and for the status of participating the job training program ("treat", if participated in the program the value equals to 1, otherwise equals to 0).

We are interested in estimating the low income proportion of 1978's income (Y). There are n=445 individuals records in this dataset, 185 of them participated in the training program and 260 of them did not participate in the training program. We assume each individual who participated in the training program had a propensity score, it's reasonable because it's depend on covariant such as years of education, age, marital status, and some indicators for African-American, Hispanic- American, and degree level. We consider the age and education as the covariant variables (x_{1j}, x_{2j}) and consider the status of participating the job training program as the indicator variable (d). We assume the income data of the people who did not participate in the training program as the missing data. We constructed the model as follows:

$$P(D = 1|x_1, x_2) = \frac{exp(\eta_0 + \eta_1 x_1 + \eta_2 x_2)}{1 + exp(\eta_0 + \eta_1 x + \eta_2 x_2)}$$

We apply the recommended JEL-based intervals to the real earning year of 1978. For low income proportion estimation, HT method and HJ method are close to each other, and

they are 36% by HJ method and 38% by HT method. In addition, we calculate confidence intervals and the length via JEL method. The JEL-based interval using HJ estimate gives a 95% confidence interval of (0.2968, 0.4385) and the length is approximate to 0.1416. For HT estimate, the 95% confidence interval is (0.3154, 0.4576) and the length is around 0.1422.

PART 5

DISCUSSIONS

In this thesis, we have proposed Jackknife empirical likelihood-based inferences for a low income proportion with missing data by using HT and HJ estimators. First of all, JEL methods work well for both HT and HJ estimators when construct the confidence interval for low income proportion with missing data. In most cases, the coverage probabilities are close to the nominal level 95%. Secondly, HJ estimator performed better than HT estimator in general. Third, as the sample size n increases, the coverage probability of HJ estimator is much closer to the nominal level. However, the coverage probability of the JEL-based HT interval appears to under-cover the true income proportions. In summary, we recommend to use HJ estimator to construct JEL-based interval for the low income proportion with the missing value. In further studies JEL method will be needed for more complicated missing data such as non-ignorable missing data problems. (Vardi, 1985 [22] and Qin, 1993 [23]).

REFERENCES

[1] T. M. Smeeding, L. Rainwater, R. Rein, M.and Hauser, and G. Schaber, "Income poverty in seven countries: initial estimates from the lis database. in poverty, in equality and income distribution in comparative perspective," *The Luxembourg Income Study*, pp. 57–76, 1992.

[2] Eurostat, "In 2010, 17% of employees in the eu were low-wage earners," *Statistics in Focus: Population and Social Conditions*, 2010.

[3] B. Zheng, "Statistical inference for poverty measures with relative poverty lines," *Journal of Econometrics*, vol. 101, pp. 337–356, 2001.

[4] G. B. Yves and C. J. Skinner, "Variance estimation for a low income proportion," *Appl. Statist.*, vol. 52, pp. 457–468, 2003.

[5] Shao and Rao, "Standard errors for low income proportions estimated from stratified multistage samples," *Sankhya Ser*, vol. 55, pp. 393–414, 1993.

[6] I. Preston, "Sampling distributions of relative poverty statistics," *Appl. Statist.*, vol. 44, no. 1, pp. 91–99, 1995.

[7] R. J. A. Little and D. B. Rubin, "Statistical analysis with missing data," *New York: Wiley*, 2002.

[8] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, pp. 581–590, 1976.

[9] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring, "Missing data methods for generalized linear models," *A Comparative Review, "Journal of American Statistical Association"*, vol. 100, pp. 332–346, 2005.

[10] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement

from a finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 663–685, Dec 1952.

[11] A. H. Dorfman and R. Valliant, "The hájek estimator revisited," *Proceedings-Section On Survey Research Methods American Statistical Association*, pp. 760–765, 1997.

[12] Owen, "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, vol. 75, no. 2, pp. 237–249, Jun 1988.

[13] J. Chen and J. Qin, "Empirical likelihood estimation for finite populations and the effective usage of auxiliary information," *Biometrika*, vol. 80, no. 1, pp. 107–116, Mar 1993.

[14] B. Y. Yang, G. S. Qin, and J. Qin, "Empirical likelihood-based inferences for a low income proportion," *The Canadian Journal of Statistics*, vol. 39, no. 1, pp. 1–16, 2010.

[15] X. H. Zhou, G. S. Qin, H. Z. Lin, and G. Li, "Inferences in censored cost regression models with empirical likelihood," *Statistica Sinica*, vol. 16, pp. 1213–1232, 2006.

[16] B. Y. Jing, J. Yuan, and W. Zhou, "Jackknife empirical likelihood, journal of the amerian," *Statistical Association*, vol. 104, pp. 1224–1232, 2009.

[17] M. Quenouille, *BiometrikaNotes on Bias in Estimation*, no. 43, pp. 353–360.

[18] F. Gokpinar and Y. A. Ozdemir, "A horvitz-thompson estimator of the population mean using inclusion probabilities of ranked set sampling," *Communications in Statistics' Theory and Methods*, vol. 41, pp. 1029–1039, 2012.

[19] J. Hájek, "Comment on "an essay on the logical foundations of survey sampling part one"," *The Foundations of Survey Sampling*.

[20] R. J. Lalonde, "Evaluating the econometric evaluations of training programs with experimental data," *Amercian Economic Review*, no. 76, pp. 604–620, 1986.

[21] R. H. Dehejia and S. Wahba, "Casual effects in nonexperimental studies: Reevaluating the evaluation of training program," *JASA*, no. 94, pp. 1053–1062, 1999.

[22] Y. Vardi, "Empirical distribution in selection bias models," *Ann. Statist.*, no. 13, pp. 178–203, 1985.

[23] J. Qin, "Empirical likelihood in biased sample problems," *Ann. Statist.*, no. 21, pp. 1182–1196, 1993.